2014

Using Differential Functioning of Items and Tests (DFIT) to Examine Targeted Differential Item Functioning

Erin L. O'Brien
Wright State University

USING DIFFERENTIAL FUNCTIONING OF ITEMS AND TESTS (DFIT) TO EXAMINE TARGETED DIFFERENTIAL ITEM FUNCTIONING

A dissertation submitted in partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

By

ERIN L. O'BRIEN
M.S. Wright State University, 2009

2014
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

January 19, 2015

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Erin L O'Brien ENTITLED Using Differential Functioning of Items and Tests (DFIT) to Examine Targeted Differential Item Functioning BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

_____

David LaHuis, Ph.D.
Dissertation Director

_____

Debra Steele-Johnson, Ph.D.
Chair, Psychology Ph.D. Program

_____

Robert E. W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

Committee on Final Examination

_____
Debra Steele-Johnson, Ph.D.

_____
Nathan Bowling, Ph.D.

_____
Robert Gilkey, Ph.D.

ABSTRACT

O'Brien, Erin L. Ph.D., Industrial/Organizational Psychology Ph.D. program, Wright State University, 2015, Using Differential Functioning of Items and Tests (DFIT) to Examine Targeted Differential Item Functioning.

Current studies of differential item functioning (DIF) look at look at how groups differ in responding to items across an entire trait continuum. This is important for detecting the presence of consistent patterns of responses across items between groups of people. Current tests of DIF are limited in that they only detect differences between groups across all levels of the trait. However, selection decisions are usually made within specific ranges of trait levels. The purpose of this research was to determine if restricting theta values in an existing framework would be better at detecting DIF as current methods for restricted ranges of the trait continuum. This Monte Carlo study used a 3 (difficulty DIF) by 4 (discrimination DIF) by 2 (canceling versus noncanceling DIF) design. Traditional differential functioning of items and tests (DFIT) framework analyses were used and then rerun using the targeted ranges of theta. The targeted ranges were defined as the 100 lowest and 100 highest theta values. Type I error rates and power analyses were examined. Results indicate that it is possible to detect DIF accurately at specific trait levels when DIF was not detected across the entire range of theta values. This research has implications for using cut scores at particular levels of a trait for items that have not been assessed using the new, targeted ranges. Limitations and future research are discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction and Purpose

Current studies of differential item functioning (DIF) have examined how groups differ in responding to items across a trait continuum. This is important for detecting the presence of consistent patterns of responses across items between groups of people. For example, it is important to determine whether there are differences in the way applicants and incumbents respond to personality tests if those tests are created using only incumbent responses. These studies have examined group differences across all levels of the trait. In this study, I describe a method for detecting group differences at specific trait levels. Current tests of DIF are limited in that they only detect differences between groups across all levels of the trait. Figure 1 presents a graph of an item that shows a difference in responses between applicants and incumbents at higher levels of emotional stability but DIF was not detected (O'Brien & LaHuis, 2011). Responses between the two groups were not found to be significant, but there is a clear difference in responses.

Selection decisions are usually made within specific ranges of trait levels. For example, an organization may only be interested in individuals who score very high on certain items or tests of Conscientiousness. In this introduction, I discuss ways of modeling item responses using item response theory (IRT), different tests of DIF, how personality measures are used in the workplace, how I plan to develop the test of targeted DIF, and finally the research questions that I will be addressing.

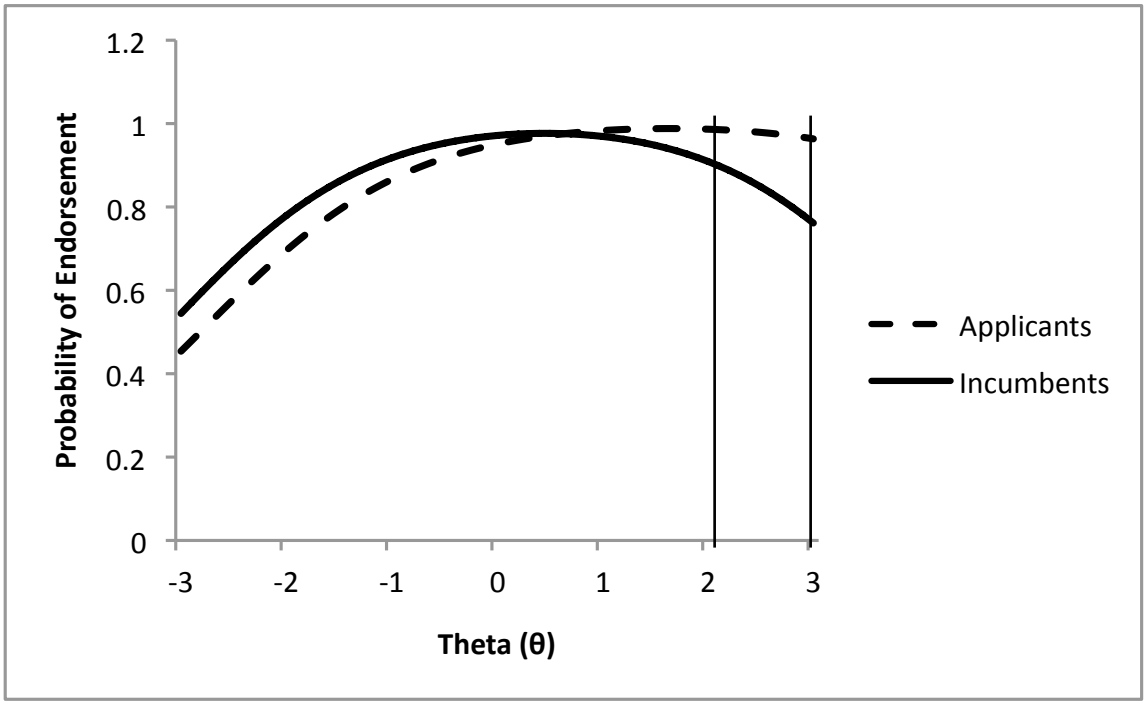

Figure 1. Example of a case when DIF was not detected across the entire continuum of a trait (theta), but there appeared to be DIF at some values of theta (indicated by the vertical lines).

Item Response Theory

Item response theory is a model-based measurement method, which evaluates the probability of response to an item based on the underlying level of the respondent. The probability of endorsing an item is a function of both person and item parameters. The person parameter indicates an individual's standing on a latent trait. Item parameters determine the shape of the item response function (IRF) and often consist of a discrimination and difficulty parameter. In most IRT models, the discrimination parameter, α, determines the slope of the curve, or how the probabilities change with the trait level (Embretson & Reise, 2000). This provides information about how well an item can discriminate between people of varying trait levels. Values of the discrimination parameter usually vary between 0.5 and 1.5 (Reise & Waller, 2002). Figure 2 plots two

IRF's with different discrimination values. These items have differing slopes on the graph. Item 1 has a smaller α value, and therefore a less steep slope. Item 2 has a larger α value, and therefore a steeper slope. Thus, Item 2 is a more discriminating item.

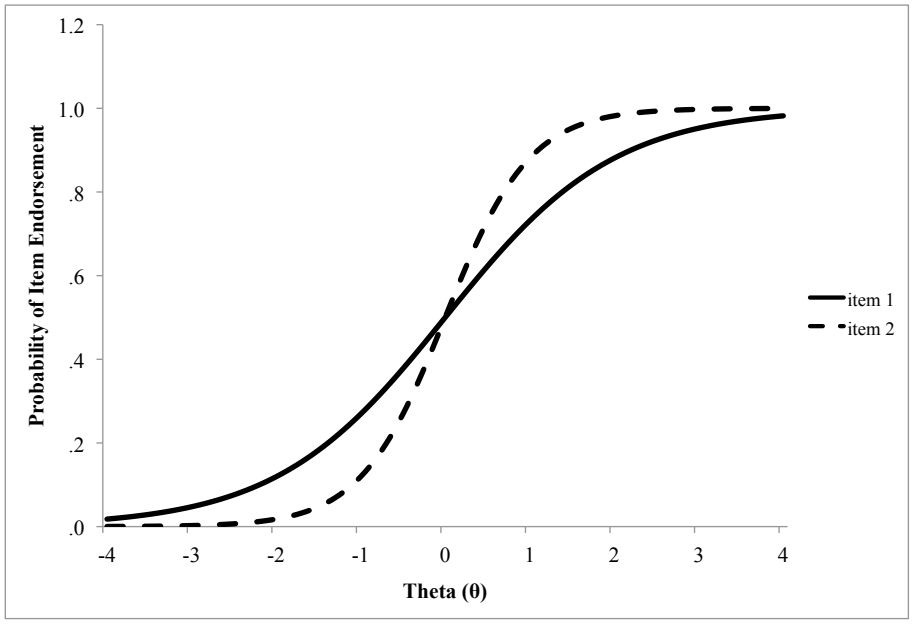

Figure 2. Example IRF's showing different discrimination parameters.

The difficulty parameter was named when IRT was almost used exclusively for multiple-choice tests with right and wrong answers. The terminology is still used today even though researchers do not think typically of personality items as being more or less difficult. The difficulty parameter, β, impacts the location on the latent trait where there is a 50% chance of endorsing the item (i.e., this parameter shifts the curve along the x-axis). Figure 3 shows two IRF's with different difficulty values for items with two response options. Item 1 has a difficulty value of 0, and Item 2 has a difficulty value of 1. Item 1 is an easier item than Item 2, or a person must have a higher level of the trait before they are likely to endorse the item. When researchers have more than two

response options, difficulty parameters can be thought of as thresholds. In general, there are m-1 β parameters where m is the number of response options.

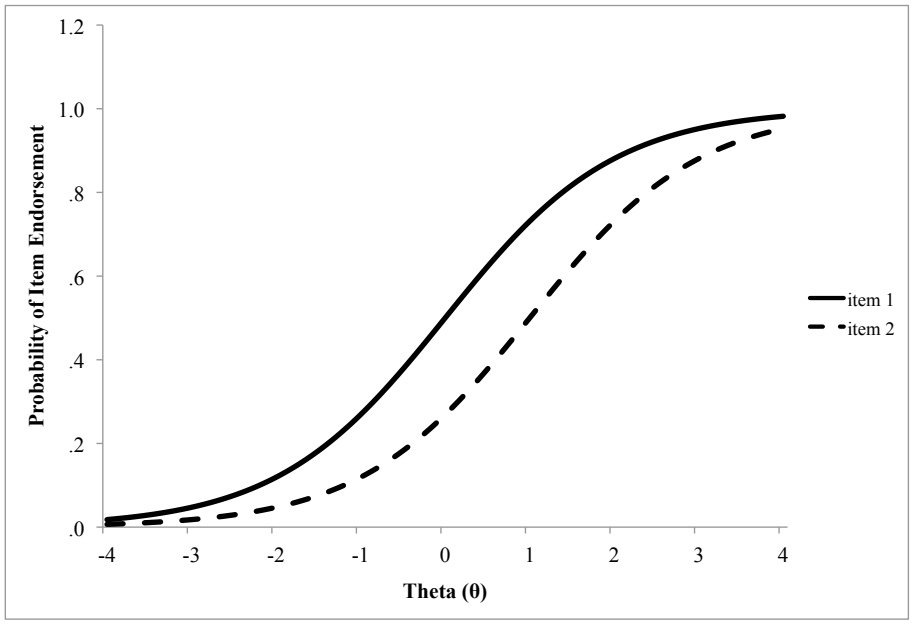

Figure 3. Example IRF's showing different difficulty parameters.

Scales can either have dichotomous or polytomous responses/items. Dichotomous items have only two possible responses. For example, the answers to a question may be 'yes' and 'no' or 'agree' and 'disagree'. Polytomous items have more than two responses. This is seen often in personality measures in which the response options range from 'Strongly Agree' to 'Strongly Disagree'. In this study, I will examine polytomous data. In the next sections, I will describe the different IRT models used for both dichotomous and polytomous data.

Dichotomous models. The most basic model in IRT is the Rasch Model (Embretson & Reise, 2000) or the one-parameter logistic model. The one-parameter model uses the difficulty parameter. The discrimination parameter is not included.

Therefore, the IRF's for this model will all have the same slope value, as seen in Figure 2. This model generally leaves the discrimination parameter set to 1.0 although some researchers set α to a predefined constant value.

One of the most common models used in personality scales is the two-parameter logistic model (2PLM). This model is very similar to the one-parameter logistic model, except with this model the discrimination parameter is allowed to vary per individual. The formula for the 2PLM is

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

where $P_{ij}(Y=1|\theta_j)$ is the probability that person *i* will endorse item *j* as a function of their trait level, θ represents the individual's trait level, α is the discrimination parameter for item i, and β is the difficulty parameter for item i. This model allows for items to be differently related to the trait level; some items may be more or less related to the trait.

Another popular IRT model is the three-parameter logistic model (3PLM). This model includes a guessing parameter. The guessing parameter is used most often with multiple choice tests, where there is a correct answer. The formula for the 3PLM is

$$P_{ij}(Y=1|\theta_j,\gamma_i)=\gamma_i+(1-\gamma_i)\frac{\exp[\alpha_i(\theta_j-\beta_i)]}{1+\exp[\alpha_i(\theta_j-\beta_i)]} \quad (2)$$

where $P_{ij}(Y=1|\theta_j,\gamma)$ is the probability that person *i* will endorse item *j* as a function of their trait level, θ represents the individual's trait level, α is the discrimination parameter for item *i*, β is the difficulty parameter for item *i*, and γ represents the guessing parameter. Researchers mostly use the guessing parameter with multiple-choice items in which individuals have a chance of getting an item correct if they just guessed. For example, a question with four possible answers has a starting value of .25 because an individual has

a 25% chance of getting the item correct if they just guessed the answer. This model is seldom used with personality measures because interpreting a guessing parameter is not very clear.

Polytomous models. There are several polytomous IRT models, including modified-GRM, partial credit model, generalized partial credit model, rating scale model, and the nominal response model. These models differ in how they calculate the probability of a response to a particular category and what type of data they are designed to analyze. For example, the nominal response model (NRM; Bock, 1972) was designed for responses that are not necessarily rated along a continuum.

For the purposes of my study, I used the graded response model (GRM; Samejima, 1969). This particular model is an extension of the 2PLM. It requires a two-step process to calculate the probability of responding for each category and is therefore referred to as an "indirect" model. The GRM is the model most commonly used with Likert-type items with multiple response options, such as those used in personality scales.

In the GRM, each item (i) has *m* number of ordered response options and *m* – 1 boundary response functions (BRF's) (Embretson & Reise, 2000). Each BRF is treated as a dichotomy. Each BRF compares the probability of endorsing option 1 versus Options 2 through 5 for a five-response option item, for example. The next BRF is the probability of choosing options 1 and 2 versus options 3, 4, and 5, and so on until there are *m* – 1 number of BRF's. The equation for a BRF is

$$P_{ix}^{*}(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ij})]}{1 + \exp[\alpha_i(\theta - \beta_{ij})]} \tag{3}$$

where θ is the trait level, α_{ij} is the discrimination parameter, and β_{ij} is the difficulty parameter. To calculate the probability of each response option, we used the following formulas

$$P_{i0}(\theta) = 1.0 - P_{i1}^{*}(\theta) \tag{4}$$

$$P_{i1}(\theta) = P_{i1}^{*}(\theta) - P_{i2}^{*}(\theta) \tag{5}$$

$$P_{i2}(\theta) = P_{i2}^{*}(\theta) - P_{i3}^{*}(\theta) \tag{6}$$

$$P_{i3}(\theta) = P_{i3}^{*}(\theta) - P_{i4}^{*}(\theta) \tag{7}$$

$$P_{i4}(\theta) = P_{i4}^{*}(\theta) - 0 \tag{8}$$

Differential Item Functioning

Definition and purpose of DIF. People who have the same knowledge of a particular topic should respond to questions about that topic similarly. For example, if men and women know the same thing about the history of World War I, then both groups should have the same probability of getting a question about that war correct. However, if men and women systematically differ in their responses to that item, then the item is said to function differently across groups. This is what is referred to as differential item functioning, or DIF.

Types of DIF. Groups' responses on an item can cause parameters to differ. For example, two groups can differ in their discrimination. Figure 2 plots two IRF's with different α parameters. The discrimination parameter is the same for both items. Item 1 has a smaller α value and therefore a less steep slope. Item 2 has a larger α value and therefore a steeper slope. This is an example of DIF with the α parameter.

The difficulty parameter, β, determines the location on θ where there is a 50% chance of endorsing the item. Figure 3 plots two IRF's with different β parameters. The slope is the same for both items. Item 1 has a lower β value and is more readily endorsed. Item 2 has a higher β value and is a more difficult item. This is an example of DIF with the β parameter. Another type of DIF we are interested in is the interaction between α and β DIF. Researchers noted that when both α and β DIF were present it showed up as β DIF. That is, α DIF was not detected very often.

IRT methods. The basic IRT procedure is to fit an IRT model to the data for two different groups, often referred to as the reference and focal groups, and test the significance of the difference between the item parameter estimates for the two groups. If the difference is significant, the item is said to exhibit DIF.

Likelihood ratio test. The Likelihood Ratio Test (LRT) is another method for detecting DIF. In general, it compares statistically significant difference between item parameters. This method can be used to only detect DIF at the item level. LRT estimates two models: a compact and an augmented model. The compact model constrains all items in the test to be equal. The compact model assumes there are no differences between the two groups. The augmented model allows at least one item to not be constrained. This allows for testing item or items that are believed to have DIF. The models are compared then using a chi-square statistic, G2. The equation is

$$G^2(df) = -2\log(likelihoodC / likelihoodA) \tag{9}$$

where C refers to the compact model and A refers to the augmented model. The metric of both models is the same as an anchor set of items that do not have DIF.

Differential functioning of items and tests (DFIT) framework. Up until the mid 1990's, researchers could use one of several techniques to detect differences in groups' responses on an item. Each technique differed in the way it calculated the group differences for an item. The drawback to all of the techniques available is their inability to detect differences across an entire set of items. Being able to detect differences across the test is important because some items may benefit one group whereas other items may benefit a different group. These items would have been marked as having DIF even if they did not make the overall test biased. Generally, researchers remove or modify an item from the entire test (Raju, van der Linden, & Fleer, 1995), and then the test is considered unbiased. Researchers developed a framework, the differential functioning of items and tests framework (DFIT), to test each item, in the same manner as previous researchers, but they included a way to also detect DIF across all the items for a particular test (Raju, van der Linden, & Fleer, 1995). This was beneficial because some items might have been an advantage for one group whereas another item might have benefitted the other group. For example, conscientiousness Item 1 might have benefitted men whereas conscientiousness Item 2 might have benefitted women. The process of being able to detect DIF across the entire test allowed for all items to remain in the test without the test being biased overall. Also, it allowed for a way to detect the effect of an item on an entire test by removing or adding it to the test.

Previous research has suggested that using the DFIT framework (Raju, van der Linden, & Fleer, 1995) is the most appropriate way to detect DIF in organizational research. This framework provides a way of detecting DIF at the item level as well as at the test level. One DIF index at the item level, the noncompensatory DIF index (NCDIF),

assumes independence from all other items of the test. What this means is that it has the ability to detect DIF for an item, even if that item is canceled out at the test level. The NCDIF index compares items expected scores. If d_i equals the difference between the probabilities of item endorsement under the focal and referent group parameters then

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

where σ and μ are standard deviations and means of d_i, respectively. NCDIF is best used if there are concerns about particular questions disadvantaging some groups (Raju, et al., 1995).

The other item level index, the compensatory DIF index (CDIF), does not assume independence of items. The CDIF index considers all items and marks the items that would lower overall test differences if removed. CDIF differs from NCDIF in that items with DIF may not be marked because they are canceled out by other items and therefore do not affect the overall test.

$$CDIF_i = Cov(d_i, D) + \mu_{di}\mu_D \quad (11)$$

Researchers often use CDIF when it is necessary to keep items with DIF in the test (Raju, et al., 1995).

The index to detect differences at the test level is called differential test functioning (DTF). DTF tests differences for the entire test, rather than just individual items. Whereas some items might function differentially, they might cancel each other out, creating a test with no DIF. Researchers calculate the difference in true scores by summing the differences for the items.

$$D = \sum^{n} d_i \tag{12}$$

Researchers calculate DTF by squaring these test level differences

$$\text{DTF} = E_F(D^2) = \sigma_D{}^2 + \mu_D{}^2 \tag{13}$$

According to Raju, van der Linden, and Fleer (1995), DTF is most appropriate to use when total test scores are used to develop and test instruments.

Item parameter replication. Researchers still are determining appropriate statistical cutoffs for the DFIT framework. Initially, researchers declared specific cutoff values for both dichotomous and polytomous items (Fleer, 1993; Flowers, et al., 1999; Raju et al., 1995). However, researchers found those values resulted in too many Type I errors. Several studies (Bolt, 2002; Flowers et al., 1999; Meade, Lautenschlager, & Johnson, 2007) encouraged using empirically derived cutoff values. However, this method was very time-consuming and requires complex calculations.

Researchers developed a method for calculating cutoffs for NCDIF statistics using dichotomous IRT models. Oshima et al. (2006) called this method the item parameter replication (IPR) method. Later, researchers extended this to polytomous data (Raju, et al., 2009). Researchers use the estimated item parameters to calculate cutoffs for each item. They do this by using the variance-covariance structures of the item parameters. Researchers then use these structures to simulate item parameters for a large number of samples. Then they calculate NCDIF statistics for each sample. Researchers then use the distribution of NCDIF values and calculate the cutoff based on a chosen alpha level. For example, if alpha were set to 0.05, the value at the 95th percentile would be the cutoff

value.

Problems with DIF

One of the problems with detecting DIF using an IRT method is that the item parameters are on different metrics making a direct comparison inaccurate. Ideally, researchers would collect the appropriate data from both the reference and focal groups, estimate their respective item parameters, and then compare the IRF's. However, the process for estimating parameters separately creates different metrics for both the reference and focal groups, making it impossible to simply compare their graphs. Researchers estimate the parameters for each group separately using the mean of each group. Because the means of each group will be different, they end up with different metrics. To account for this discrepancy, researchers need to link the item parameters.

Previous research using applicants and incumbents linked the reference group parameters to the focal group parameter metric (Raju et al., 1995; Robie et al., 2001). Linking constants need to be obtained using an iterative process. Researchers compute the constants using the ICC method based on item response functions. The calculation removes items with DIF and then repeats the ICC method until the same items are identified as having DIF on consecutive iterations.

LRT and DFIT

LRT compares two models, is based on a chi-square statistic, and uses a test of statistical significance. On the other hand, the DFIT framework compares IRF's and uses empirically-derived cutoff values to detect DIF. LRT uses chi-square statistic in determining significance, which makes it more likely to detect DIF for larger sample sizes (Braddy, 2004).

DFIT can detect DIF at test level whereas LRT can detect DIF only at the item level. The DFIT framework requires the extra step of linking to put all the item parameters onto the same metric. LRT handles this problem with the anchor set of items. DFIT framework is capable of detecting DIF at the item level as well as at the test level. LRT cannot test for DIF at the item level.

Clark and LaHuis (2011) found that LRT was better at detecting DIF with the alpha, or discrimination parameter, whereas the DFIT framework was better at detecting differences at higher levels of the β, or difficulty parameter. They found that the DFIT framework resulted in higher Type I error rates with unequal sample sizes. The authors found that LRT had greater power to detect DIF for the discrimination parameter when there was a large change in α, and LRT was better at detecting β parameter DIF. Whereas LRT has many advantages, I will be using the NCDIF framework. I can specify the ranges of theta that I am interested in analyzing with the NCDIF framework. LRT does not allow for analysis of specific ranges of theta.

Purpose

The purpose of this research was to create a way to target DIF within specific levels of a trait. I used the NCDIF framework, but instead of using the entire range of theta values I used a specific range of theta values. For example, researchers are interested often in differences at higher levels of the personality trait Conscientiousness. Instead of analyzing the entire range of thetas for this trait, I looked only at the range of thetas of interest.

Research Questions

To determine whether this revised method is effective, I addressed several research questions.

Research Question 1: How well does this new procedure detect only α DIF? Previous methods of DIF detection have not been able to distinguish α DIF accurately. Clark and LaHuis (2012) found that the Likelihood Ratio Test did a poor job of correctly identifying α DIF and over-identified β DIF when there was no β DIF present. One of the questions my research aimed to answer was whether this new method of targeting specific ranges of theta would be better at identifying only α DIF.

Research Question 2: How well does this new procedure detect only β DIF? Previous research has indicated that detecting β DIF is fairly easy (Clark & LaHuis, 2012). The main concern in my study was whether that remains true for the new, targeted areas. I wanted to make sure that this new method did not decrease the ability to detect β DIF.

Research Question 3: How well does this new method detect only canceling DIF?

Research Question 4: Does this revised method detect DIF better than the current DIF measure? The new, targeted method of DIF detection should be better than overall DIF because of its ability to target areas that are more prone to differences. DIF is more likely to cancel out across the entire item when the full range of theta values is analyzed. The new, targeted method should do a better job of detecting DIF.

Research Question 5: Is there a difference in DIF detection between the highest and lowest targeted areas? I created the targeted areas as the highest and lowest values of theta. I aimed to identify whether this new method is better at detecting DIF for the highest or the lowest theta values.

Method

Design

I created a 3 (difficulty DIF) by 4 (discrimination DIF) by 2 (canceling versus noncanceling DIF) design. I created 1000 simulee responses for each sample and I created 100 replications of each sample in each of the conditions. Based on Clark and LaHuis (2012), I created several DIF conditions for the difficulty parameter by adding values (0, 0.4, and 1.0) to the β parameter for the first four items. I created several DIF conditions for the discrimination parameter by adding values (0, 0.3, 0.5, and 0.7) to the α parameter for the first four test items (Meade & Lautenschlager, 2004). For example, I generated item parameters for each condition. Then I added a value, 0.3 for example, to the first four item's α parameters and kept the remaining six items with their original parameter values.

I created canceling DIF and noncanceling DIF conditions. The canceling DIF condition created DIF for the first two items for the referent group and the next two items for the focal group. For example, I added .3 to the α parameter for Items 1 and 2 for the referent group and added .3 to α parameter for Items 3 and 4 for the focal group. This cancels DIF across the entire test. The noncanceling DIF conditions simply created DIF for the first four items for only the focal group.

Data Generation

I used the computer program R for all of the data creation. Appendix A shows the computer code used for data creation. First, I created estimated θ values for each simulee from a random normal distribution with a mean of 0 and a standard deviation of 1. Next, I generated item parameters to obtain the discrimination and difficulty parameters. I

created the discrimination parameters using a random uniform distribution with a mean of .5 and a standard deviation of 2 (Meade, Lautenschlager, & Johnson, 2004). Because the data for this study were polytomous, I needed to create difficulty parameters for one minus the total number of response options. I created the first difficulty parameter using a random normal distribution with a mean of -1.7 and a standard deviation of .45 (Meade & Lautenschlager, 2004). Adding the constant 1.2 to the previous parameter created on each consecutive difficulty parameter. Then I calculated probabilities of endorsing each response option using the GRM equation for both the referent and focal groups. Finally, I generated random numbers between 0 and 1. I compared these values to the previously calculated 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.

Data Analysis

First, I estimated the item parameters. Then, I extracted the variances and covariances from the item parameters. Before I ran the DFIT analyses, I linked the parameters using the iterative Stocking and Lord (1983) method.

Next, I used the IPR method to determine the cutoff values for each item (Oshima, Raju, & Nanda, 2006). The IPR method uses the variance-covariance structures of estimated item parameters to produce separate cutoffs for each item. The variances and covariances are used to simulate item parameters for a large number of samples. NCDIF statistics are calculated for each sample, and then the distribution of the NCDIF values for each item is used to determine the cutoff. The value that is the 95th percentile was the

cutoff value for an α level of .05. Appendices B and C show the computer code used for creating the cutoffs.

I conducted the DFIT analyses by estimating person parameters for each condition. These parameters indicate each person's estimated level of that trait. I calculated traditional DFIT statistics using the entire range of theta. Then, I calculated the targeted DIF by only inputting the focal group thetas falling in the specified ranges. The targeted ranges were defined as the 100 lowest and 100 highest theta values. Then, I compared the DIF results from the traditional DIF analyses with the results of the targeted DIF analyses.

Kim, Cohen, Alagoz, and Kim (2007) recommended using Type I error and power rates for analyses with large sample sizes. I analyzed the Type I error rates. I calculated these rates by totaling the number of non-DIF items that were not expected to have DIF by the number of non-DIF specified items. I counted the number of items that had DIF from the six items that were not specified to have DIF and divided by six. These values were averaged across each of the 100 replications. I calculated power similarly, except I examined the number of DIF-identified items out of the total number of items created to have DIF.

Results

I created the theta ranges based on the highest 100 and lowest 100 theta values. The lowest 100 values ranged from -2.82 to -1.93 with a mean of -1.57 (M = -1.57, SD = .42). The highest 100 values ranged from 1.17 to 2.52 with a mean of 1.54 (M = 1.54, SD = .27).

To answer the first research question, does the new procedure detect only α DIF, I examined the Type I error rates and Power analyses when β DIF was zero. This is important because previous methods have not been able to accurately detect α DIF by itself. Table 1 shows the power and Type I error calculations when β DIF was zero. The highest 100 theta values and the entire theta range had acceptable Type I error rates. However, the highest 100 values had slightly lower rates. The highest 100 theta values had Type I error rates ranging from 0.04 to 0.05. The entire theta range had Type I error rates ranging from 0.05 to 0.06. The lowest 100 theta values did not have acceptable Type I error rates. The lowest 100 theta values Type I error rates ranged from 0.07 to 0.09. Whereas none of the conditions had acceptable power rates (80% or higher), the highest 100 theta values when α DIF was 0.7 and the entire theta range when α DIF was 0.7 had the highest power values at 39.13% and 30.25% respectively. The lowest power rates occurred for the entire theta range and the lowest 100 theta values when α DIF was 0.3 at 9.62% and 11.00% respectively. The highest 100 theta values and the entire theta range when α DIF was 0.7 had the strongest combination of Type I error and power. However, the power rates did not reach acceptable levels for either condition.

Table 1.

Power and Type I error calculations of α DIF when β DIF was zero.

	α DIF		
Theta Range	0.3	0.5	0.7
Entire range	9.62 (.06)	19.50 (.05)	30.25 (.06)
Highest 100	12.63 (.05)	25.38 (.04)	39.13 (.04)
Lowest 100	11.00 (.08)	15.50 (.07)	25.12 (.09)

Note: $N = 1800$, $n = 200$; Power rates presented as percentages; Type I error rates in parentheses.

To answer the second research question, does the new procedure detect only β DIF, I examined the Type I error rates and Power analyses when α DIF was zero. Table 2 shows the power and Type I error calculations when α DIF was zero. All three theta ranges (entire range, highest 100, and lowest 100) had acceptable Type I error rates when β DIF was 1.0, with all of the Type I error rates at 0.05. The lowest 100 theta values did not have acceptable Type I error rates when β DIF was 0.4, with an error rate of 0.07. All three theta ranges (entire range, highest 100, and lowest 100) had acceptable power rates when β DIF was 1.0. The power rates were 100.00%, 99.50%, and 96.50% respectively. None of the theta ranges had acceptable power rates when β DIF was 0.4. The power rates were 64.25% for the entire range, 59.12% for the highest 100, and 38.50% for the lowest 100 theta values. The entire theta range when β DIF is 1.0 had the best combination of Type I error and power rates (though the highest and lowest 100 have acceptable rates as well), while the lowest 100 theta values when β DIF was 0.4 had the worst Type I error and power rates.

Table 2.

Power and Type I error calculations of β DIF when α DIF was zero.

	β DIF	
Theta Range	0.4	1.0
Entire range	64.25 (.04)	100.00 (.05)
Highest 100	59.12 (.04)	99.50 (.05)
Lowest 100	38.50 (.07)	96.50 (.05)

Note: $N = 1200$, $n = 200$; Power rates presented as percentages; Type I error rates in parentheses.

To answer the third research question, how well does the new procedure detect only canceling DIF, I examined the Type I error rates and Power analyses after collapsing

across α and β DIF. Table 3 shows the results of the Power and Type I error rates analyses. The highest 100 and the entire theta range both had acceptable Type I error rates for both the canceling and noncanceling DIF conditions. The Type I error rates ranged from 0.04 to 0.05 for the highest theta ranges and they ranged from 0.05 to 0.06 for the entire theta range. The lowest 100 theta range did not have acceptable Type I error rates, with values ranging from 0.07 to 0.08 for both the noncanceling and canceling conditions, respectively. Whereas none of the conditions had acceptable power rates (0.80 or higher), the entire theta range and lowest 100 in the canceling DIF conditions had the highest power rates at 60.83% and 58.17% respectively. The highest 100 theta values had the lowest power rates for both the canceling and noncanceling conditions at 52.06% and 53.04% respectively. The entire theta range for both noncanceling and canceling DIF conditions had the best combination of Type I error and power rates, though the power rates were below acceptable levels. The highest and lowest 100 theta values had the worst power and Type I error rates respectively.

Table 3.

Power and Type I error calculations for noncanceling and canceling DIF.

	Canceling DIF	
Theta Range	Noncanceling	Canceling
Entire range	57.10 (.05)	60.83 (.06)
Highest 100	53.04 (.04)	52.06 (.05)
Lowest 100	57.25 (.07)	58.17 (.08)

Note: $N = 7200$, $n = 1200$; Power rates presented as percentages; Type I error rates in parentheses.

To answer the fourth research question, does the new procedure detect DIF better than the traditional method, I examined the Type I error rates and Power analyses across all conditions. Table 4 shows the results of the Type I error rates for all conditions and

Table 5 shows the Power analysis results. The highest 100 theta values had consistently acceptable Type I error rates ranging from 0.02 to 0.05. The only unacceptable rate occurred for the noncanceling β at 1.0 and α at 0.7 condition where the error rate was 0.09. The entire theta range had consistently acceptable Type I error rates for the noncanceling DIF conditions (except when β was 1.0 and α was 0.5 and 0.7). The Type I error rates ranged from 0.03 to 0.07 for the canceling DIF conditions. The lowest 100 theta values had the worst Type I error rates, with only one condition having an acceptable rate of 0.05 (canceling DIF, β was 1.0, and α was 0). The Type I error rates ranged from 0.05 to 0.12 for the lowest 100 theta values.

The only acceptable power rates occurred when β DIF was 1.0 for all three theta ranges. The entire range of theta had power rates of 100% for all conditions when β DIF was 1.0. The highest 100 theta values had power rates that ranged from 95-100%. The lowest 100 theta values had power rates that ranged from 81-100%. All other β DIF conditions had unacceptable power rates ranging from 6-76%. The power rates were particularly low when β DIF was 0. The best DIF detection occurred when β DIF was 1.0 for the highest 100 theta values, with the exception of the canceling condition when α was 0.7. The entire theta range was best when β DIF was 0 and α was 0.3 for the noncanceling condition or when β DIF was 1.0 and α was 0 for the canceling condition. The lowest 100 was best for the canceling condition when β DIF was 1.0 and α DIF was 0.

Table 4.

Type I error calculations of all conditions.

		α DIF							
		Non-Canceling				Canceling			
Theta Range	β DIF	0	0.3	0.5	0.7	0	0.3	0.5	0.7
Entire Range	0	0.04	0.05	0.05	0.05	0.06	0.07	0.05	0.06
	0.4	0.05	0.04	0.05	0.03	0.03	0.05	0.07	0.07
	1.0	0.06	0.05	0.06	0.10	0.04	0.07	0.06	0.06
Highest 100	0	0.03	0.05	0.04	0.03	0.04	0.05	0.05	0.05
	0.4	0.04	0.04	0.05	0.02	0.04	0.05	0.05	0.05
	1.0	0.05	0.04	0.04	0.09	0.05	0.05	0.04	0.04
Lowest 100	0	0.06	0.08	0.08	0.09	0.08	0.07	0.06	0.09
	0.4	0.08	0.07	0.06	0.06	0.06	0.06	0.08	0.07
	1.0	0.06	0.06	0.07	0.12	0.05	0.10	0.10	0.09

Note: $N = 7200$, $n = 100$

Table 5.

Power calculations of all conditions.

		α DIF							
		Non-Canceling				Canceling			
Theta Range	β DIF	0	0.3	0.5	0.7	0	0.3	0.5	0.7
Entire Range	0	-	6	14	20	-	13	25	40
	0.4	62	59	59	60	66	57	62	63
	1.0	100	100	100	100	100	100	100	100
Highest 100	0	-	9	13	17	-	13	18	33
	0.4	37	62	73	74	41	53	68	76
	1.0	98	100	99	100	95	98	99	99
Lowest 100	0	-	8	19	31	-	16	32	47
	0.4	61	45	38	41	58	36	36	42
	1.0	100	98	97	95	99	91	85	81

Note: $N = 7200$, $n = 100$; Power rates presented as percentages.

To answer the fifth research question, is there a difference in DIF detection between the highest and lowest targeted areas, I examined the Type I error rates and Power analyses across all conditions for the highest 100 and lowest 100. Tables 4 and 5 show the results of the Type I error and Power analyses for the two groups. The results indicate that the lowest 100 theta values had better power rates when β DIF and α DIF were 0. The only time this was not true was for the noncanceling condition when β DIF was 0 and α DIF was 0.3. It is important to note that the power rates were not acceptable except when β DIF was 1.0. The highest 100 theta values did, however, have acceptable Type I error rates while the lowest 100 theta values consistently did not.

Discussion

I found evidence that it is possible to detect DIF accurately at specific trait levels instead of across the entire trait continuum. Figure 4 shows an example item in which DIF was not detected across the entire range of thetas, but there was DIF at both the higher and lower trait levels. The new method of targeting specific levels of the trait was best when β DIF was 1.0. Overall, the new method was better when using the highest 100 theta values compared to the lowest 100 theta values. Individuals with low levels of a trait are less likely to endorse an item. It appears that the item parameters do not change enough at these low levels of a trait to detect differences. Though the results were not terribly strong, there is evidence to suggest that modifying the ranges of theta that are being analyzed has value.

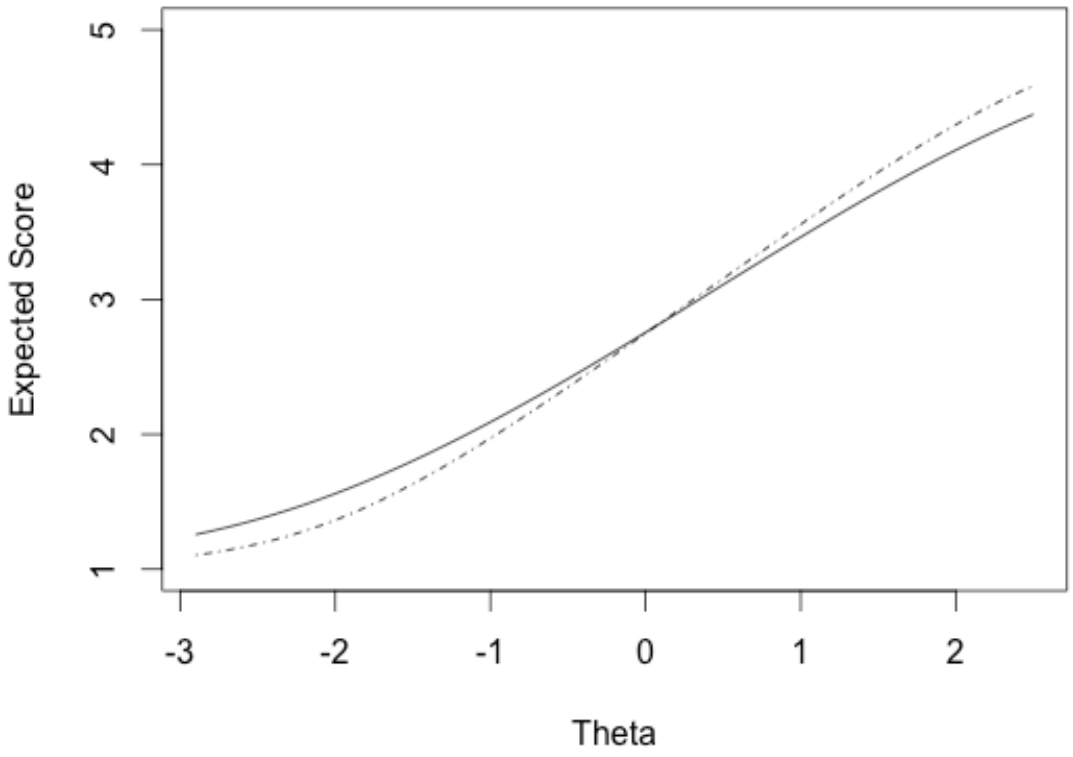

Figure 4. Example item with no DIF detected across the entire range of thetas, but with DIF detected at both high and low levels of theta.

The new method was not able to detect α DIF when β DIF was not present, compared to previous research. Clark and LaHuis (2012) found power rates of 22-23% when α DIF was 0.25 and 51-72% when α DIF was 0.5. The current study found power rates of 12% when α DIF was 0.3 and 25% when α DIF was 0.5. This is disappointing and difficult to explain given the data generation was similar in both studies.

The new targeted method was as good at detecting β DIF as the traditional method, though the results for the lowest 100 theta values were not as strong. There did not appear to be any difference in canceling DIF when comparing the new method to the traditional method.

Implications

Often organizations will determine a particular score an applicant must obtain on personality assessments to be considered in the hiring process. These cut scores help the organization to eliminate applicants from the large stack of applications they often receive. Usually a validation study is done to determine cut scores, in which incumbent responses and performance scores are used to create the personality assessments used in the selection process (Cascio & Aguinis, 2011). The current study aimed at preventing discrimination when using these cut scores. Current studies of DIF do not distinguish differences in responses between groups of people at particular levels of a trait. If people with the same level of a trait are responding differently at a specific level of that trait, then the item is discriminating between groups. For example, men and women with the same level of conscientiousness might respond to the item "I try to follow the rules" differently at higher levels of conscientiousness, whereas men might be less likely to endorse the item. An organization might be interested in hiring individuals who are more likely to endorse this item, which means men would be less likely to be hired even though they possess the same level of the trait. Current measures of DIF are not able to detect such subtle differences, but the current study found support that it is possible to detect these differences. This will be useful for organizations that are interested in creating fair, accurate selection tools or improving existing measures.

Limitations and Future Research

Perhaps the biggest limitation of the study was the low power rates for all the conditions where β DIF was not 1.0. This was true even when the entire range of theta values was used.

Future research needs to examine more appropriate cutoff values. Instead of examining only at the highest and lowest trait levels, researchers should investigate varying theta ranges. It would be practical to investigate cutoffs that matched specific hiring criteria. For example, if an organization makes hiring decisions at particular levels of a trait, it would be useful to determine whether this method works for those precise trait levels. The current study aimed to determine whether it was possible to create a targeted test of DIF. Given that I found evidence to support the application, it would be beneficial for researchers to determine cutoffs that make sense for each project, rather than simply using the cutoffs we provided.

Previous research using Monte Carlo data simulation showed little difference in DIF detection between 10 and 20 items (e.g., Clark & LaHuis, 2012). Also, researchers have indicated that personality tests usually only have about 10 items per trait (e.g., Collins, Raju, & Edwards, 2000; Meade, Lautenschlager, & Johnson, 2007; Stark, Chernyshenko, & Drasgow, 2004), making it more applicable to personality traits. However, some researchers have preferred to use a greater number of items when assessing DIF (e.g., Flowers, Oshima, & Raju, 1999; Zumbo, 1999) Future research should examine results when more items are used for each scale.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (3-24). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Angoff, W. H. & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-106.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1, Pt. 1), 29-51. doi:10.1007/BF02291411

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

Cascio, W. F. & Aguinis, H. (2011). Applied Psychology in Human Resource Management. Upper Saddle River: New Jersey.

Clark, P. C., & LaHuis, D. M. (2012). An examination of power and type I errors for two differential item functioning indices using the graded response model. *Organizational Research Methods*, *15*(2), 229-246. doi:10.1177/1094428111403815

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Heibattolah, B. & Ferrara, S. (1989). Proceedings from Annual Meeting of the American Educational Research Association, 1989:

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6*, 249-260.

Ironson, G. H., Guion, R. M., & Ostrander, M. (1982). Adverse impact from a psychometric perspective. *Journal of Applied Psychology, 67*, 419-432.

Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2), 93-116.

Linn, R. L. & Drasgow F. (1987). Implications of the Golden Rule settlement for the test construction. *Educational Measurement: Issues and Practice, 6*, 13-17.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388. doi:10.1177/1094428104268027

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

O'Brien, E. & LaHuis, D. M. (2011). Do applicants and incumbents respond to personality items similarly? A comparison of dominance and ideal point response models. *International Journal of Selection and Assessment, 19*, 113-122.

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

Osterlind, S. J, & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

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., 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

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*(1), 45-58. doi:10.1177/014662169001400105

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144. doi:10.1111/j.1745-3984.1990.tb00738.x

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*.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89,* 497-508.

Stocking, M. L., , & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. doi:10.1177/014662168300700208

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. JAP Volume 82

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

Appendix A.

```
setwd ('~/Documents/erindiss2')

nsample=1
for(a in 1:nsample){
  if (a==1) nss=1000
nit=1
for(b in 1:nit){
  if (b==1) nit=10

ndifficulty=1
for(c in 1:ndifficulty){
  if (c==1) ncdiff=0
  if (c==2) ncdiff=0.4
  if (c==3) ncdiff=1.0


ndiscrimination=1
for(d in 1:ndiscrimination){
  if (d==1) ndisc=0
  if (d==2) ndisc=0.3
  if (d==2) ndisc=0.5
  if (d==4) ndisc=0.7

ncanceldiff=1
for (e in 1:ncanceldiff){
  if (e==1) ncan=1
  if (e==2) ncan=-1

for (xx in 1:1000){
  rthetas2 = data.frame(theta=rnorm(nss,0,1))
  fthetas = data.frame(theta=rnorm(nss,0,1))
  rthetas=rbind(rthetas2,fthetas)
  rrandvals=data.frame(v1=fthetas$theta)
  rrandvals$v1 =NULL
  itemparms=NULL
  itemparms$item= 1:10
  itemparms$a=rnorm(10,1.25, .07)
  itemparms$b1= runif(10,-2.5,-1.5)
  itemparms$b2= itemparms$b1+1.2
  itemparms$b3= itemparms$b2+1.2
  itemparms$b4= itemparms$b3+1.2

  itemparms=as.data.frame(itemparms)

  rrandvals$u1=runif(nss,0,1)
  rrandvals$u2=runif(nss,0,1)
  rrandvals$u3=runif(nss,0,1)
  rrandvals$u4=runif(nss,0,1)
  rrandvals$u5=runif(nss,0,1)
  rrandvals$u6=runif(nss,0,1)
  rrandvals$u7=runif(nss,0,1)
  rrandvals$u8=runif(nss,0,1)
  rrandvals$u9=runif(nss,0,1)
```

```
  rrandvals$u10=runif(nss,0,1)

  rrandvals=rbind(rrandvals,rrandvals)

  fitemparms=as.data.frame(itemparms)

  #change focal item parms to manipulate DIF I am guessing first 4
items?

  #Disc parms
  fitemparms[1,2]=fitemparms[1,2]+ndisc
  fitemparms[2,2]=fitemparms[2,2]+ndisc
  fitemparms[3,2]=fitemparms[3,2]+ndisc*ncan
  fitemparms[4,2]=fitemparms[4,2]+ndisc*ncan

  #B1 - not sure what B's we are manipulating but we can take this
code and fix it.
  fitemparms[1,3]=fitemparms[1,3]+ncdiff
  fitemparms[2,3]=fitemparms[2,3]+ncdiff
  fitemparms[3,3]=fitemparms[3,3]+ncdiff*ncan
  fitemparms[4,3]=fitemparms[4,3]+ncdiff*ncan

  fitemparms[1,6]=fitemparms[1,6]+ncdiff
  fitemparms[2,6]=fitemparms[2,6]+ncdiff
  fitemparms[3,6]=fitemparms[3,6]+ncdiff*ncan
  fitemparms[4,6]=fitemparms[4,6]+ncdiff*ncan

  #referent group data generation

  rthetas$brf1_1=(exp(itemparms[1,2]*(rthetas$theta-
itemparms[1,3])))/(1+exp(itemparms[1,2]*(rthetas$theta-
itemparms[1,3])))
  rthetas$brf2_1=(exp(itemparms[1,2]*(rthetas$theta-
itemparms[1,4])))/(1+exp(itemparms[1,2]*(rthetas$theta-
itemparms[1,4])))
  rthetas$brf3_1=(exp(itemparms[1,2]*(rthetas$theta-
itemparms[1,5])))/(1+exp(itemparms[1,2]*(rthetas$theta-
itemparms[1,5])))
  rthetas$brf4_1=(exp(itemparms[1,2]*(rthetas$theta-
itemparms[1,6])))/(1+exp(itemparms[1,2]*(rthetas$theta-
itemparms[1,6])))

  rthetas$brf1_2=(exp(itemparms[2,2]*(rthetas$theta-
itemparms[2,3])))/(1+exp(itemparms[2,2]*(rthetas$theta-
itemparms[2,3])))
  rthetas$brf2_2=(exp(itemparms[2,2]*(rthetas$theta-
itemparms[2,4])))/(1+exp(itemparms[2,2]*(rthetas$theta-
itemparms[2,4])))
  rthetas$brf3_2=(exp(itemparms[2,2]*(rthetas$theta-
itemparms[2,5])))/(1+exp(itemparms[2,2]*(rthetas$theta-
itemparms[2,5])))
  rthetas$brf4_2=(exp(itemparms[2,2]*(rthetas$theta-
itemparms[2,6])))/(1+exp(itemparms[2,2]*(rthetas$theta-
itemparms[2,6])))
```

```
  rthetas$brf1_3=(exp(itemparms[3,2]*(rthetas$theta-
itemparms[3,3])))/(1+exp(itemparms[3,2]*(rthetas$theta-
itemparms[3,3])))
  rthetas$brf2_3=(exp(itemparms[3,2]*(rthetas$theta-
itemparms[3,4])))/(1+exp(itemparms[3,2]*(rthetas$theta-
itemparms[3,4])))
  rthetas$brf3_3=(exp(itemparms[3,2]*(rthetas$theta-
itemparms[3,5])))/(1+exp(itemparms[3,2]*(rthetas$theta-
itemparms[3,5])))
  rthetas$brf4_3=(exp(itemparms[3,2]*(rthetas$theta-
itemparms[3,6])))/(1+exp(itemparms[3,2]*(rthetas$theta-
itemparms[3,6])))

  rthetas$brf1_4=(exp(itemparms[4,2]*(rthetas$theta-
itemparms[4,3])))/(1+exp(itemparms[4,2]*(rthetas$theta-
itemparms[4,3])))
  rthetas$brf2_4=(exp(itemparms[4,2]*(rthetas$theta-
itemparms[4,4])))/(1+exp(itemparms[4,2]*(rthetas$theta-
itemparms[4,4])))
  rthetas$brf3_4=(exp(itemparms[4,2]*(rthetas$theta-
itemparms[4,5])))/(1+exp(itemparms[4,2]*(rthetas$theta-
itemparms[4,5])))
  rthetas$brf4_4=(exp(itemparms[4,2]*(rthetas$theta-
itemparms[4,6])))/(1+exp(itemparms[4,2]*(rthetas$theta-
itemparms[4,6])))


  rthetas$brf1_5=(exp(itemparms[5,2]*(rthetas$theta-
itemparms[5,3])))/(1+exp(itemparms[5,2]*(rthetas$theta-
itemparms[5,3])))
  rthetas$brf2_5=(exp(itemparms[5,2]*(rthetas$theta-
itemparms[5,4])))/(1+exp(itemparms[5,2]*(rthetas$theta-
itemparms[5,4])))
  rthetas$brf3_5=(exp(itemparms[5,2]*(rthetas$theta-
itemparms[5,5])))/(1+exp(itemparms[5,2]*(rthetas$theta-
itemparms[5,5])))
  rthetas$brf4_5=(exp(itemparms[5,2]*(rthetas$theta-
itemparms[5,6])))/(1+exp(itemparms[5,2]*(rthetas$theta-
itemparms[5,6])))

  rthetas$brf1_6=(exp(itemparms[6,2]*(rthetas$theta-
itemparms[6,3])))/(1+exp(itemparms[6,2]*(rthetas$theta-
itemparms[6,3])))
  rthetas$brf2_6=(exp(itemparms[6,2]*(rthetas$theta-
itemparms[6,4])))/(1+exp(itemparms[6,2]*(rthetas$theta-
itemparms[6,4])))
  rthetas$brf3_6=(exp(itemparms[6,2]*(rthetas$theta-
itemparms[6,5])))/(1+exp(itemparms[6,2]*(rthetas$theta-
itemparms[6,5])))
  rthetas$brf4_6=(exp(itemparms[6,2]*(rthetas$theta-
itemparms[6,6])))/(1+exp(itemparms[6,2]*(rthetas$theta-
itemparms[6,6])))
```

```
  rthetas$brf1_7=(exp(itemparms[7,2]*(rthetas$theta-
itemparms[7,3])))/(1+exp(itemparms[7,2]*(rthetas$theta-
itemparms[7,3])))
  rthetas$brf2_7=(exp(itemparms[7,2]*(rthetas$theta-
itemparms[7,4])))/(1+exp(itemparms[7,2]*(rthetas$theta-
itemparms[7,4])))
  rthetas$brf3_7=(exp(itemparms[7,2]*(rthetas$theta-
itemparms[7,5])))/(1+exp(itemparms[7,2]*(rthetas$theta-
itemparms[7,5])))
  rthetas$brf4_7=(exp(itemparms[7,2]*(rthetas$theta-
itemparms[7,6])))/(1+exp(itemparms[7,2]*(rthetas$theta-
itemparms[7,6])))

  rthetas$brf1_8=(exp(itemparms[8,2]*(rthetas$theta-
itemparms[8,3])))/(1+exp(itemparms[8,2]*(rthetas$theta-
itemparms[8,3])))
  rthetas$brf2_8=(exp(itemparms[8,2]*(rthetas$theta-
itemparms[8,4])))/(1+exp(itemparms[8,2]*(rthetas$theta-
itemparms[8,4])))
  rthetas$brf3_8=(exp(itemparms[8,2]*(rthetas$theta-
itemparms[8,5])))/(1+exp(itemparms[8,2]*(rthetas$theta-
itemparms[8,5])))
  rthetas$brf4_8=(exp(itemparms[8,2]*(rthetas$theta-
itemparms[8,6])))/(1+exp(itemparms[8,2]*(rthetas$theta-
itemparms[8,6])))

  rthetas$brf1_9=(exp(itemparms[9,2]*(rthetas$theta-
itemparms[9,3])))/(1+exp(itemparms[9,2]*(rthetas$theta-
itemparms[9,3])))
  rthetas$brf2_9=(exp(itemparms[9,2]*(rthetas$theta-
itemparms[9,4])))/(1+exp(itemparms[9,2]*(rthetas$theta-
itemparms[9,4])))
  rthetas$brf3_9=(exp(itemparms[9,2]*(rthetas$theta-
itemparms[9,5])))/(1+exp(itemparms[9,2]*(rthetas$theta-
itemparms[9,5])))
  rthetas$brf4_9=(exp(itemparms[9,2]*(rthetas$theta-
itemparms[9,6])))/(1+exp(itemparms[9,2]*(rthetas$theta-
itemparms[9,6])))

  rthetas$brf1_10=(exp(itemparms[10,2]*(rthetas$theta-
itemparms[10,3])))/(1+exp(itemparms[10,2]*(rthetas$theta-
itemparms[10,3])))
  rthetas$brf2_10=(exp(itemparms[10,2]*(rthetas$theta-
itemparms[10,4])))/(1+exp(itemparms[10,2]*(rthetas$theta-
itemparms[10,4])))
  rthetas$brf3_10=(exp(itemparms[10,2]*(rthetas$theta-
itemparms[10,5])))/(1+exp(itemparms[10,2]*(rthetas$theta-
itemparms[10,5])))
  rthetas$brf4_10=(exp(itemparms[10,2]*(rthetas$theta-
itemparms[10,6])))/(1+exp(itemparms[10,2]*(rthetas$theta-
itemparms[10,6])))

  rthetas[1001:2000,2]=(exp(fitemparms[1,2]*(rthetas[1001:2000,1]-
fitemparms[1,3])))/(1+exp(fitemparms[1,2]*(rthetas[1001:2000,1]-
fitemparms[1,3])))
```

```
  rthetas[1001:2000,3]=(exp(fitemparms[1,2]*(rthetas[1001:2000,1]-
fitemparms[1,4])))/(1+exp(fitemparms[1,2]*(rthetas[1001:2000,1]-
fitemparms[1,4])))
  rthetas[1001:2000,4]=(exp(fitemparms[1,2]*(rthetas[1001:2000,1]-
fitemparms[1,5])))/(1+exp(fitemparms[1,2]*(rthetas[1001:2000,1]-
fitemparms[1,5])))
  rthetas[1001:2000,5]=(exp(fitemparms[1,2]*(rthetas[1001:2000,1]-
fitemparms[1,6])))/(1+exp(fitemparms[1,2]*(rthetas[1001:2000,1]-
fitemparms[1,6])))

  rthetas[1001:2000,6]=(exp(fitemparms[2,2]*(rthetas[1001:2000,1]-
fitemparms[2,3])))/(1+exp(fitemparms[2,2]*(rthetas[1001:2000,1]-
fitemparms[2,3])))
  rthetas[1001:2000,7]=(exp(fitemparms[2,2]*(rthetas[1001:2000,1]-
fitemparms[2,4])))/(1+exp(fitemparms[2,2]*(rthetas[1001:2000,1]-
fitemparms[2,4])))
  rthetas[1001:2000,8]=(exp(fitemparms[2,2]*(rthetas[1001:2000,1]-
fitemparms[2,5])))/(1+exp(fitemparms[2,2]*(rthetas[1001:2000,1]-
fitemparms[2,5])))
  rthetas[1001:2000,9]=(exp(fitemparms[2,2]*(rthetas[1001:2000,1]-
fitemparms[2,6])))/(1+exp(fitemparms[2,2]*(rthetas[1001:2000,1]-
fitemparms[2,6])))

  rthetas[1001:2000,10]=(exp(fitemparms[3,2]*(rthetas[1001:2000,1]-
fitemparms[3,3])))/(1+exp(fitemparms[3,2]*(rthetas[1001:2000,1]-
fitemparms[3,3])))
  rthetas[1001:2000,11]=(exp(fitemparms[3,2]*(rthetas[1001:2000,1]-
fitemparms[3,4])))/(1+exp(fitemparms[3,2]*(rthetas[1001:2000,1]-
fitemparms[3,4])))
  rthetas[1001:2000,12]=(exp(fitemparms[3,2]*(rthetas[1001:2000,1]-
fitemparms[3,5])))/(1+exp(fitemparms[3,2]*(rthetas[1001:2000,1]-
fitemparms[3,5])))
  rthetas[1001:2000,13]=(exp(fitemparms[3,2]*(rthetas[1001:2000,1]-
fitemparms[3,6])))/(1+exp(fitemparms[3,2]*(rthetas[1001:2000,1]-
fitemparms[3,6])))

  rthetas[1001:2000,14]=(exp(fitemparms[4,2]*(rthetas[1001:2000,1]-
fitemparms[4,3])))/(1+exp(fitemparms[4,2]*(rthetas[1001:2000,1]-
fitemparms[4,3])))
  rthetas[1001:2000,15]=(exp(fitemparms[4,2]*(rthetas[1001:2000,1]-
fitemparms[4,4])))/(1+exp(fitemparms[4,2]*(rthetas[1001:2000,1]-
fitemparms[4,4])))
  rthetas[1001:2000,16]=(exp(fitemparms[4,2]*(rthetas[1001:2000,1]-
fitemparms[4,5])))/(1+exp(fitemparms[4,2]*(rthetas[1001:2000,1]-
fitemparms[4,5])))
  rthetas[1001:2000,17]=(exp(fitemparms[4,2]*(rthetas[1001:2000,1]-
fitemparms[4,6])))/(1+exp(fitemparms[4,2]*(rthetas[1001:2000,1]-
fitemparms[4,6])))


ritemprob = data.frame(theta=rthetas[,1])

  ritemprob$p1_1= 1.0-rthetas$brf1_1
  ritemprob$p2_1= rthetas$brf1_1-rthetas$brf2_1
  ritemprob$p3_1= rthetas$brf2_1-rthetas$brf3_1
```

```
ritemprob$p4_1= rthetas$brf3_1-rthetas$brf4_1
ritemprob$p5_1= rthetas$brf4_1-0

ritemprob$p1_2= 1.0-rthetas$brf1_2
ritemprob$p2_2= rthetas$brf1_2-rthetas$brf2_2
ritemprob$p3_2= rthetas$brf2_2-rthetas$brf3_2
ritemprob$p4_2= rthetas$brf3_2-rthetas$brf4_2
ritemprob$p5_2= rthetas$brf4_2-0


ritemprob$p1_3= 1.0-rthetas$brf1_3
ritemprob$p2_3= rthetas$brf1_3-rthetas$brf2_3
ritemprob$p3_3= rthetas$brf2_3-rthetas$brf3_3
ritemprob$p4_3= rthetas$brf3_3-rthetas$brf4_3
ritemprob$p5_3= rthetas$brf4_3-0

ritemprob$p1_4= 1.0-rthetas$brf1_4
ritemprob$p2_4= rthetas$brf1_4-rthetas$brf2_4
ritemprob$p3_4= rthetas$brf2_4-rthetas$brf3_4
ritemprob$p4_4= rthetas$brf3_4-rthetas$brf4_4
ritemprob$p5_4= rthetas$brf4_4-0

ritemprob$p1_5= 1.0-rthetas$brf1_5
ritemprob$p2_5= rthetas$brf1_5-rthetas$brf2_5
ritemprob$p3_5= rthetas$brf2_5-rthetas$brf3_5
ritemprob$p4_5= rthetas$brf3_5-rthetas$brf4_5
ritemprob$p5_5= rthetas$brf4_5-0

ritemprob$p1_6= 1.0-rthetas$brf1_6
ritemprob$p2_6= rthetas$brf1_6-rthetas$brf2_6
ritemprob$p3_6= rthetas$brf2_6-rthetas$brf3_6
ritemprob$p4_6= rthetas$brf3_6-rthetas$brf4_6
ritemprob$p5_6= rthetas$brf4_6-0


ritemprob$p1_7= 1.0-rthetas$brf1_7
ritemprob$p2_7= rthetas$brf1_7-rthetas$brf2_7
ritemprob$p3_7= rthetas$brf2_7-rthetas$brf3_7
ritemprob$p4_7= rthetas$brf3_7-rthetas$brf4_7
ritemprob$p5_7= rthetas$brf4_7-0

ritemprob$p1_8= 1.0-rthetas$brf1_8
ritemprob$p2_8= rthetas$brf1_8-rthetas$brf2_8
ritemprob$p3_8= rthetas$brf2_8-rthetas$brf3_8
ritemprob$p4_8= rthetas$brf3_8-rthetas$brf4_8
ritemprob$p5_8= rthetas$brf4_8-0


ritemprob$p1_9= 1.0-rthetas$brf1_9
ritemprob$p2_9= rthetas$brf1_9-rthetas$brf2_9
ritemprob$p3_9= rthetas$brf2_9-rthetas$brf3_9
ritemprob$p4_9= rthetas$brf3_9-rthetas$brf4_9
ritemprob$p5_9= rthetas$brf4_9-0

ritemprob$p1_10= 1.0-rthetas$brf1_10
```

```
  ritemprob$p2_10= rthetas$brf1_10-rthetas$brf2_10
  ritemprob$p3_10= rthetas$brf2_10-rthetas$brf3_10
  ritemprob$p4_10= rthetas$brf3_10-rthetas$brf4_10
  ritemprob$p5_10= rthetas$brf4_10-0


  itemdat = data.frame(subid=1:2000)

itemdat[rrandvals[,1]<ritemprob[,2], "item1"]<-1
itemdat[rrandvals[,1]>ritemprob[,2] &
rrandvals[,1]<(ritemprob[,2]+ritemprob[,3]), "item1"]<-2
itemdat[rrandvals[,1]>(ritemprob[,2]+ritemprob[,3]) &
rrandvals[,1]<(ritemprob[,2]+ritemprob[,3]+ritemprob[,4]),
"item1"]<-3
itemdat[rrandvals[,1]>(ritemprob[,2]+ritemprob[,3]+ritemprob[,4]) &
rrandvals[,1]<(ritemprob[,2]+ritemprob[,3]+ritemprob[,4]+ritemprob[,
5]), "item1"]<-4
itemdat[rrandvals[,1]>
(ritemprob[,2]+ritemprob[,3]+ritemprob[,4]+ritemprob[,5]),
"item1"]<-5

itemdat[rrandvals[,2]<ritemprob[,7], "item2"]<-1
itemdat[rrandvals[,2]>ritemprob[,7] &
rrandvals[,2]<(ritemprob[,7]+ritemprob[,8]), "item2"]<-2
itemdat[rrandvals[,2]>(ritemprob[,7]+ritemprob[,8]) &
rrandvals[,2]<(ritemprob[,7]+ritemprob[,8]+ritemprob[,9]),
"item2"]<-3
itemdat[rrandvals[,2]>(ritemprob[,7]+ritemprob[,8]+ritemprob[,9]) &
rrandvals[,2]<(ritemprob[,7]+ritemprob[,8]+ritemprob[,9]+ritemprob[,
10]), "item2"]<-4
itemdat[rrandvals[,2]>
(ritemprob[,7]+ritemprob[,8]+ritemprob[,9]+ritemprob[,10]),
"item2"]<-5


itemdat[rrandvals[,3]<ritemprob[,12], "item3"]<-1
itemdat[rrandvals[,3]>ritemprob[,12] &
rrandvals[,3]<(ritemprob[,12]+ritemprob[,13]), "item3"]<-2
itemdat[rrandvals[,3]>(ritemprob[,12]+ritemprob[,13]) &
rrandvals[,3]<(ritemprob[,12]+ritemprob[,13]+ritemprob[,14]),
"item3"]<-3
itemdat[rrandvals[,3]>(ritemprob[,12]+ritemprob[,13]+ritemprob[,14])
&
rrandvals[,3]<(ritemprob[,12]+ritemprob[,13]+ritemprob[,14]+ritempro
b[,15]), "item3"]<-4
itemdat[rrandvals[,3]>
(ritemprob[,12]+ritemprob[,13]+ritemprob[,14]+ritemprob[,15]),
"item3"]<-5

itemdat[rrandvals[,4]<ritemprob[,17], "item4"]<-1
itemdat[rrandvals[,4]>ritemprob[,17] &
rrandvals[,4]<(ritemprob[,17]+ritemprob[,18]), "item4"]<-2
itemdat[rrandvals[,4]>(ritemprob[,17]+ritemprob[,18]) &
rrandvals[,4]<(ritemprob[,17]+ritemprob[,18]+ritemprob[,19]),
"item4"]<-3
```

```
itemdat[rrandvals[,4]>(ritemprob[,17]+ritemprob[,18]+ritemprob[,19])
&
rrandvals[,4]<(ritemprob[,17]+ritemprob[,18]+ritemprob[,19]+ritempro
b[,20]), "item4"]<-4
itemdat[rrandvals[,4]>
(ritemprob[,17]+ritemprob[,18]+ritemprob[,19]+ritemprob[,20]),
"item4"]<-5

itemdat[rrandvals[,5]<ritemprob[,22], "item5"]<-1
itemdat[rrandvals[,5]>ritemprob[,22] &
rrandvals[,5]<(ritemprob[,22]+ritemprob[,23]), "item5"]<-2
itemdat[rrandvals[,5]>(ritemprob[,22]+ritemprob[,23]) &
rrandvals[,5]<(ritemprob[,22]+ritemprob[,23]+ritemprob[,24]),
"item5"]<-3
itemdat[rrandvals[,5]>(ritemprob[,22]+ritemprob[,23]+ritemprob[,24])
&
rrandvals[,5]<(ritemprob[,22]+ritemprob[,23]+ritemprob[,24]+ritempro
b[,25]), "item5"]<-4
itemdat[rrandvals[,5]>
(ritemprob[,22]+ritemprob[,23]+ritemprob[,24]+ritemprob[,25]),
"item5"]<-5

itemdat[rrandvals[,6]<ritemprob[,27], "item6"]<-1
itemdat[rrandvals[,6]>ritemprob[,27] &
rrandvals[,6]<(ritemprob[,27]+ritemprob[,28]), "item6"]<-2
itemdat[rrandvals[,6]>(ritemprob[,27]+ritemprob[,28]) &
rrandvals[,6]<(ritemprob[,27]+ritemprob[,28]+ritemprob[,29]),
"item6"]<-3
itemdat[rrandvals[,6]>(ritemprob[,27]+ritemprob[,28]+ritemprob[,29])
&
rrandvals[,6]<(ritemprob[,27]+ritemprob[,28]+ritemprob[,29]+ritempro
b[,30]), "item6"]<-4
itemdat[rrandvals[,6]>
(ritemprob[,27]+ritemprob[,28]+ritemprob[,29]+ritemprob[,30]),
"item6"]<-5

itemdat[rrandvals[,7]<ritemprob[,32], "item7"]<-1
itemdat[rrandvals[,7]>ritemprob[,32] &
rrandvals[,7]<(ritemprob[,32]+ritemprob[,33]), "item7"]<-2
itemdat[rrandvals[,7]>(ritemprob[,32]+ritemprob[,33]) &
rrandvals[,7]<(ritemprob[,32]+ritemprob[,33]+ritemprob[,34]),
"item7"]<-3
itemdat[rrandvals[,7]>(ritemprob[,32]+ritemprob[,33]+ritemprob[,34])
&
rrandvals[,7]<(ritemprob[,32]+ritemprob[,33]+ritemprob[,34]+ritempro
b[,35]), "item7"]<-4
itemdat[rrandvals[,7]>
(ritemprob[,32]+ritemprob[,33]+ritemprob[,34]+ritemprob[,35]),
"item7"]<-5

itemdat[rrandvals[,8]<ritemprob[,37], "item8"]<-1
itemdat[rrandvals[,8]>ritemprob[,37] &
rrandvals[,8]<(ritemprob[,37]+ritemprob[,38]), "item8"]<-2
```

```
itemdat[rrandvals[,8]>(ritemprob[,37]+ritemprob[,38]) &
rrandvals[,8]<(ritemprob[,37]+ritemprob[,38]+ritemprob[,39]),
"item8"]<-3
itemdat[rrandvals[,8]>(ritemprob[,37]+ritemprob[,38]+ritemprob[,39])
&
rrandvals[,8]<(ritemprob[,37]+ritemprob[,38]+ritemprob[,39]+ritempro
b[,40]), "item8"]<-4
itemdat[rrandvals[,8]>
(ritemprob[,37]+ritemprob[,38]+ritemprob[,39]+ritemprob[,40]),
"item8"]<-5

itemdat[rrandvals[,9]<ritemprob[,42], "item9"]<-1
itemdat[rrandvals[,9]>ritemprob[,42] &
rrandvals[,9]<(ritemprob[,42]+ritemprob[,43]), "item9"]<-2
itemdat[rrandvals[,9]>(ritemprob[,42]+ritemprob[,43]) &
rrandvals[,9]<(ritemprob[,42]+ritemprob[,43]+ritemprob[,44]),
"item9"]<-3
itemdat[rrandvals[,9]>(ritemprob[,42]+ritemprob[,43]+ritemprob[,44])
&
rrandvals[,9]<(ritemprob[,42]+ritemprob[,43]+ritemprob[,44]+ritempro
b[,45]), "item9"]<-4
itemdat[rrandvals[,9]>
(ritemprob[,42]+ritemprob[,43]+ritemprob[,44]+ritemprob[,45]),
"item9"]<-5

itemdat[rrandvals[,10]<ritemprob[,47], "item10"]<-1
itemdat[rrandvals[,10]>ritemprob[,47] &
rrandvals[,10]<(ritemprob[,47]+ritemprob[,48]), "item10"]<-2
itemdat[rrandvals[,10]>(ritemprob[,47]+ritemprob[,48]) &
rrandvals[,10]<(ritemprob[,47]+ritemprob[,48]+ritemprob[,49]),
"item10"]<-3
itemdat[rrandvals[,10]>(ritemprob[,47]+ritemprob[,48]+ritemprob[,49]
) &
rrandvals[,10]<(ritemprob[,47]+ritemprob[,48]+ritemprob[,49]+ritempr
ob[,50]), "item10"]<-4
itemdat[rrandvals[,10]>
(ritemprob[,47]+ritemprob[,48]+ritemprob[,49]+ritemprob[,50]),
"item10"]<-5

  ritemdat = itemdat[1:1000,]
  fitemdat  =itemdat[1001:2000,]

  ritemfile=paste('rdata','_',nit,
'_',ncdiff,'_',ndisc,'_',ncan,'_', xx,'.csv',sep="")
  fitemfile=paste('fdata','_',nit,
'_',ncdiff,'_',ndisc,'_',ncan,'_', xx,'.csv',sep="")
  write.csv(fitemdat,fitemfile)
  write.csv(ritemdat,ritemfile)

}
}}}}
}
```

Appendix B.

```
obs_ncdif= matrix(nrow = 100,ncol=13)
sm_ncdif = matrix(nrow = 100,ncol=13)
lg_ncdif = matrix(nrow = 100,ncol=13)
setwd ('~/Documents/erindiss2')
counter = 0
library(ltm)
library(DFIT)

for (nn in 1:100){
nsample=1
for(a in 1:nsample){
  if (a==1) nss=2000
    nit=1
  for(b in 1:nit){
    if (b==1) nit=10

    ndifficulty=1
    for(c in 1:ndifficulty){
      if (c==1) ncdiff=0
      if (c==2) ncdiff=0.4
      if (c==3) ncdiff=1.0

      ndiscrimination=1
      for(d in 1:ndiscrimination){
        if (d==2) ndisc=0
        if (d==2) ndisc=0.3
        if (d==1) ndisc=0.5
        if (d==4) ndisc=0.7

        ncanceldiff=1
        for (e in 1:ncanceldiff){
          if (e==1) ncan=1
          if (e==2) ncan=-1

          for (xx in 1:100){

          ritemfile=paste('rdata','_',nit,
'_',ncdiff,'_',ndisc,'_',ncan,'_', xx,'.csv',sep="")
          fitemfile=paste('fdata','_',nit,
'_',ncdiff,'_',ndisc,'_',ncan,'_', xx,'.csv',sep="")
          fitemdat =   read.csv(fitemfile)
          ritemdat=  read.csv(ritemfile)
          ritemdat=ritemdat-1
          fitemdat=fitemdat-1

          ritemgrm
=ltm::grm(ritemdat[,3:12],constrained=F,Hessian=T)
          fitemgrm
=ltm::grm(fitemdat[,3:12],constrained=F,Hessian=T)
          fiobs_ncdifarmt = coef(fitemgrm)
          fiobs_ncdifarm_a = fiobs_ncdifarmt[,5]
```

```
          fiobs_ncdifarm =
cbind(fiobs_ncdifarm_a,fiobs_ncdifarmt[,1:4])
          fitemscore<-
factor.scores(ritemgrm,resp.patterns=NULL,method=c("EAP"))
          fthetas<-fitemscore$score.dat

          riobs_ncdifarmt = coef(ritemgrm)
          riobs_ncdifarm_a = riobs_ncdifarmt[,5]
          riobs_ncdifarm =
cbind(riobs_ncdifarm_a,riobs_ncdifarmt[,1:4])

          common=matrix((c(5,6,7,8,9,10,5,6,7,8,9,10)),ncol=2)
          pars=list(riobs_ncdifarm,fiobs_ncdifarm)
          names(pars)=c('ritem','fitem')
          x<-list(pars,common)
          names(x)=c('pars','common')


          pm<-as.poly.mod(10,model = "grm")
          xpars<-
as.irt.pars(x$pars,x$common,cat=list(rep(5,10),rep(5,10)),poly.mod=l
ist(pm,pm))
          link.out<-plink(xpars,rescale="SL")
          transparm=link.pars(link.out)

          parmlist = (list('focal' = transparm$group1, 'reference' =
transparm$group2))
          obs_ncdif[xx+counter*100,1:10]=Ncdif(parmlist,irtModel =
"grm", focalAbilities = fthetas$z1)
          obs_ncdif[xx+counter*100,11]= ncdiff
          obs_ncdif[xx+counter*100,12]= ndisc
          obs_ncdif[xx+counter*100,13]= ncan

          sm_ncdif[xx+counter*100,1:10]=Ncdif(parmlist,irtModel =
"grm", focalAbilities = fthetas[1:100,13])
          sm_ncdif[xx+counter*100,11]= ncdiff
          sm_ncdif[xx+counter*100,12]= ndisc
          sm_ncdif[xx+counter*100,13]= ncan

          lg_ncdif[xx+counter*100,1:10]=Ncdif(parmlist,irtModel =
"grm", focalAbilities = fthetas[(nrow(fthetas)-
100):nrow(fthetas),13])
          lg_ncdif[xx+counter*100,11]= ncdiff
          lg_ncdif[xx+counter*100,12]= ndisc
          lg_ncdif[xx+counter*100,13]= ncan


        }
  counter = counter + 1

      }}}
}
}}
#obs_ncdif = read.csv('results/lg_ncdif unicorn.csv')
lgcutoffs = read.csv('results/cutoffs100.csv')
```

```
lgcutoffs = lgcutoffs[1:100,-1]
obs_ncdif=as.data.frame(obs_ncdif)
obs_ncdif$i1dif=0
obs_ncdif$i2dif=0
obs_ncdif$i3dif=0
obs_ncdif$i4dif=0
obs_ncdif$i5dif=0
obs_ncdif$i6dif=0
obs_ncdif$i7dif=0
obs_ncdif$i8dif=0
obs_ncdif$i9dif=0
obs_ncdif$i10dif=0

obs_ncdif[obs_ncdif[,1]>quantile(lgcutoffs[,1],.95),'i1dif']<-1
obs_ncdif[obs_ncdif[,2]>quantile(lgcutoffs[,2],.95),'i2dif']<-1
obs_ncdif[obs_ncdif[,3]>quantile(lgcutoffs[,3],.95),'i3dif']<-1
obs_ncdif[obs_ncdif[,4]>quantile(lgcutoffs[,4],.95),'i4dif']<-1
obs_ncdif[obs_ncdif[,5]>quantile(lgcutoffs[,5],.95),'i5dif']<-1
obs_ncdif[obs_ncdif[,6]>quantile(lgcutoffs[,6],.95),'i6dif']<-1
obs_ncdif[obs_ncdif[,7]>quantile(lgcutoffs[,7],.95),'i7dif']<-1
obs_ncdif[obs_ncdif[,8]>quantile(lgcutoffs[,8],.95),'i8dif']<-1
obs_ncdif[obs_ncdif[,9]>quantile(lgcutoffs[,9],.95),'i9dif']<-1
obs_ncdif[obs_ncdif[,10]>quantile(lgcutoffs[,10],.95),'i10dif']<-1


#write.csv(obs_ncdif,'results/obs_ncdif unicorn.csv')
#write.csv(obs_ncdif,'results/sm_ncdif unicorn.csv')
#write.csv(obs_ncdif,'results/lg_ncdif unicorn.csv')

library(plyr)

ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i1dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i2dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i3dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i4dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i5dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i6dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i7dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i8dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i9dif))
ddply(obs_ncdif,.(V11, V12,V13), summarise,mean=mean(i10dif))



mean(obs_ncdif$i7dif)
```

Appendix C.

```
cutoffs = matrix(nrow =1000,ncol=10)
smcutoffs = matrix(nrow =1000,ncol=10)
lgcutoffs = matrix(nrow =1000,ncol=10)

setwd ('~/Documents/erindiss')
library(ltm)
library(DFIT)
nsample=1
for(a in 1:nsample){
  if (a==1) nss=2000
  nit=1
  for(b in 1:nit){
    if (b==1) nit=10
    ndifficulty=1
    for(c in 1:ndifficulty){
      if (c==1) ncdiff=0
      if (c==2) ncdiff=0.4
      if (c==3) ncdiff=1.0

      ndiscrimination=1
      for(d in 1:ndiscrimination){
        if (d==1) ndisc=0
        if (d==2) ndisc=0.3
        if (d==2) ndisc=0.5
        if (d==4) ndisc=0.7

        ncanceldiff=1
        for (e in 1:ncanceldiff){
          if (e==1) ncan=1
          if (e==2) ncan=-1

          for (xx in 1:1000){

          ritemfile=paste('rdata','_',nit,
'_',ncdiff,'_',ndisc,'_',ncan,'_', xx,'.csv',sep="")
          fitemfile=paste('fdata','_',nit,
'_',ncdiff,'_',ndisc,'_',ncan,'_', xx,'.csv',sep="")
          fitemdat = read.csv(fitemfile)
          ritemdat=  read.csv(ritemfile)
          ritemdat=ritemdat-1
          fitemdat=fitemdat-1

          ritemgrm
=ltm::grm(ritemdat[,3:12],constrained=F,Hessian=T)
          fitemgrm
=ltm::grm(fitemdat[,3:12],constrained=F,Hessian=T)
          fitemparmt = coef(fitemgrm)
          fitemparm_a = fitemparmt[,5]
          fitemparm = cbind(fitemparm_a,fitemparmt[,1:4])
          fitemscore<-
factor.scores(ritemgrm,resp.patterns=NULL,method=c("EAP"))
          fthetas<-fitemscore$score.dat
```

```
          ritemparmt = coef(ritemgrm)
          ritemparm_a = ritemparmt[,5]
          ritemparm = cbind(ritemparm_a,ritemparmt[,1:4])
          common=matrix((c(5,6,7,8,9,10,5,6,7,8,9,10)),ncol=2)
          pars=list(ritemparm,fitemparm)
          names(pars)=c('ritem','fitem')
          x<-list(pars,common)
          names(x)=c('pars','common')


          pm<-as.poly.mod(10,model = "grm")
          xpars<-
as.irt.pars(x$pars,x$common,cat=list(rep(5,10),rep(5,10)),poly.mod=l
ist(pm,pm))
          link.out<-plink(xpars,rescale="SL")
          transparm=link.pars(link.out)

          parmlist = (list('focal' = transparm$group1, 'reference' =
transparm$group2))
          cutoffs[xx,]=Ncdif(parmlist,irtModel = "grm",
focalAbilities = fthetas$z1)
          smcutoffs[xx,]=Ncdif(parmlist,irtModel = "grm",
focalAbilities = fthetas[1:100,13])
          lgcutoffs[xx,]= Ncdif(parmlist,irtModel = "grm",
focalAbilities = fthetas[(nrow(fthetas)-100):nrow(fthetas),13])
          }
        }}}}
}

write.csv(cutoffs,'~/Documents/erindiss/results/cutoffs1000.csv')
write.csv(smcutoffs,'~/Documents/erindiss/results/smcutoffs1000.csv'
)
write.csv(lgcutoffs,'~/Documents/erindiss/results/lgcutoffs1000.csv'
)
```