June 2017

The Empirical Selection of Anchor Items Using a Multistage Approach

Brandon Craig
University of South Florida, bdcraig@mail.usf.edu

The Empirical Selection of Anchor Items Using a Multistage Approach

by

Brandon Craig

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Curriculum and Instruction
with an emphasis in
Measurement and Evaluation
Department of Educational and Psychological Studies
College of Education
University of South Florida

Major Professor: Eun Sook Kim, Ph.D.
John Ferron, Ph.D.
Jeffrey Kromrey, Ph.D.
Kristine Hogarty, Ph.D.

Date of Approval:
June 13, 2017

Keywords: DIF, differential item functioning, PROC IRT, Rasch

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The purpose of this study was to determine if using a multistage approach for the empirical selection of anchor items would lead to more accurate DIF detection rates than the anchor selection methods proposed by Kopf, Zeileis, & Strobl (2015b). A simulation study was conducted in which the sample size, percentage of DIF, and balance of DIF were manipulated. The outcomes of interest were true positive rates, false positive rates, familywise false positive rates, anchor contamination rates, and familywise anchor contamination rates. Results showed the proposed multistage methods produced lower anchor contamination rates than the non-multistage methods under some conditions, but there were generally no meaningful differences in true positive and false positive rates.

CHAPTER ONE:

INTRODUCTION

The *Standards for Education and Psychological Testing* describes bias as "construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees" (AERA, APA, & NCME, 1999, p. 76). One source of bias is when an item performs differently for two groups of test takers after controlling for the construct ability of the test takers. This source of bias is labeled differential item functioning (DIF). When a test contains items with DIF, the results from the test may be biased, resulting in inaccurate estimates of the test takers' ability levels and compromising any conclusions that can be inferred from the results of the test (Hidalgo, Galindo-Garre, & Gómez-Benito, 2015; Li & Zumbo, 2009).

Several techniques have been developed or used to detect DIF including Mantel-Haenszel (Holland & Thayer, 1986), logistic regression (Rogers & Swaminathan, 1993), structural equation modeling (Finch, 2005), and item response theory (Wang, 2004). One essential component of these techniques is controlling for the construct ability of the test taker. However, accurately estimating the construct ability can be difficult. Generally, the researcher does not have a priori knowledge of the test taker's ability level and must use the same test being investigated for DIF to also estimate the construct ability of the test taker. This approach to investigating DIF creates a problem. The researcher cannot accurately estimate the construct ability of the examinees if the test contains items with DIF; however, the researcher cannot

accurately identify items with DIF without an accurate estimate of the examinees' construct abilities.

Several researchers have attempted to address this problem by developing techniques to locate DIF-free items to be used as anchor items when testing for DIF (González-Betanzos & Abad, 2012; Khalid & Glas, 2014; Kopf, Zeileis, & Strobl, 2015a; Kopf, Zeileis, & Strobl, 2015b; Meade & Wright, 2012; Shih, Liu, & Wang, 2014; Shih & Wang, 2009; Wang & Shih, 2010; Wang, Shih, & Sun, 2012; Wang, Shih, & Yang, 2009; Woods, 2009). Anchor items are items whose parameters are held constant between groups. Non-anchor items are items whose parameters are allowed to vary between groups. If an anchor is made up entirely of DIF-free items, that anchor can then be used to accurately test non-anchor items for DIF (Wang, 2004). However, locating DIF-free items to use in the anchor can be challenging, especially when there is a high percentage of items with DIF that all favor the same group.

Recently, Kopf et al. (2015a, 2015b) conducted two extensive simulation studies examining different methods used to empirically select anchor items for the purpose of DIF analyses. These methods were evaluated based on the proportions of false positives and true positives. False positives are defined as the proportion of DIF-free items identified as having DIF. True positives are defined as the proportion of DIF items identified as having DIF. The two best performing methods examined by Kopf et al. can be abbreviated as C4-SA(MPT) and IF-SA(MTT).

C4-SA(MPT) uses each item as a single anchor (SA) to preliminarily test every other item for DIF resulting in k-1 test statistics for each item, where k is the number of items on the test. Items are then ranked by the number of times the p-value for the item is above the mean p-value threshold (MPT). Items above MPT the greatest number of times are presumed to be the

most likely DIF-free items and ranked the highest. The four (C4) highest ranking items are then chosen to be the anchor for the DIF analysis.

IF-SA(MTT) also uses SA to preliminarily test every other item for DIF; however, items are ranked by the number of times the absolute test statistic for the item is below the mean test statistic threshold (MTT). The items below MTT the greatest number of times are presumed to be the most likely to be DIF-free and ranked the highest. The highest ranking item is chosen to be the anchor, and all other items are tested for DIF. If the number of items presumed to be DIF-free, defined as the number of items without a significant DIF test, is longer than the current anchor length, the next highest ranking item based on the original ranking is added to the anchor. The DIF test is iteratively repeated and new items added to the anchor until it is not shorter than the number of items presumed to be DIF-free.

While both methods worked better than the other anchor selection methods they were compared to, in terms of false positive and true positive rates they generally did not perform as well as the DIF-free sets of anchor items they were compared to when the test contained 40% of items with DIF favoring a single group. Based on these results, there is a need to develop a method of empirically selecting anchor items that works well even in cases of extreme DIF contamination such as 40% of items with DIF favoring a single group. It is theorized that adding a multistage approach to the two methods Kopf et al. (2015b) found to be most effective will result in more accurate DIF detection in cases of extreme DIF contamination.

The proposed multistage approach consists of the following stages and stopping criteria: (Stage 1) Use the anchor selection method to identify anchors and test all non-anchor items for DIF. During this stage all items are anchor item candidates. (Stage 2) Remove any items showing statistically significant DIF in the previous stage from the pool of anchor item

candidates, rerun the anchor selection method, and use the newly identified anchor to test all non-anchor items for DIF. (Stage *i*) Continue this procedure until the same set of anchor items are identified in two consecutive stages. Use that set of anchor items to conduct a final DIF test on all non-anchor items.

The goal of the multistage approach is to create multiple iterative stages during which the percentage of DIF within the test was reduced in each successive stage. Because the effectiveness of C4-SA(MPT) and IF-SA(MTT) improves as the percentage of DIF decreases, it was hypothesized the final anchor items selected by the multistage approach should be more likely to be DIF-free, which should result in more accurate false positive and true positive rates during the final DIF test.

A simulation study was conducted to compare the multistage approach to C4-SA(MPT) and IF-SA(MTT). The multistage approach in combination with the two prior methods was abbreviated as MS[C4-SA(MPT)] and MS[IF-SA(MTT)]. For comparison purposes, two sets of DIF-free anchor items were also implemented. These four anchor item selection methods and the two DIF-free anchor items were compared by applying them to datasets simulated using the following manipulated variables: three sample sizes for the reference/focal groups (500/500, 750/750, 1000/1000), four percentages of DIF (0%, 10%, 20%, 40%), and two balances of DIF (one-sided, balanced). A 20-item test was generated under the Rasch model with a DIF magnitude of 0.4, mean group difference where the reference group is 1 standard deviation higher than the focal group, and item parameters identical to the parameters used by Kopf et al. (2015b). Note that balanced DIF means that half of the DIF items favor the reference group and half favor the focal group. In the one-sided balance all DIF items favor the reference group. This study was fully crossed, resulting in 21 conditions because the balance of DIF is not applicable

when the test contains zero percent DIF (3x3x2 DIF conditions and three 0% DIF conditions). Four hundred datasets were generated for each condition. All four anchor item selection methods were then applied to the 8,400 datasets, along with two types of DIF-free anchors. The outcomes of interest were the false positive rates, true positive rates, familywise false positive rates, anchor contamination, and familywise anchor contamination.

Summary

Problem

Current methods for empirically selecting anchor items tend not to work well in conditions where a large percentage of DIF items all favor the same group (Kopf et al., 2015a; Kopf et al., 2015b). Under these conditions, items with DIF are sometimes selected as anchor items, which leads to inaccurate DIF detection in terms of increased false positive and decreased true positive rates.

Purpose

The purpose of this study was to determine if adding a multistage approach to the empirical selection of anchor items would lead to more accurate DIF detection. The proposed multistage approach expanded previously researched anchor methods by adding iterative steps to the procedure during which items primarily identified as having DIF are removed from the pool of potential anchor items. The goal was to reduce the percentage of DIF in the pool of potential anchor items which would theoretically lead to an improved rate of selecting DIF-free anchor items. Using only DIF-free items in the anchor has been shown to improve DIF detection accuracy rates when compared to a contaminated anchor (Wang, 2004).

Hypotheses

1. The multistage anchor selection methods will have higher true positive rates, lower false positive rates, lower familywise false positive rates, lower anchor contamination, and lower familywise anchor contamination than the non-multistage methods.
2. The anchor selection methods using IF will have higher true positive rates but also higher false positive rates than anchor selection methods using C4.
3. Familywise false positive rates will be greater than .05 for most, or all, conditions.

Questions

1. Will any of the studied methods result in DIF detection rates equal to the DIF detection rates for the DIF-free anchors for all conditions?
2. Will there be a difference in the anchor contamination rates between the IF and C4 methods?

Definitions of Frequently Used Terms

Definitions of terms frequently used throughout this paper are provided below in alphabetical order.

Anchor Class: A component of the framework for classifying anchor item selection methods which defines the length of the anchor and the overall approach used when identifying anchor items (Kopf et al., 2015a; Kopf et al., 2015b).

Anchor Contamination Rate: Outcome of interested which is calculated by diving the total number of DIF items within a set of anchor items by the total number of items in the anchor.

Anchor Items: Items whose parameters are constrained to be equal between groups when testing for DIF.

Anchor Method: A component of the framework for classifying anchor item selection methods which is defined as the combination of the anchor class and anchor selection strategy (Kopf et al., 2015a; Kopf et al., 2015b).

Anchor Strategy: A component of the framework for classifying anchor item selection methods which determines how anchors are chosen for a particular anchor class (Kopf et al., 2015a; Kopf et al., 2015b).

Average Signed Area (ASAR): Represents the magnitude of DIF for the entire test. Is calculated by summing the signed area (SAR) for each item and dividing by the total number of items (Raju, 1988).

Contaminated Anchor: A set of anchor items that contains one or more items with DIF.

Differential Item Functioning (DIF): A test item performs differently for groups of test takers after controlling for the construct ability.

False Positive Rate: Outcome of interest in the proposed simulation which is calculated by dividing the number of DIF-free items identified as having DIF by the total number of DIF-free items.

Familywise Anchor Contamination Rate: Outcome of interest which is calculated by dividing the number of sets of anchor items which contain one or more DIF items by the total number of sets of anchor items within a condition.

Familywise False Positive Rate: Outcome of interest which is calculated by dividing the number of simulated tests in which at least one DIF-free item was identified as having DIF by the total number of simulated tests within a condition.

Item Difficulty: Refers to how challenging a test item is for a test taker to provide a correct response.

Item Discrimination: Refers to how well an item separates test takers with high ability levels from test takers with low ability levels.

Item Response Theory (IRT): Method for analyzing and calculating test results based on non-linear, latent trait modeling.

Non-Uniform Differential Item Functioning: Type of DIF which occurs when the item discrimination is different between groups.

Rasch Model: A type of IRT model where the item discrimination for each item is constrained to 1.

Signed Area (SAR): Represents the magnitude of DIF for a single item and is equal to the area between the item characteristic curves for two groups (Raju, 1988).

True Positive Rate: Outcome of interest which is calculated by dividing the number of DIF items identified as having DIF by the total number of DIF items

Uniform Differential Item Functioning: Type of DIF which occurs when the item discrimination is the same for each group but the item difficulty is different.

CHAPTER TWO:

LITERATURE REVIEW

Overview

This chapter explains DIF, some of the consequences of DIF in test score interpretations, common techniques used to identify DIF, the role of anchor items, the challenges associated with identifying DIF-free anchor items, methods developed to locate DIF-free anchor items, limitations of those methods, the need for further research, the methods for empirically selecting anchor items being investigated in this study, and selected components of the study designs of reviewed simulation studies focused on the empirical selection of anchor items.

Differential Item Functioning (DIF)

What is DIF?

Differential item functioning (DIF) occurs when a test item performs differently for subgroups of test takers, generally referred to as the reference and focal groups in a two-group DIF analysis, while holding the construct ability constant. The formula representing DIF for a dichotomously scored item, where one equals a correct response and zero equals an incorrect response, is shown in Equation 1 (Penfield & Camilli, 2006):

$$P(Y = 1|\theta, G = R) \neq P(Y = 1|\theta, G = F) \quad (1)$$

Where

$P(Y = 1)$ *is the probablity of a correct response*,

θ *is the construct ability*,

$G\ is\ the\ group\ membership,$

$R\ is\ the\ reference\ group$, and

F $is\ the\ focal\ group.$

A test item represented by the above formula is considered to have DIF because the conditional probability of a correct response is not the same for the reference and focal groups even when holding the construct ability constant between the two groups.

Taking into account the construct ability is an essential component of any DIF analysis. Often there are mean group differences in test scores between subgroups, and these differences may show up at the item level if only the overall probability of a correct response is examined. However, a difference in the overall probability of a correct response is not considered bias as long as that difference is not present once the construct ability is taken into account (Warne, Yoon, & Price, 2014).

DIF is often described as being uniform or non-uniform, depending on whether or not the difference between groups is identified in the item's difficulty or the item's discrimination. Item difficulty refers to how challenging the item is for a test taker to provide a correct response. In general, a test taker has a lower probability, which is dependent on the test taker's ability level, of providing a correct response on a more difficult item than a less difficult item. Item discrimination refers to how well an item separates test takers with high ability levels from test takers with low ability levels.

Uniform DIF, shown in Figure 1, occurs when the item discrimination is the same for the reference and focal groups but the item difficulty is different. In the figure below, the item difficulty can be measured by identifying the ability level at which the test taker has a 50% probability of answering the item correctly. The reference group needs an ability level of 0 to

have a 50% probability of answering the item correctly. The focal group needs an ability level of 0.5 to have a 50% probability of answering the item correctly. This difference in ability levels between groups needed to have the same probability of a correct response is labeled as DIF. Although the item difficulty differs between groups in this example, the item discrimination, represented by the slope of the line, is the same, meaning this item displays uniform DIF.

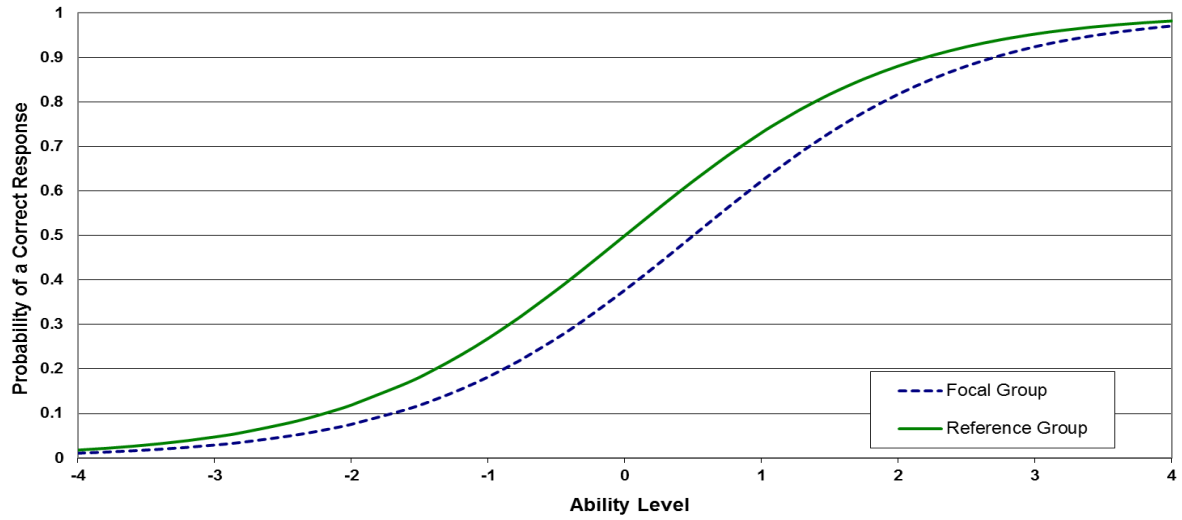

Figure 1. Uniform differential item functioning for a dichotomously scored item.

Non-uniform DIF, shown in Figure 2, occurs when the item discrimination is different between the two groups. The difference in item discrimination is represented by the difference in the slope of the lines for the two groups. Additionally, although the ability level needed to have a 50% probability of a correct response is the same in this example, under non-uniform DIF the difficulty level may also differ between groups.

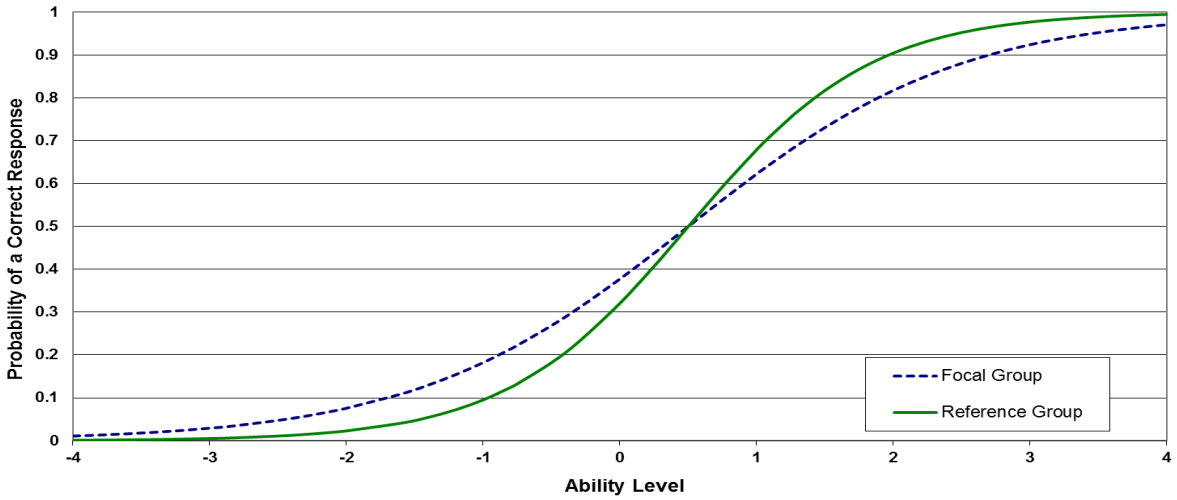

Figure 2. Non-uniform differential item functioning for a dichotomously scored item.

DIF can also occur in ordinal response items such as a Likert scale. Uniform DIF for a four-level ordinal response item is displayed in Figure 3. In this example, the slopes for each level of the response are the same between groups, but the difficulty level differs.

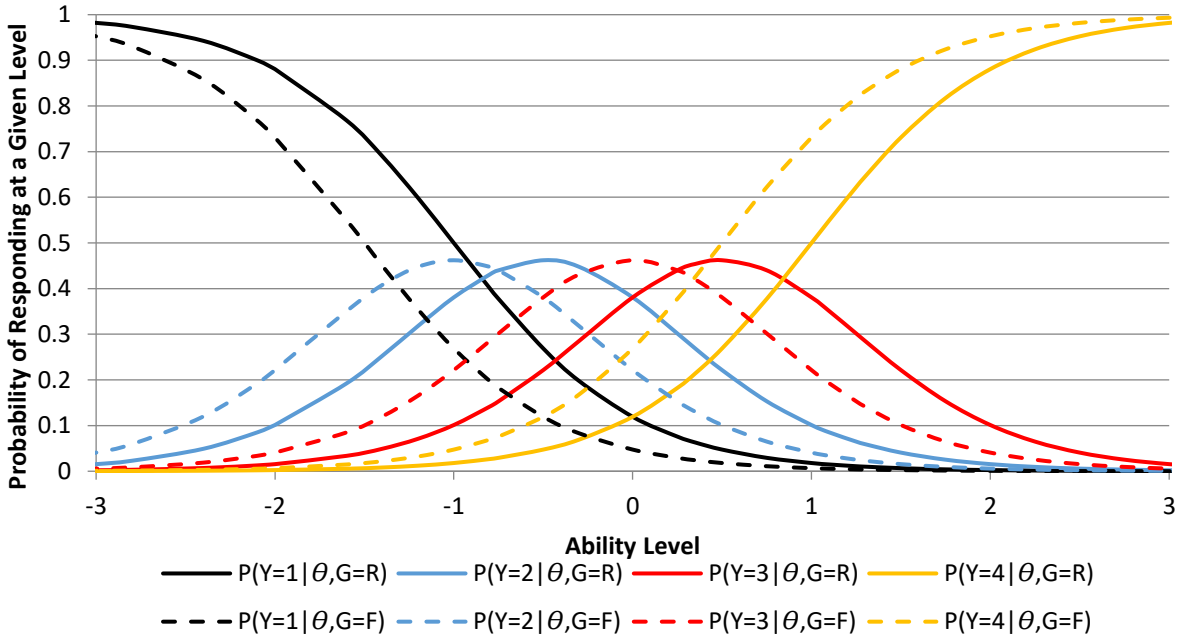

Figure 3. Uniform differential item functioning for a four-level ordinal response item.

Note. θ represents the ability level, G=R represents reference group membership, G=F represents focal group membership.

Why is DIF important?

The presence of DIF is a concern across a variety of subgroups including gender (Murray, Booth, & McKenzie, 2015; Steinmayr, Bergold, Margraf-Stiksrud, & Freund, 2015), race (Roth, Dilworth-Anderson, Jin Huang, Gross, & Gitlin, 2015), English language learner status (Fidalgo, Alavi, & Amirian, 2014), and cognitive status (Fieo et al., 2015). Additionally, the widespread and high stakes use of tests, which include employment decisions (Lievens & Patterson, 2011), college admissions (Shaffer & McCabe, 2013), psychological measurement (Ready & Veague, 2014), teacher and school evaluation (Croft, Roberts, & Stenhouse, 2016), and professional licensure (Jarl, Heinemann, Linder, & Hermansson, 2015), cause DIF to be a major concern for all stakeholders concerned with the results of the test. Because of these concerns, it is imperative that the test accurately measure the target construct; however, when DIF is present on a test, it can impact the results, which may lead to erroneous conclusions (Hidalgo et al., 2015).

Several simulation studies have examined the impact of DIF on overall test results. Li & Zumbo (2009) examined how Type 1 error rates and effect sizes were impacted under a variety of conditions. They found that when all DIF items favored the same group, the Type 1 error rates and the effect sizes between groups were inflated. The inflation got worse as the magnitude of DIF increased. When DIF was balanced between groups or when there were very few items with DIF on test and those items only had small DIF, that Type 1 error rates and effect sizes were close to or within acceptable levels.

Hidalgo et al. (2015) conducted a simulation study in which they examined how the presence of DIF impacts test score interpretations based on cut scores. They simulated a 20-item test using the Rasch model, a constant DIF magnitude of 0.8, and equal mean ability scores

between groups. They varied the sample size, the percentage of items with DIF on the test, and the cut scores. The researchers determined that as the percentage of DIF on the test increased there was greater misclassification on test score interpretations made based on cut scores, and there were greater differences in mean scores between groups.

The negative impact of DIF on test score interpretations has been found in empirical studies as well. Steinmayr et al. (2015) examined gender differences on a general knowledge test given to German high school students. The researchers found that although males scored higher on the test than females, many items contained DIF that favored males. When those items were removed, the overall mean difference in scores between males and females was reduced. Another study examined a random selection of 15,000 Korean test takers. This study determined that regardless of whether an item displaying DIF favors the reference or focal group, the overall score may be impacted (Pae & Park, 2006). An analysis of SAT results showed that removing an item displaying DIF had the largest impact on the total score of the subgroup that was most disadvantaged by the item (Zhang, Dorans, & Matthews-López, 2005). The common findings in all of these studies is that the presence of DIF has a negative impact in the overall score of test takers, which can impact inferences made based on those scores.

When is DIF important?

Another important aspect of DIF is the magnitude of the effect. Like all tests of statistical significance, the statistical tests used to detect DIF are influenced by sample size, which may lead to a situation where although an item displays statistically significant DIF, the magnitude of the DIF is not large enough to be practically significant. Examining the area between the item response functions, like the ones shown in Figures 1 and 2, is one method of determining the effect size. For an item displaying DIF, the signed area (SAR) between the item characteristic

curves for the reference and focal groups can be represented using Equation 2 for the 3-PL IRT model (Raju, 1988):

$$SAR = (1 - c)(b_F - b_R) \quad (2)$$

Where

c *is the psuedo guessing parameter,*

b_R *is the difficulty parameter for the reference group,* and

b_F *is the difficulty parameter for the focal group.*

Although SAR for an individual item can have an impact on the overall test score, an even more important consideration is the average signed area (ASAR) shown in Equation 3 (Wang & Su, 2004).

$$ASAR = \Sigma_{i=1}^{k} \frac{SAR_i}{k} \quad (3)$$

Where

SAR_i *is the signed area for item* i*, and*

k *is the number of items on the test.*

When ASAR is 0, either the test contains no DIF items, or there are items with DIF but those items cancel each other out. For example, if one item has an SAR of -1 and a second item has an SAR of 1, those items will cancel each other out so that the overall score of the test is neither biased against the reference or the focal group even though individual items display DIF. However, in empirical studies of DIF most items usually favor the reference group, and the further away ASAR is from zero, the greater the impact DIF will have on the overall score (Wang & Su, 2004).

DIF from a Test Developer's Prospective

While the presence of DIF items is a concern due to the negative impact DIF can have on the interpretation of test scores, there is also a danger in the over-identification of DIF items, especially from the perspective of a test developer working in K-12 public education.

When an item has DIF, one of the most common methods of dealing with that item is to remove the item from the assessment (Cho, Sun, & Lee, 2016). In some situations an item may be revised so that it no longer has DIF; unfortunately, it is often difficult to determine the cause of DIF within an item. Recent research has begun to focus on identifying the causes of DIF (Balluerka, Plewis, Gorostiaga, & Padilla, 2013; Benitez, Padilla, Montesinos, & Sireci 2016; Huang, & Sheeran, 2011). However, there is not any established consensus as to the causes of DIF for different subgroups, and efforts to revise items are not guaranteed to be successful. Allalouf (2003) used a team of experts to revise items identified as having DIF. After administering the revised items, only 32% of the items had a significant decrease in the amount of DIF, and 24% of the items had an increase in DIF. These results may not be acceptable to many test developers, and their best option may be to remove the item from the test as well as future use. However, throwing out test items is not desirable given the time and cost of developing items.

Item development procedures include creating test item specifications, recruiting and training qualified item writers, writing the items, review of the items for alignment, bias, and content, field testing, and psychometric analysis (Florida Department of Education, 2016). This process is time consuming and expensive. Over the four years from 2010-2014, the Florida Department of Education distributed nearly 20 million dollars of federal grant money for the development of test items for 145 courses, which equals nearly $140,000 for item development

per course. These 145 represent only a small portion of the nearly 2,000 courses that exist in Florida's K-12 education system (Florida Department of Education, 2014; U.S. Department of Education, 2011). Per-item-cost development estimates range from range $1,500 to $2,500 (Rudner, 2007). These costs for item development are staggering compared to the budgets allotted for public school students. According to Darling-Hammond and Adamson (2013) states spend an average of $10,000 per student, but only 20 dollars of that amount is budgeted for assessment. Much of that 20 dollars is spent on assessment procedures such as administration and scoring, leaving less for item development. For test developers, especially those working in public school districts, it is very costly to remove an item from an item bank; therefore, it is important to minimize the items that are incorrectly identified as having DIF.

Detecting DIF

Many techniques have been developed to detect DIF. These techniques can be divided into two broad categories: observed score methods and latent ability methods. Observed score methods rely on a measure such as the number of items answered correctly to estimate the construct ability of the test taker and include Mantel Haenszel and logistic regression. In contrast, latent ability methods use a latent variable to estimate the construct ability of the test taker and include structural equation modeling techniques such as the multiple indicators, multiple causes model as well as item response theory-based techniques.

Mantel-Haenszel

Mantel-Haenszel (MH) is one of the most common methods used in DIF analyses (Rogers & Swaminathan, 1993). As shown in Table 1, MH for dichotomous items is calculated by creating *k* 2 x 2 contingency tables for each studied item, where *k* is the number of levels of the construct ability (Holland & Thayer, 1986). The construct ability is usually defined by total

score and *k* is usually equal to the total number of items on the test. A weighted average of the odds ratios for each of the contingency tables is then calculated to detect DIF. When the reference group equals 0 and the focal group equals 1, an odds ratio greater than 1 indicates that the item favors the focal group, and an odds ratio less than 1 indicates that the item favors the reference group.

Table 1
Contingency table for detecting differential item functioning in a dichotomous item using Mantel-Haenszel

	Item Response Level		
Group	$y = 1$	$y = 0$	Total
Reference	n_{R1k}	n_{R0k}	n_{R+k}
Focal	n_{F1k}	n_{F0k}	n_{F+k}
Total	n_{+1k}	n_{+0k}	n_{++k}

Note. *k* equals the level of the construct ability.

MH can be extended to test for DIF in polytomously scored items by using the generalized Mantel-Haenszel (GMH) procedure. As shown in Table 2, GMH is calculated by creating k 2 x *T* contingency tables for each studied item, where *k* is the number of levels of the construct ability and *T* is the number of response options for the studied item (Michaelides, 2008; Zwick, Donoghue, & Grima, 1993).

Table 2
Contingency table for detecting differential item functioning in a polytomous item using generalized Mantel-Haenszel

	Item Response Level					
Group	y_1	y_2	y_3		y_T	Total
Reference	n_{R1k}	n_{R2k}	n_{R3k}	…	n_{RTk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k}	…	n_{FTk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	…	n_{+Tk}	n_{++k}

Note. *k* equals the level of the construct ability.

To test the null hypothesis that there is no association between group membership and responses, Mantel's chi-square statistic is calculated as shown in Equation 4 (Michaelides, 2008):

$$Mantel's\ \chi^2 = \frac{(\sum_J \sum_T n_{FTk} y_T - \sum_J \frac{n_{F+k}}{n_{++k}} \sum_T n_{+Tk} y_T)^2}{\sum_J VAR(\sum_T n_{FTk} y_T)} \quad (4)$$

MH has been shown to work well in detecting uniform DIF when the sample size was greater than 500, the percent of items with DIF was fewer than 20%, the data fit the Rasch model, and when a scale purification procedure was incorporated (Guilera, Gomez-Benito, Hidalgo, & Sanchez-Meca, 2013). However, MH is not sensitive to non-uniform DIF, and a method such as logistic regression is better for detecting non-uniform DIF (Rogers & Swaminathan, 1993).

Logistic Regression

Logistic Regression (LR), like MH, generally uses a measure such as the total number of items answered correctly to define the construct ability (Rogers & Swaminathan, 1993). For items with a dichotomous outcome, three separate binary logistic regression models, as shown in Equations 5, 6, and 7, are created for each item. The change in the -2 log likelihood (-2LL) statistic between models is then calculated to determine statistical significance. The outcome statistic follows a chi-square distribution with one degree of freedom. A change in -2LL between Equations 5 and 6 greater than 3.84 indicates statistically significant uniform DIF. A change in -2LL between Equations 6 and 7 greater than 3.84 indicates statistically significant non-uniform DIF.

$$\ln[\frac{P(Y_i = 1|G, X)}{1 - P(Y_i = 1|G, X)}] = \beta_0 + \beta_1 X \quad (5)$$

$$\ln[\frac{P(Y_i = 1|G,X)}{1 - P(Y_i = 1|G,X)}] = \beta_0 + \beta_1 X + \beta_2 G \quad (6)$$

$$\ln[\frac{P(Y_i = 1|G,X)}{1 - P(Y_i = 1|G,X)}] = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG \quad (7)$$

Where

$P(Y_i = 1)$ *is the probability of a correct response for item i,*

X *is the construct ability*, and

G *is the group membership.*

Logistic regression (LR) has been shown to be as powerful as MH in detecting uniform DIF but can also be used to detect non-uniform DIF (Rogers & Swaminathan, 1993). Additionally, the model can be extended to test for DIF in polytomously scored items (Apinyapibal, Lawthong, & Kanjanawasee, 2015). Ordinal logistic regression can be used to test for polytomously scored items using Equations 8, 9, and 10 (Jafari, 2014):

$$\ln[\frac{P(Y_i \leq k|G,X)}{1 - P(Y_i > k|G,X)}] = \beta_0 + \beta_1 X \quad (8)$$

$$\ln[\frac{P(Y_i \leq k|G,X)}{1 - P(Y_i > k|G,X)}] = \beta_0 + \beta_1 X + \beta_2 G \quad (9)$$

$$\ln[\frac{P(Y_i \leq k|G,X)}{1 - P(Y_i > k|G,X)}] = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG \quad (10)$$

Where

$P(Y_i \leq k)$ *is the cumulative probability of a k response for item i,*

X *is the construct ability, and*

G *is the group membership.*

However LR, like MH, is criticized because it utilizes the observed score of the test taker to estimate the construct ability rather than the latent score (Millsap & Everson, 1993). Structural

equation modeling and item response theory both make use of the latent ability estimate for DIF analysis.

Multiple Indicators, Multiple Causes Models

Structural equation modeling (SEM) techniques rely on a latent variable to estimate the construct ability. SEM includes confirmatory factor analysis (CFA) and multiple indicators multiple causes (MIMIC) models. CFA models have been shown to be equivalent to two parameter item response theory models (Takane & Leeuw, 1987). Similarly, MIMIC models have been shown to be effective at detecting DIF when the data fit two parameter item response theory models (Finch, 2005).

The MIMIC model is a type of structural equation model that can be divided into a measurement and structural component (Joreskog & Goldberger, 1975). The measurement component can be represented by Equation 11 which models the continuous latent response, y_i^*, for item *i* (Wang et al., 2009). Because y_i^* cannot be directly measured, it is estimated by the observed response for item i, y_i, using Equation 12, when y_i is dichotomous.

$$y_i^* = \lambda_i \theta + \beta_i G + \varepsilon_i \quad (11)$$

Where

y_i^* *is the continuous latent response for item i,*

θ *is the latent ability variable,*

G *is the grouping variable,*

λ_i *is the factor loading,*

β_i *is the effect of the grouping variable on* y_i^**, and*

ε_i *is the error with a standard normal distribution.*

$$y_i = \begin{cases} 1, & if\ y_i^* > \tau_i \\ 0, & otherwise \end{cases} \quad (12)$$

Where

τ_i *is the threshold parameter for item* i.

The structural component is represented by Equation 13:

$$\theta = \gamma G + \zeta \quad (13)$$

Where

γ *is the regression coefficient for the impact of* G *on* θ, *and*

ζ *is the residual.*

The variable β_i is used to detect DIF. If $\beta_i = 0$ there is no uniform DIF in item *i*. If $\beta_i \neq 0$, then uniform DIF is present in item *i*.

MIMIC models are easily expanded to test for polytomous response outcomes such as a Likert scale by making the following adjustment to y_i shown in Equation 14 (Wang & Shih, 2010):

$$y_i = \begin{cases} 0, & if\ y_i^* \leq \tau_{i1} \\ 1, & if\ \tau_{i1} < y_i^* \leq \tau_{i2} \\ . & . \\ . & . \\ . & . \\ J, & if\ y_i^* > \tau_{iJ} \end{cases} \quad (14)$$

Where

τ_{iJ} *is the threshold for endorsing item* i *at level* J.

Item Response Theory

Like SEM, item response theory (IRT) uses a latent variable to estimate the construct ability rather than an observed score. Three of the most common IRT modes are the one

parameter (1PL), two parameter (2PL), and three parameter (3PL) which are represented using Equation 15 (Özdemir, 2015):

$$p(Y_i = 1|\theta) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_i - b_i)]}{1+\exp[a_i(\theta_i - b_i)]} \quad (15)$$

where

$p(Y_i = 1|\theta)$ *is the probablity of a correct response on item i given theta,*

θ *is the latent ability,*

c_i *is the psuedo guessing parameter,*

a_i *is the item discrimination,* and

b_i *is the item difficulty.*

For a 1PL model a_i is held constant for all items, c_i is set to zero, and only b_i is estimated. For a 2PL model c_i is set to zero and both a_i and b_i are estimated.

To test for DIF in polytomous items, a graded response model (GRM) may be used (Cohen, Kim, & Baker, 1993). The GRM estimates the cumulative probability of endorsing an item at or above a given threshold as shown in Equation 16:

$$p(Y_i \geq j|\theta) = \frac{\exp[a_i(\theta_i - b_{ij})]}{1 + \exp[a_i(\theta_i - b_{ij})]} \quad (16)$$

Where

$p(Y_i \geq j|\theta)$ *is the probablity of a correct rsponse on item i at or above j given* θ,

θ *is the latent ability,*

j *is the observed score for item i,*

a_i*is the item discrimination,* and

b_{ij} *is the item difficulty at level* j.

There are multiple approaches for using IRT to test for DIF including Lord's chi-square, differential functioning of items and tests, and the likelihood ratio test (Tay, Meade, & Cao, 2015). Several studies have confirmed the effectiveness of using the IRT likelihood ratio (IRT-LR) approach for the detection of DIF (Atar & Kamata, 2011; Kabasakal, Arsan, Gök, & Kelecioğlu, 2014; Pei & Li, 2010). IRT-LR examines the differences in model fit between the constrained and less constrained model using Equation 17:

$$G^2 = 2 \ln\left(\frac{L_c}{L_f}\right) \tag{17}$$

Where

G^2 *is the test statistic with an approximate chi square distribution,*

L *is the likelihood function,*

c *is the contrained model*, and

f *is the less constrained model.*

In the constrained model, the item parameters for the studied item are constrained equal between groups. In the less constrained model, the item parameters for the studied item are allowed to vary between groups. However, IRT-LR is computationally intensive since it requires two models to test an item for DIF. The Wald test can also be used to test an item for DIF but only requires one model in which the item parameters for the studied item are allowed to vary. The Wald test for uniform DIF within the IRT framework is shown in Equation 18 (Kopf et al., 2015b):

$$W_i = \frac{b_{iG=R} - b_{iG=F}}{\sqrt{Var(b_{iG=R}) + Var(b_{iG=F})}} \tag{18}$$

Where

W_i *is the Wald test statistic,*

b_{iG} *is the item difficulty for item i and group G*, and

$Var(b_{iG})$ *is the variance of item difficulty for item i and group G*.

Anchor Items

For the purpose of this study, anchor items are defined as items whose parameters are constrained to be equal between groups, while the parameters for non-anchor items are allowed to vary between groups. The functioning of anchor items can be shown using the formula for the Rasch model, shown in Equation 19, which is a special case of the 1PL model where the *a* parameter is constrained to be 1 for all items and all groups (Kopf et al., 2015b).

$$p(Y_{iG} = 1|\theta_G) = \frac{\exp[(\theta_i - b_{iG})]}{1+\exp[(\theta_i - b_{iG})]} \quad (19)$$

Where

$p(Y_{iG} = 1|\theta_G)$ *is the probablity of a correct response on item i given theta for group G*,

θ *is the latent ability*, and

b_{iG} *is the item difficulty for item i and group G*.

For anchor items, the *b* parameter estimate for item *i* for the focal group is equal to the *b* parameter estimate for item *i* for the reference group, $b_{iG=F} = b_{iG=R}$. Conversely, for non-anchor items, $b_{iG=F} \neq b_{iG=R}$.

A commonly used technique for identifying DIF is to use all other items (AO), other than the item being studied for DIF, as the anchor. AO has been shown to work adequately when few items have DIF or when the overall DIF on the test is balanced between the reference and the focal group. However, when the percentage of DIF is large or the direction of DIF is not balanced, AO has been shown to have suboptimal false positive and true positive rates (Wang et al., 2009; Woods, 2009).

The reason for these suboptimal rates in certain conditions is that AO will include items with DIF in the anchor any time any item other than the studied item contains DIF. When items with DIF are included in the anchor, labeled as a contaminated anchor, the ability level estimates are biased, which leads to inaccurate parameter estimates (Wang, 2004). Ultimately, contaminated anchors used in a DIF analysis lead to higher false positive and lower true positive rates (Wang et al., 2009). As a result of this problem, many researchers have begun exploring methods to empirically identify DIF-free items that can be used as anchor items for DIF analysis.

Framework

A framework for classifying the numerous methods examined in published literature for empirically selecting anchor items was proposed by Kopf et al. (2015a, 2015b) and slightly adapted for this study. This framework consists of the anchor class, anchor selection strategy, and anchor method.

Anchor class

The anchor class defines the length of the anchor and the overall approach used when identifying anchor items. Generally, the anchor class can be categorized as either a constant anchor length or an iterative scale purification. In the reviewed literature, the constant anchor methods seek to identify a predetermined number of anchor items such as one, four, or 20% of the total number of items (Woods, 2009; Shih & Wang, 2009). The iterative scale purification classes seek to identify anchors without assuming a predefined length but rather relying on a stopping criterion to determine when the desired number of anchors have been identified. Examples of the iterative class include the scale purification procedure examined by Wang et al. (2009) in which an iterative procedure is used to identify anchor items. This procedure stops when all non-anchor items display statistically significant DIF and all anchor items do not

display statistically significant DIF. A complete list of anchor classes examined in the current literature is included in Table 3.

Table 3
Anchor class abbreviations and descriptions

Abbreviation	Anchor Class	Description
C*i*	Constant	The anchor is a predefined, constant length equal to *i*.
I	Iterative	Items are iteratively added and removed from the anchor based on the anchor selection strategy.
IB	Iterative Backward	Items are iteratively removed from the anchor based on the anchor selection strategy.
IF	Iterative Forward	Items are iteratively added to the anchor based on the anchor selection strategy.

Anchor selection strategy

The anchor selection strategy determines how anchors are chosen for a particular anchor class. Anchor selection strategies used in the reviewed literature are shown in Table 4. These strategies may rank candidate anchor items using statistics such as the number of statistically significant DIF tests (Kopf et al., 2015a) or the mean absolute DIF index (Shih & Wang, 2009). Alternatively, the anchor selection strategy may add and remove items to and from the anchor using a statistic such as a significant DIF test (Wang et al., 2009). The calculation of these test statistics can occur either using all other items (AO) as the anchor or using each item as a single anchor (SA), respectively labeled as Type (I) and Type (II) by Kopf et al. (2015a, 2015b). However, as Type (I) and Type (II) are less intuitively descriptive than AO and SA, the latter two labels are used to describe these two approaches. Procedures using AO obtain one test statistic per studied item, while procedures using SA obtain $k - 1$ test statistics per studied item, where k is the number of items on the test.

Table 4

Anchor selection strategies and descriptions

Anchor Selection Strategy	Description
AO(LAT)	lowest absolute test statistic
AO(LES)	lowest effect size
AO(LM)	Lagrange multiplier test statistic
AO(MaxA)	largest difficulty (a) parameter
AO(SIBL)	simultaneous linear item bias test
AO(SIBN)	simultaneous non-linear item bias test
AO(ST)	significant DIF test
SA(MADI)	mean absolute DIF index
SA(MP)	mean p-value
SA(MPT)	number of times above the mean p-value threshold
SA(MT)	mean test statistic
SA(MTT)	number of times below mean test statistic threshold
SA(NST)	number of significant tests

Note. AO uses all other items as anchors. SA uses each item as a single anchor resulting in *k*-1 tests for each item, where *k* is the number of items on the test. Candidate anchor items are added or removed from the anchor based on the anchor selection strategy.

Anchor method

The anchor method is defined as the combination of the anchor class and anchor selection strategy. For example, the constant four anchor class (C4) can be combined with an AO selection strategy based on the lowest absolute test statistic (LAT) to create the C4-AO(LAT) method. Alternatively, C4 can be combined with a SA selection strategy also based on LAT to create the C4-SA(LAT) method. The choice of test statistic can also be changed. Instead of using LAT, a researcher may wish to base anchor selection on the number of significant tests (NST) and create the C4-SA(NST) method.

Previously Studied Methods

Iterative scale purification class

Among the iterative scale purification class, I-AO(ST) has been included in the greatest number of studies (Gonzalez-Betanzos & Abad, 2011; Shih et al., 2014; Wang et al., 2009; Wang et al., 2012; Wang & Shih, 2010). I-AO(ST) is conducted using the following iterations: (1) Use AO to test every item for DIF. (2) Retest every item for DIF using AO but exclude from the anchor items any item showing a statistically significant DIF test (ST) in the first iteration. (3) Retest every item for DIF using AO but exclude from the anchor items any item showing a statistically significant DIF test (ST) in the second iteration. Any items not showing statistically significant DIF in this step that showed statistically significant DIF in the previous step are added back to the anchor. (4) This procedure stops when the same items are displaying statistically significant DIF in two consecutive iterations (Wang et al., 2009).

The iterative scale purification class also includes methods that are either strictly backward (IB) or forward (IF) which were explored by Kopf et al. (2015a, 2015b) in two published articles. An example of an IB method is IB-AO(ST) which removed items from the anchor based on ST when using AO. IB-AO(ST) differed from I-AO(ST) because IB-AO(ST) only removed items from the anchor while I-AO(ST) both removed from and added items to the anchor.

The IF methods studied used both AO and SA in the selection strategy. For example, IF-AO(LAT) ranked items based on the lowest absolute test (LAT) statistic calculated using AO, where the highest ranked item had the smallest absolute test statistic and presumably was the most likely item to be DIF-free. After each new item was added to the anchor, the remaining non-anchor items were tested for DIF using the current anchor. If the number of items not

showing statistically significant DIF was greater than the number of items in the current anchor, the next highest ranked item based on the original ranking was added to the anchor. IF-SA(NST) was similar to the previously described method except that items were ranked based on the number of statistically significant DIF tests (NST) when using SA. Items with the fewest number of statistically significant DIF tests were ranked the highest.

In terms of false positive and true positive rates, the IF anchor class was shown to work better than the IB anchor class, and the SA anchor selection strategy was shown to work better than the AO anchor selection strategy when the test had a large percentage of DIF items favoring the same group (Kopf et al., 2015a). The reason for these findings was likely due to the fact that both IB and AO began by using all other items as anchors, which produced biased results when there was a high percentage of DIF items favoring the same group. However, the authors did find that the performance of IF-SA(NST) was largely influenced by sample size, which was attributed to the fact that statistical significance is also influenced by sample size.

Kopf et al. (2015b) further explored the IF anchor class by researching IF-SA(MP), IF-SA(MPT), IF-SA(MT), IF-SA(MTT), and IF-SA(NST), which incorporated the mean p-value (MP), number of times an item was above the mean p-value threshold (MPT), mean absolute test statistic (MT), number of times an item was below the mean test statistic threshold (MTT), and the number of significant tests (NST) into the anchor selection strategy. The authors were attempting to find an anchor selection strategy that would not be impacted by sample size the way NST was. The new methods worked better than NST, with IF-SA(MTT) working the best in terms of low false positive and high true positive rates. However, IF-SA(MTT) still had false positive rates greater than .05 when the percentage of DIF was 40% and all DIF items favored the reference group.

Constant anchor length class

Many studies have explored different methods within the constant anchor length class (Gonzalez-Betanzos & Abad, 2011; Kopf et al., 2015a; Kopf et al., 2015b; Meade & Wright, 2012; Shih et al., 2014; Shih & Wang, 2009; Wang et al., 2012; Wang & Shih, 2010; Woods, 2009). All of the methods within this class rely on ranking items using the anchor selection strategy and then selecting *i* items to use as the final anchors, where *i* is defined by the anchor class. The variations of these methods have explored the number of items in the anchor, the use of AO and SA to rank items, and the test statistic used to rank these items.

Anchor length

Generally, longer anchors lead to higher true positive and lower false positive rates when the anchor is pure (Shih & Wang, 2009). In their study that examined anchor lengths of one, two, four, and ten items, Shih and Wang (2009) concluded that an anchor length of four is long enough to ensure adequate true positive and false positive rates. Similarly, Meade and Wright (2012) examined anchor lengths of one, three, and five items and concluded that an anchor length of five produced acceptable results. However, both sets of researchers used very liberal criteria for defining acceptable false positive rates. Shih and Wang (2009) reported false positive rates in the range of .03 to .08 with a standard deviation of .01 to .04. Meade and Wright (2012) reported false positive rates with a five item anchor of .074, .075, and .061.

AO vs. SA

When ranking items to select the desired number of anchors within the constant anchor class, some studies have exclusively used AO (Woods, 2009; Wang et al., 2012; Meade & Wright, 2012; Shih et al., 2014). Other studies have exclusively used SA (Shih & Wang, 2009; Wang & Shih, 2010). In the currently published literature on the empirical selection of anchor

items, only Kopf et al. (2015a, 2015b) has examined both SA and AO in the same study, allowing for a direct comparison of the two selection strategies. The researchers concluded that when DIF was balanced, AO outperformed SA, although SA was generally acceptable. In conditions with unbalanced DIF, SA outperformed AO. This finding is likely due to the fact that AO produces biased results when a test contains many DIF items favoring a single group (Wang, 2004).

Test statistic used

Several test statistics used to rank items within the constant anchor length class have also been explored in the published research. However, few studies have directly compared these methods, which is necessary in order to determine which test statistic produces the most desirable results. Meade and Wright (2012) compared rankings based on the lowest absolute test statistic (LAT), lowest effect size (LES), and largest discrimination parameter (MaxA). The choice of MaxA was based on the work of Rivas, Stark, and Chernyshenko (2009), who concluded that higher discriminating items in the anchor produced more optimal true positive and false positive rates than did lower discriminating anchor items. Meade and Wright (2012) determined that the MaxA method outperformed the other methods in their study. However, as Woods (2009) noted, there is no known relationship between DIF and item discrimination. If the goal of empirically selecting anchor items is to ensure a pure anchor, there is not a clear theoretical reason to believe that selecting items based on discrimination would help to ensure that a pure anchor is found. However, Meade and Wright's (2009) results indicate that there may be an advantage to using item discrimination in the anchor selection process. Unfortunately, the authors do not give a theoretical reason to explain their results, nor do they clearly define certain aspects of their study, such as how the items simulated with DIF were chosen. It may be the case

that their study design was responsible for MaxA outperforming the other methods, or it may be that there is a valid reason why MaxA is more likely to choose non-DIF items as anchors. Further research into this issue is needed, although it is beyond the scope of this study.

Kopf et al. (2015b) also directly compared different test statistics to rank potential anchor items. They examined the largest absolute test statistic (LAT), mean p-value (MP), mean test statistic (MT), number of significant tests (NST), number of times the item was above the mean p-value threshold (MPT), and the number of times the item was below the mean test statistic threshold (MTT). Within the constant anchor length class, Kopf et al. (2015b) determined that MPT performed best, although false positive rates were above .05 when tests contained 40% DIF items favoring the reference group.

Need for Further Research

Although much progress has been made in research focused on the empirical selection of anchor items, a method that works well when there is a high percentage of items with DIF favoring a single group has not been found. Generally, as the percentage of DIF increases, the false positive rate increases and the true positive rate decreases (Kopf et al., 2015a; Kopf et al., 2015b; Meade & Wright, 2012; Shih et al., 2014; Shih & Wang, 2009; Wang et al., 2012; Wang & Shih, 2010; Woods, 2009). When testing an empirical dataset for DIF, there is no a priori knowledge about what percentage of items have DIF and which group those items favor. Because of this issue, it is ideal for researchers to have an anchor selection method that works well in the most extreme circumstances, such as a large percentage of items with DIF favoring a single group, and also works well when DIF is balanced or nonexistent.

To date, Kopf et al. (2015a, 2015b) have conducted the most extensive studies on the empirical selection of anchor items. Based on their work, the two most promising methods in

terms of acceptable false positive and true positive rates are IF-SA(MTT) and C4-SA(MPT). While these methods perform well when DIF is balanced or the DIF is unbalanced but the percentage of DIF is 20% or less, they still have suboptimal false positive and true positive rates when 40% of items have DIF and favor the reference group. However, it seems that this process can be addressed by adding a multistage (MS) approach to the current methods: MS[IF-SA(MTT)] and MS[C4-SA(MPT)].

Proposed Methods

The MS approach is conducted using successive stages. In the first stage the anchor method is used to select anchor items and then test all non-anchor items for DIF using the selected anchor. In this stage all items are anchor item candidates and are used as an SA. In the second stage the anchor method is conducted again to select new anchor items and test all non-anchor items for DIF using the newly selected anchor. In this stage only items without a statistically significant DIF test in the previous stage are anchor item candidates and used as an SA. This process is continued until the same set of anchor items are selected in two consecutive stages. That set of anchor items is then used to test all non-anchor items for DIF.

In theory, the percentage of DIF within the candidate anchor items will be reduced in each successive stage, which should improve the accuracy of the anchor selection method. For example, a particular anchor item selection method may have a true positive rate of .50 and a false positive rate of .10 when a test contains 40% DIF items. On a fifty-item test with 40% DIF, a researcher could expect that 3 out of 30 DIF-free items and 10 out of 20 DIF items would be identified as having DIF, leaving 37 items not identified as having DIF. Ten, or 27%, of those 37 remaining items can be expected to have DIF. If the anchor item selection method is conducted on the remaining 37 items, the true positive and false positive rates should improve, since the

percentage of DIF is now less. The newly selected anchor items are now less likely to have DIF and can be used to test all non-anchor items for DIF. The items that are then expected to be DIF-free can be reevaluated as potential anchor items. This process can be continued until the same set of anchor items is identified in two consecutive rounds of testing. Once the final set of anchor items is identified, all non-anchor items can then be tested for DIF.

Conceptually, the MS method is similar to the DIF-Free-then-DIF method proposed by Wang et al. (2012). Within the current framework for classifying anchor item selection methods, the DIF-Free-then-DIF approach can be labeled as C4-(I-AO(ST))(LAT). This method uses the lowest absolute test statistic (LAT) to select four anchor items similar to C4-AO(LAT). However, where C4-AO(LAT) simply ranks items based on LAT obtained by using AO, C4-(I-AO(ST))(LAT) uses I-AO(ST) to first remove items displaying DIF from the pool of candidate anchor items and then calculates and ranks items based on LAT using only the current pool of candidate anchor items. Wang et al. (2012) compared this technique to C4-AO(LAT) and found that C4-(I-AO(ST))(LAT) had better false positive and true positive rates. Theoretically, the reason for the improved performance can be attributed to the fact that using I-AO(ST) prior to calculating and ranking items based on LAT lowered the percentage of DIF in the test, which resulted in more accurate estimates of LAT and rankings. Using the same logic, it can then be expected that the two newly proposed methods will outperform the methods previously proposed by Kopf et al. (2015b).

Methodological Details of Prior Studies

Eleven articles examining the empirical selection of anchor items were located and reviewed (González-Betanzos & Abad, 2012; Khalid & Glas, 2014; Kopf et al., 2015a; Kopf et al., 2015a; Meade & Wright, 2012; Shih et al., 2014; Shih & Wang, 2009; Wang & Shih, 2010;

Wang et al,, 2012; Wang et al., 2009; Woods, 2009). The most commonly manipulated variables were sample size, mean group difference between the reference and focal groups, number of items on the test, percentage of DIF on the test, magnitude of DIF, type of DIF, balance of DIF, and the models used to generate the data. The levels for each of these variables used in the reviewed studies are presented in Table 5, in addition to the number of replications. A brief discussion of each of these variables and their relationship with detection of DIF is presented below.

Sample Size

Sample size has been explored both in terms of overall sample size as well as equal or unequal sample sizes between the reference and focal groups. Equal sample sizes examined ranged from 100 each for the reference and focal groups (Khalid & Glas, 2014; Meade & Wright, 2012) to 1,500 for each group (Kopf et al., 2015a; Kopf et al., 2015b; Woods, 2009). All reviewed studies which examined unequal sample sizes simulated the reference group larger than the focal group (González-Betanzos & Abad, 2012; Kopf et al., 2015a; Meade & Wright, 2012; Shih et al., 2014; Wang et al., 2009; Woods, 2009;). These unequal sample sizes ranged from 500/100 to 5,000/2,000 for the reference/focal groups, respectively.

In general, in terms of detecting items with DIF, false positive rates decrease and true positive rates increase as sample sizes increase. The impact of unequal sample sizes is less clear because the studies that examined unequal sample sizes did not also simulate equal sample sizes that had the same overall sample size as the unequal sample size conditions. For example, Kopf et al. (2015a) simulated six equal sample size conditions and five unequal sample size conditions. However, none of the total sample sizes, determined by adding the sample size for the reference group to the sample size for the focal group, for the unequal condition were equal

to the total sample size for the equal condition. Without holding the total sample size constant for these two conditions, it is difficult to determine the impact, if any, of equal vs. unequal sample sizes between the reference and focal groups.

Mean Group Difference

The simulated mean group differences, often labeled as impact, in the reviewed studies ranged from 0.0 to 1.0 standard deviations. The majority of studies simulated a mean group difference, versus no mean group difference. Only two studies varied the magnitude of the mean group difference, both of which used 0.0 and 1.0 standard deviations (Shih et al. 2014; Wang et al., 2012). In both of these studies the false positive rates were generally higher and the true positive rates were generally lower when there was a mean group difference present indicating that the presence of a mean group difference has a negative impact on DIF detection rates.

Number of Items on the Test

The number of items on the test examined in the reviewed studies ranged from 10 (Wang & Shih, 2010; Woods, 2009) to 80 (Kopf et al., 2015a). Only three studies did not vary the number of items on the test (Woods, 2009; Meade & Wright 2012; Kopf at al., 2015b). The impact the number of items on the test has on DIF detection is unclear. In some studies a larger number of items on the test was generally positively correlated with improved false positive and true positive DIF detection rates (Shih et al., 2014). However, other studies found opposite or mixed results. For example, the results of Wang et al. (2009) showed that both false positive rates and true positive rates generally were lower for shorter tests; however, the results from Wang and Shih (2010) usually showed the same pattern but had many conditions in which this pattern did not hold. These mixed results may indicate that the impact the number of items on the test has on DIF detection rates is dependent on other study conditions.

Table 5

Study design for selected variables in reviewed studies

Study	Sample Size[a]		Mean Group Difference[b]	Number of Items	Percentage of DIF	Type of DIF	Balance of DIF	Data Generation Model	Absolute Magnitude of DIF		Replications
	Equal	Unequal							a	b	
González-Betanzos & Abad, 2012	500, 1000	1000/500	0.0	15	27	U, NU	O	2PL	.15, .40	.25, .50	100
Khalid & Glas, 2014	100, 400, 1000	--	0.0	10, 20, 40	0, 10, 20, 30, 40	U, NU	O	1PL, 2PL, 3PL	.5	.5, 1.0	100
Kopf et al., 2015a	250, 500, 750, 1000, 1250, 1500	500/250, 750/500, 1000/750, 1250/1000, 1500/1250	1.0	20, 40, 60, 80	15, 30, 45	U	B, O	1PL	--	.6	2000
Kopf et al., 2015b	250, 500, 750, 1000, 1250, 1500	--	1.0	40	0, 10, 25, 40	U	B, O	1PL	--	.4	1000
Meade & Wright, 2012	100, 250, 500	500/250, 5000/2000	0.5	20	5, 10, 20, 40, 75	U, NU	O	GR	.5	.4, .8	300, 100
Shih, et al., 2014	250, 500	500/250, 1000/500	0.0, 1.0	20, 60	0, 10, 20, 30, 40	U	O	2PL, 3PL	--	*N*(.4,.01)	100
Shih & Wang, 2009	500, 1000, 1500	--	0.5	20, 30, 40	0, 10, 20, 30, 40	U	B, O, D	1PL, 2PL, 3PL	--	.4	100
Wang et al., 2009	500	500/100, 1000/500	1.0	20, 50	0, 10, 20, 30, 40	U	B, O	2PL, 3PL	--	.6	100
Wang, et al., 2012	250, 500, 1000	--	0.0, 1.0	20, 40	10, 20, 30, 40	U	B, O	2PL	--	*N*(.6,.01)	100
Wang & Shih, 2010	500		1.0	10, 20, 30	0, 10, 20, 30, 40	U	O	PC, GPC	--	.25	100
Woods, 2009	--	1500/500	0.4	10, 20, 40	0, 20, 50, 80	U, NU	R	GR	.3,.4,.5,.6,.7	≤ 1.52	100

Note. [a]Equal sample sizes are per group. Unequal sample sizes are for the reference/focal groups. [b]Differences are in standard deviations. U=uniform, N=non-uniform, B=balanced, O=one-sided, D=dominant, R=random, PC=partial credit, GPC=generalized partial credit, GR=graded response.

Percentage of DIF on the Test

The percentage of DIF on the test is calculated by dividing the number of items with DIF by the total number of items on the test. For example, a 40-item test with four DIF items would have 10% DIF. Only one of the reviewed studies did not vary the percentage of DIF (González-Betanzos & Abad, 2012). The impact of the percentage of DIF is largely influenced by the balance of DIF. On tests where there is the same number of DIF items favoring the reference groups as the focal group, assuming the magnitude of DIF is equal, the percentage of DIF has little impact on the detection of DIF. However, when all items with DIF favor a single group, a larger percentage of DIF leads to higher false positive and lower true positive rates when detecting items with DIF.

Wang & Su (2004) attributed these observations to the average signed area (ASAR). When ASAR is equal to zero the items with DIF have little impact on DIF detection. However, greater magnitudes of ASAR lead to higher false positive and lower true positive rates when detecting items with DIF. Assuming the item level magnitude of DIF is constant, balanced DIF conditions will have an ASAR of zero, and in unbalanced DIF conditions ASAR will be positively correlated with the percentage of items with DIF.

Magnitude of DIF

The magnitude of DIF refers to the difference in the item parameters between the reference and focal groups. Most of the reviewed studies which explored the empirical selection of anchor items did not vary the magnitude of DIF but simply specified a constant magnitude for each item parameter being simulated with DIF. Three studies did vary the magnitude of DIF, all of which showed true positive rates were generally higher in conditions in which DIF items had a larger magnitude of DIF than conditions with a smaller magnitude of DIF (González-Betanzos &

Abad, 2012; Khalid & Glas, 2014; Meade & Wright, 2012). The relationship between DIF magnitude and false positive rates was less clear. Meade and Wright (2012) did not report their false positive rates in a way that allows any inference in the relationship between DIF magnitude and false positive rates to be made. Khalid and Glas's (2014) results showed no clear relationship between the magnitude of DIF and false positive rates. González-Betanzos and Abad's (2012) results showed that relationship between the magnitude of DIF and false positive rates was dependent on the anchor item method. When all other items were used as an anchor, conditions with a large magnitude of DIF had a higher false positive rate than conditions with a smaller magnitude of DIF. However, once an anchor method was applied, conditions with a larger magnitude of DIF had a smaller false positive rate than conditions with a smaller magnitude of DIF.

Type of DIF

The type of DIF examined in the reviewed studies was either uniform or non-uniform. All the reviewed studies examined uniform DIF. Three studies also examined non-uniform DIF (González-Betanzos & Abad, 2012, Meade & Wright, 2012; Woods, 2009). The impact of type of DIF on false positive and true positive DIF detection rates is unclear. González-Betanzos & Abad (2012) found that uniform DIF generally had a higher true positive DIF detection rate but also a larger false positive DIF detection rates than non-uniform. Conversely, both Woods (2009) and Meade and Wright (2012) found non-uniform DIF generally had higher true positive rates than uniform DIF. In regards to false positive rates, Woods's (2009) results showed no clear difference between uniform and non-uniform DIF, while Meade and Wright (2012) did not report their false positive rates in a way which allows the reader to determine if there is any impact due to type of DIF. There were major differences in the designs of these studies which

may impact the results and make comparisons between studies difficult. For example, González-Betanzos & Abad (2012) used a two-parameter IRT model to generate their data, while both Woods (2009) and Meade and Wright (2012) used graded response models.

Balance of DIF

The balance of DIF refers to whether items with DIF favor the reference or focal group. There were two commonly used balances of DIF used in the reviewed simulation studies: balanced and one-sided. Balanced DIF occurs when half of the items with DIF favor the reference group and the other half favor the focal group. One-sided DIF refers to conditions where all the DIF items favor a single group. With the exception of Woods (2009), all of the reviewed studies included a one-sided condition in their simulation. Five studies also included a balanced condition.

The impact of the balance of DIF can be explained by ASAR. When DIF is balanced, ASAR is zero which leads to lower false positive rates and higher true positive rates when detecting DIF. However, when DIF is one-sided, ASAR is not zero which leads to higher false positive rates and lower true positive rates (Wang & Su, 2004).

Models Used to Generate Data

The models used to generate the data in the reviewed studies were either the one-parameter, two-parameter, three-parameter, graded response, partial credit, or generalized partial credit IRT models. Only five studies used more than one model to generate data. The most obvious impact the model used to generate had on DIF detection was dependent on the model used to detect DIF. True and false positive DIF detection rates were improved when the model used to detect DIF was the same or equivalent to the model used to generate the data as opposed

to when the model used to detect DIF was not equivalent or the same (Shih & Wang, 2009; Wang et al, 2009).

Number of Replications

Eight of the reviewed studies only used 100 replications for each of their study conditions. Meade and Wright (2012) used 300 replications for their primary simulation and 100 replications for a secondary simulation. Kopf et al. (2015a) used 2000 replications in their first study and then used 1000 replications in their second study (Kopf et al., 2015b). There were no justifications provided for these numbers of replications.

CHAPTER 3:

STUDY DESIGN

The purpose of this study was to determine if two newly proposed methods used for the empirical selection of anchor items for DIF analyses outperform two currently established methods proposed by Kopf et al. (2015b). Two sets of DIF-free anchor items were also used for comparison purposes. Slight changes were made to the stopping criteria in the methods proposed by Kopf et al. (2015b) due to differences in the specification of anchor items. These changes are discussed in detail later in this chapter.

The goal of the study was to support or refute three hypotheses: (1) The multistage anchor selection methods will have higher true positive rates, lower false positive rates, lower familywise false positive rates, lower anchor contamination, and lower familywise anchor contamination than the non-multistage methods. (2) The anchor selection methods using IF will have higher true positive rates but also higher false positive rates than anchor selection methods using C4. (3) Familywise false positive rates will be greater than .05 for most, or all, conditions.

Additionally, this study addressed two study questions: (1) Will any of the studied methods result in DIF detection rates equal to the DIF detection rates for the DIF-free anchors for all conditions? (2) Will there be a difference in the anchor contamination rates between the IF and C4 methods?

A simulation study was conducted to address these hypotheses and question. The Rasch model was used for data generation. The sample size, percentage of DIF, and direction of DIF

were manipulated. The Rasch model and the Wald test were used for the selection of anchor items and DIF analysis, which were applied using PROC IRT with the marginal likelihood estimator in SAS 9.4 TS1M4. A detailed description of mean p-value threshold (MPT), mean test statistic threshold (MTT), all four anchor items selection methods, two DIF-free anchors, data generation, manipulated variables, and outcome variables are provided below. Also, two key changes made to the anchor items specification and the stopping criteria used by Kopf et al. (2015b) are discussed.

Determining MPT and MTT

To determine MPT, each item was first preliminarily tested for DIF using SA resulting in k-1 DIF tests for each item, where k was the number of items on the test. Then the item-level mean p-values were ranked from large to small and the ($[0.5 \cdot k]$)-th ranked mean p-value (Kopf et al, 2015b, p. 35) was identified. For example, on the twenty-item test, there were 20 item-level mean p-values. Each of these mean p-values was the mean of the 19 DIF tests for each individual item. Those item-level mean p-values were ranked from large to small, and the 10^{th} ranked item-level mean p-value was identified as the MPT.

Similar to MPT, MTT was determined by ranking the item-level mean test statistics from large to small and identifying the ($[0.5 \cdot k]$)-th ranked mean test statistic. The absolute value of each test statistic was used to calculate the item-level mean.

Anchor Selection Methods

C4-SA(MPT)

One of the two methods proposed by Kopf et al. (2015b) used a constant anchor length of four items (C4) which were selected by using each item as a single anchor (SA) and ranked items based on the number of times the items were above the mean p-value threshold (MPT). This

anchor selection method was abbreviated as C4-SA(MPT), and a detailed description of this method is provided below.

1. Use each item as a single anchor (SA) and test every other item for DIF resulting in *k*-1 DIF tests for each item, where *k* is the number of items on the test. These DIF tests are conducted using the Wald test.
2. Calculate the number of times the *k*-1 p-values for each item is above the MPT.
3. Rank items based on the number of times the p-value for each item is above the MPT. Items above MPT the greatest number of times are theorized to be the items most likely to be DIF-free and are ranked the highest.
4. Choose the four highest ranked items (C4) to serve as the anchor and test all non-anchor items for DIF using the Wald test.

IF-SA(MTT)

The second method proposed by Kopf et al. (2015b) used an iterative forward scale purification (IF) class approach to select anchor items. Candidate anchor items were ranked by using each item as a single anchor (SA) and ranking items based on the number of times the items were below the mean test statistic threshold (MTT). Items were then added to the anchor one at a time based on this ranking as long as the number of non-anchor items not displaying statistically significant DIF was longer than the current anchor. This anchor selection method was abbreviated as IF-SA(MTT), and a detailed description of this method is provided below.

1. Use each item as a single anchor (SA) and test every other item for DIF resulting in *k*-1 DIF tests for each item, where *k* is the number of items on the test. These DIF tests are conducted using the Wald test.

2. Calculate the number of times the *k*-1 absolute Wald test statistic for each item is below the MTT.
3. Rank items based on the number of times the test statistic for each item is below the MTT. Items below MTT the greatest number of times are theorized to be the items most likely to be DIF-free and are ranked the highest.
4. Choose the highest ranked item to serve as the anchor and test all non-anchor items for DIF using the Wald test.
5. If the number of non-anchor items not displaying statistically significant DIF is longer than the current anchor, add the next highest ranked item based on the rankings in step 3 to the anchor and retest all non-anchor items for DIF.
6. Repeat step 5 until all the number of non-anchor items not displaying statistically significant DIF is not longer than the current anchor.
7. Use the final set of anchor items to test all non-anchor items for DIF.

The two newly proposed methods extended the methods proposed by Kopf et al. (2015b) by adding multiple iterative stages to each method. The newly proposed multistage (MS) methods were abbreviated as MS[C4-SA(MPT)] and MS[IF-SA(MTT)]. Detailed procedures for conducting each of these methods are provided below. The maximum number of stages per replication was limited to 10 in order to deal with the possibility of encountering infinite loops in the simulation.

MS[C4-SA(MPT)]

Stage 1

1. Use each item as a single anchor (SA) and test every other item for DIF resulting in *k*-1 DIF tests for each item, where *k* is the number of items on the test. These DIF tests are conducted using the Wald test.
2. Calculate the number of times the *k*-1 p-values for each item is above the MPT.
3. Rank items based on the number of times the p-value for each item is above the MPT. Items above MPT the greatest number of times are theorized to be the items most likely to be DIF-free and are ranked the highest.
4. Choose the four highest ranked items (C4) to serve as the anchor and test all non-anchor items for DIF using the Wald test.

Stage 2

5. Repeat steps 1-4, but items displaying statistically significant DIF in step 4 are not included as an SA, tested for DIF by any other SA, or included in the ranking based on MPT. However, these items, along with all other non-anchor items, are tested for DIF when step 4 is repeated during stage 2. In other words, stage 2 identifies four new items to serve as an anchor, and all non-anchor items are tested for DIF using that anchor.

Stage *i*

6. Repeat step 5, *i* times, until the same set of anchor items are identified in two consecutive stages. Use the final set of anchor items to test all non-anchor items for DIF.

MS[IF-SA(MTT)]

Stage 1

1. Use each item as a single anchor (SA) and test every other item for DIF resulting in *k*-1 DIF tests for each item, where *k* is the number of items on the test. These DIF tests are conducted using the Wald test.
2. Calculate the number of times the *k*-1 absolute Wald test statistics for each item is below the MTT.
3. Rank items based on the number of times the test statistic for each item is below the MTT. Items below MTT the greatest number of times are theorized to be the items most likely to be DIF-free and are ranked the highest.
4. Choose the highest ranked item to serve as the anchor and test all non-anchor items for DIF using the Wald test.
5. If the number of non-anchor items not displaying statistically significant DIF is longer than the current anchor, add the next highest ranked item based on the rankings in step 3 to the anchor and retest all non-anchor items for DIF.
6. Repeat step 5 until the number of non-anchor items not displaying statistically significant DIF is not longer than the current anchor.
7. Use the final set of anchor items to test all non-anchor items for DIF.

Stage 2

8. Repeat steps 1-7, but items displaying statistically significant DIF in step 7 are not included as an SA, tested for DIF by any other SA, or included in the ranking based on MTT. However, these items, along with all other non-anchor items, are tested for DIF

when step 7 is repeated during stage 2. In other words, stage 2 identifies new items to serve as an anchor, and all non-anchor items are tested for DIF using that anchor.

Stage *i*

9. Repeat step 8, *i* times, until the same set of anchor items are identified in two consecutive stages. Use the final set of anchor items to test all non-anchor items for DIF.

DIF-free Anchors

Two sets of DIF-free anchors, designed to mirror the anchor length of the C4 and IF anchor classes, were used to identify DIF items in each study condition. The C4-DIF-free anchor consisted of four items for all percentages of DIF. The anchor length for the IF-DIF-free anchor was dependent on the percentage of DIF as shown in the table below. These anchor lengths were based on the stopping criteria used for the IF anchor class applied in this study. If the IF anchor class accurately identified every DIF item, the anchor length would be half the length of the number of non-DIF items. The items within each DIF-free anchor were randomly chosen separately for each replication from the set of simulated DIF-free items.

Table 6
Anchor Lengths for IF-DIF-free Anchor by Percentage of DIF

Percentage of DIF	Anchor Length
0	10
10	9
20	8
40	6

Data Generation

Statistical Software

SAS 9.4 TS1M4 was used to generate the datasets, apply the anchor method, and perform the final DIF test.

Model Used to Generate Data

All data were generated under the Rasch model. Although there were numerous models that could have been used to generate the data, Kopf et al. (2015b) used the Rasch model for their study. Since this study was a direct extension of their study, using the Rasch model allowed for a more direct comparison between study results.

Mean Group Difference

Ability levels for the reference and focal groups were randomly generated from a normal distribution with a mean of 0 and -1, respectively. Both groups had a standard deviation of 1. This design simulated real life situations where often the mean ability level of the reference group is greater than the focal group (Wang, 2004). Additionally, when there is a difference in the mean ability between groups, accurate DIF detection tends to be more challenging than when there is no mean ability difference (Shih et al., 2014; Wang et al., 2012;). Using these ability level distributions allows this study to examine the performance of the proposed methods in conditions that both simulate real life and are more challenging than a condition without mean ability differences.

Test Length

The test length was 20 items. This length was chosen because it has often been used in similar simulation studies (Khalid & Glas, 2014; Kopf et al., 2015a; Shih et al., 2014; Shih & Wang, 2009; Wang et al., 2009; Wang et al., 2012; Wang & Shih, 2010; Woods, 2009). Additionally, actual cognitive instruments often use a similar number of items (Morris, Lee & Barnes, 2008; Parslow, Christensen, Griffiths & Groves, 2006; Sekercioglu, Bayat & Bakir, 2014).

Item Parameters

The difficulty parameters for the items used in this simulation, displayed in Table 7, are identical to the parameters used by Kopf et al. (2015b) and Wang et al. (2012). In both of those studies, the researchers simulated the first *X*% percentage of items with DIF, where *X* is the percentage of items with DIF. However, this design ensured that under many values of *X* only items with a low difficulty parameter are simulated with DIF. In the current study, the same difficulty parameters were used and the first *X*% percentage of items were simulated with DIF; however, the order of the difficulty parameters was randomized so that all difficulty parameters have an equal probability of being simulated with DIF. Additionally, 20 of the 40 items were randomly chosen for each replication in the study.

Table 7
Difficulty parameters used by Kopf et al. (2015b) and Wang et al. (2012)

Item	*b*-parameter	Item	*b*-parameter	Item	*b*-parameter	Item	*b*-parameter
1	-2.522	11	0.295	21	-2.198	31	0.116
2	-1.902	12	0.778	22	-1.621	32	0.273
3	-1.351	13	1.514	23	-0.761	33	0.840
4	-1.092	14	1.744	24	-1.179	34	0.745
5	-0.234	15	1.951	25	-0.610	35	1.485
6	-0.317	16	-1.152	26	-0.291	36	-1.208
7	0.037	17	-0.526	27	0.067	37	0.189
8	0.268	18	1.104	28	0.706	38	0.345
9	-0.571	19	0.961	29	-2.713	39	0.962
10	0.317	20	1.314	30	0.213	40	1.592

DIF Magnitude

For items with DIF, the difference between the difficulty parameters for reference and focal groups was 0.4. When items favored the reference group, 0.4 was added to the difficulty parameter for the focal group. When items favored the focal group, 0.4 was subtracted from the

difficulty parameter for the focal group. This magnitude was used in previous simulation studies (Kopf et al., 2015b; Rogers & Swaminathan, 1993).

Manipulated Variables

Sample Size

Three sample sizes were included in this study for the reference/focal groups: 500/500, 750/750, and 1000/1000. In general, larger sample sizes lead to lower false positive and higher true positive rates during DIF detection (González-Betanzos & Abad, 2012; Khalid & Glas, 2014; Kopf et al., 2015a; Kopf et al., 2015b; Shih et al., 2014; Shih & Wang, 2009; Wang et al., 2010). However, as real life testing situations may include small and large sample sizes, it was useful to include a variety of sample sizes in this simulation.

Percentage of DIF

Four percentages of DIF were included in this study: 0%, 10%, 20%, and 40%. In previous simulation studies higher percentages of DIF have led to higher false positive and lower true positive rates, especially when all DIF items favor the same group (Khalid & Glas, 2014; Kopf et al., 2015a; Kopf et al., 2015b; Meade & Wright, 2012; Shih et al., 2014; Shih & Wang, 2009; Wang et al., 2009; Wang et al., 2012; Wang & Shih, 2010; Woods, 2009). The 40% DIF condition is of particular interest because that is the condition under which the methods proposed by Kopf et al. (2015b) performed the worst.

Balance of DIF

This study included a balanced and a one-sided DIF condition. In the balanced condition half the DIF items favored the reference group, and half favored the focal group. In the one-sided condition all DIF items favored the reference group. Typically, most anchor selection methods work well under the balanced condition. Under the one sided condition, many of the previously

proposed methods for the empirical selection of anchor items have not performed as well, especially when the test contains higher percentages of DIF (González-Betanzos & Abad, 2012; Khalid & Glas, 2014; Kopf et al., 2015a; Kopf et al., 2015b; Meade & Wright, 2012; Shih et al., 2014; Shih & Wang, 2009; Wang et al., 2009; Wang et al., 2012; Wang & Shih, 2010).

Number of Replications

The number of replications for this study was determined using Bradley's (1978) criteria for liberal robustness which he defined as 0.5α, where α is the nominal Type 1 error rate. Using a common α of .05, the formula for the standard error of a binomial distribution, shown in Equation 20, was used to determine the number of replications needed to obtain estimates of the outcome variables ±.025 of the true value. Under a binomial distribution, the greatest amount of variance will occur when the observed value is equal to .5; therefore, .5 was used to determine that 400 replications were needed to obtain acceptable estimates of the outcome variables.

$$SE = \sqrt{\frac{p(1-p)}{n}} \tag{20}$$

Where

SE *is the standard error of a binomial distribution,*

p *is the proportion of a certain outcome, and*

n *is the number of replications.*

Outcomes

The performance of the anchor selection methods were evaluated using false positive rates, true positive rates, familywise false positive rates, anchor contamination rates, and familywise anchor contamination rates. With the exception of familywise false positive rates, these outcomes have been used in other simulations studies examining the empirical selection of

anchor items (Gonzalez-Betanzos & Abad, 2011; Kopf et al., 2015a; Kopf et al., 2015b; Meade & Wright, 2012; Shih et al., 2014; Shih & Wang, 2009; Wang et al., 2012; Wang & Shih, 2010; Woods, 2009). Familywise false positive rates were included because testing multiple items for DIF within the same item test may result in inflated false positive rates.

False Positive Rate

The false positive rate was defined as the proportion of DIF-free items showing statistically significant DIF during the final DIF test.

True Positive Rate

The true positive rate was defined as the proportion of DIF items showing statistically significant DIF during the final DIF test.

Familywise False Positive Rate

The familywise false positive rate was defined as the proportion of replication with at least one DIF-free item within the 20-item test shows statistically significant DIF during the final DIF test.

Anchor Contamination Rate

The anchor contamination rate was defined as the proportion of DIF items in the final anchor.

Familywise Anchor Contamination Rate

The familywise anchor contamination rate was defined as the proportion of replications with a final anchor that contained at least one DIF item.

Changes to Kopf et al.'s (2015b) Anchor Methods

Two changes were made to the methods used by Kopf et al. (2015b) due to differences in the software used in each study. These differences were the specification of anchor items and the stopping criteria for the IF anchor class.

Kopf et al. (2015b) constrained the mean difficulty parameter for all anchor items to 0. The b-parameter of one of the anchor items was constrained to 0 to identify the model. This allowed the b-parameters for other anchor items within the anchor to vary between groups. These parameter estimates could then be tested for DIF along with the non-anchor items. In this current study, the b-parameters for anchor items were constrained to be equal between groups. This meant that none of the anchor items could be tested for DIF; therefore, these items were treated as having been identified as DIF-free for the calculation of outcome variables. This change to the specification of anchor items led to the need to change the stopping criteria for the IF anchor class.

When applying the IF anchor class, Kopf et al. (2015b) stopped adding items to the anchor when the length of the anchor was equal to or greater than the number of items not being identified as having DIF. For a 20-item test with 40% DIF, if the IF anchor class worked perfectly then all 12 non-DIF items would be included in the anchor. This anchor class could only work perfectly if the 12 non-DIF items were the 12 highest ranked items using SA(MTT). This was rarely the case, and items with DIF were often included in the anchor. However, because the specification of anchor items used by Kopf et al. (2015b) allowed all but one anchor item to be tested for DIF, even DIF items included in the anchor could be identified as having DIF.

When the same stopping criteria was combined with the specification of anchor items used in this study, there were two adverse effects. The true positive rate decreased, and the anchor contamination rate increased. These adverse effects were directly attributable to the fact that all items in the anchor were treated as DIF-free items. Table 8 shows the anchor lengths resulting from different combinations of the stopping criteria and anchor item specifications used in Kopf et al.'s (2015b) study and this current study.

Table 8
Anchor lengths from applying different stopping criteria and anchor item specifications on a five item test with 40% differential item functioning

Item Rank Using SA(MTT)	Simulated DIF	Anchor Length				
		1	2	3	4	5
1	No	Anchor	Anchor	Anchor	Anchor	Anchor
2	No		Anchor	Anchor	Anchor	Anchor
3	Yes			Anchor	Anchor	Anchor
4	Yes				Anchor	Anchor
5	No					Anchor
Stopping Criteria	**Anchor Item Specification**	**Test of Stopping Criteria for Each Anchor Length**				
Kopf et al.[1]	Kopf et al.[2]	1 ≥ 3 False	2 ≥ 3 False	3 ≥ 3 True	--	--
Kopf et al.[1]	Current Study[3]	1 ≥ 3 False	2 ≥ 3 False	3 ≥ 4 False	4 ≥ 5 False	5 ≥ 5 True
Current Study[4]	Current Study[3]	1 ≥ 2 False	2 ≥ 1 True	--	--	--

Note. The results in this table assume that any item tested for DIF was correctly identified as either a DIF or non-DIF item. [1]The number of items in the anchor is greater than or equal to the number of DIF-free items. [2]The mean b-parameters of the anchor items was constrained to 0, which allowed all but 1 anchor item to be tested for DIF. [3]The b-parameters of the anchor items were constrained equal between groups which did not allow the anchor items to be tested for DIF. [4]The number of items in the anchor is greater than or equal to the number of DIF-free non-anchor items.

This table shows that the item ranking using SA(MTT) has an item without DIF ranked after two items with DIF. Applying the stopping criteria and anchor item specification used by Kopf et al. (2015b) results in a three-item anchor since item three can still be identified as having

DIF when it is in the anchor. However, combining the stopping criteria used by Kopf et al. (2015b) with the anchor item specification used in this study results in a five-item anchor, which would mean that no items would be identified as DIF since the anchor item specification used in this study does not allow anchor items to be tested for DIF. By changing the stopping criteria in this study so that the number of items in the anchor is compared to the number of non-anchor items identified as DIF-free and combining that stopping criteria with the anchor item specification used in this study, a two-item anchor is identified. Preliminary analysis showed that this combination of stopping criteria and anchor item selection resulted in higher true positive rates and lower anchor contamination rates than combining Kopf et al.'s (2015b) stopping criteria with the current anchor item specification.

It should also be noted that the stopping criteria used in this study limited the length of the anchor to half the number of DIF-free items, assuming the presence of DIF was accurately identified. The stopping criteria used by Kopf et al. (2015b) could result in an anchor length equal to the number of items in the test. For example, on a 20-item test with 0% DIF, the stopping criteria used in this study could result in an anchor no longer than 10 items since an anchor length of 10 would be as long as the remaining 10 items being identified as 10 DIF-free items. However, Kopf et al.'s (2015b) stopping criteria would result in a 20-item anchor since the items in the anchor are included in the count of DIF-free items, and the stopping criteria would not be met until there was a 20-item anchor.

CHAPTER 4:

RESULTS

The main outcomes of interest for this study were true positive rates, false positive rates, familywise false positive rates, anchor contamination rates, and familywise anchor contamination rates. The results for each of those outcomes are presented in this chapter in five separate tables. Major trends or findings within each of those outcomes are noted in the text that follows. Additionally, the mean, minimum, and maximum observed anchor lengths for the IF anchor selection methods are reported. A discussion of these results, including the ways in which the findings support or refute the hypotheses and research questions for this study, are presented in Chapter 5.

True Positive Rates

Table 9 shows the true positive DIF detection rates for all conditions in the simulation. As expected, true positive rates were higher for larger sample sizes, IF anchor selection methods versus C4 anchor selection methods, and the balanced DIF conditions versus the unbalanced DIF conditions. The difference in true positive rates was negligible between the DIF-free, multistage, and non-multistage methods under most conditions. There was a difference in true positive rates between anchor methods under the one-sided 40% percent DIF condition. Under this condition, the DIF-free methods had higher true positive rates than both the non-multistage and multistage methods. Additionally, the multistage methods had slightly higher true positive rates than the non-multistage methods, especially with larger sample sizes.

Table 9

True positive rates by sample size, percentage of DIF, balance of DIF, and anchor method

Percentage of DIF			Balanced DIF			One-Sided DIF		
			Sample Size Per Group					
		Anchor Method	500	750	1000	500	750	1000
	10	C4-DIF-free	.46	.69	.80	.41	.66	.78
		C4-SA(MPT)	.46	.70	.83	.41	.67	.80
		MS[C4-SA(MPT)]	.46	.68	.82	.40	.66	.80
		IF-DIF-free	.49	.72	.84	.43	.69	.82
		IF-SA(MTT)	.51	.74	.84	.44	.70	.82
		MS[IF-SA(MTT)]	.50	.72	.84	.44	.68	.82
	20	C4-DIF-free	.49	.66	.81	.47	.67	.78
		C4-SA(MPT)	.49	.68	.83	.43	.65	.79
		MS[C4-SA(MPT)]	.48	.68	.82	.43	.63	.79
		IF-DIF-free	.50	.69	.85	.49	.69	.82
		IF-SA(MTT)	.52	.70	.85	.45	.66	.80
		MS[IF-SA(MTT)]	.51	.70	.84	.46	.66	.81
	40	C4-DIF-free	.49	.67	.82	.47	.66	.79
		C4-SA(MPT)	.49	.67	.80	.28	.51	.68
		MS[C4-SA(MPT)]	.48	.68	.81	.31	.54	.73
		IF-DIF-free	.50	.70	.83	.49	.68	.81
		IF-SA(MTT)	.51	.68	.81	.28	.47	.63
		MS[IF-SA(MTT)]	.51	.69	.83	.32	.55	.74

False Positive Rates

False positive rates, shown in Table 10, were well controlled under all conditions, and arguably even slightly conservative under most conditions since the majority of the false positive rates were .01 and .02. Note that Kopf et al. (2015b) also found false positive rates to be under .05 for many of the methods used in their simulation, so the results in this simulation study were not out of line with the results of other similar studies.

Table 10

False positive rates by sample size, percentage of DIF, balance of DIF, and anchor method

Percentage of DIF	Anchor Method	Balanced DIF			One-Sided DIF		
		Sample Size Per Group					
		500	750	1000	500	750	1000
0	C4-DIF-free	.02	.02	.02	--	--	--
	C4-SA(MPT)	.02	.02	.02	--	--	--
	MS[C4-SA(MPT)]	.02	.02	.01	--	--	--
	IF-DIF-free	.01	.01	.01	--	--	--
	IF-SA(MTT)	.02	.02	.02	--	--	--
	MS[IF-SA(MTT)]	.02	.02	.02	--	--	--
10	C4-DIF-free	.02	.02	.02	.02	.02	.02
	C4-SA(MPT)	.02	.02	.02	.02	.02	.02
	MS[C4-SA(MPT)]	.02	.02	.02	.02	.02	.02
	IF-DIF-free	.01	.01	.01	.01	.01	.01
	IF-SA(MTT)	.02	.02	.02	.02	.02	.02
	MS[IF-SA(MTT)]	.02	.02	.02	.02	.02	.02
20	C4-DIF-free	.02	.02	.02	.02	.02	.02
	C4-SA(MPT)	.02	.02	.01	.02	.02	.02
	MS[C4-SA(MPT)]	.02	.02	.02	.02	.02	.02
	IF-DIF-free	.01	.01	.01	.01	.01	.01
	IF-SA(MTT)	.02	.02	.02	.02	.02	.02
	MS[IF-SA(MTT)]	.03	.02	.02	.03	.02	.02
40	C4-DIF-free	.02	.02	.01	.02	.02	.01
	C4-SA(MPT)	.03	.02	.02	.06	.05	.05
	MS[C4-SA(MPT)]	.03	.02	.02	.07	.05	.05
	IF-DIF-free	.01	.01	.01	.01	.01	.01
	IF-SA(MTT)	.02	.02	.02	.06	.06	.05
	MS[IF-SA(MTT)]	.03	.02	.02	.07	.06	.05

There were no clear differences in false positive rates between anchor methods or with the DIF-free anchors except under the one-sided 40% DIF condition. Under that condition, there was a noticeable difference in false positive rates between all four anchor selection methods and

the DIF-free anchors. However, even with this difference the false positive rates for all anchor methods were well controlled.

Table 11

Familywise false positive rates by sample size, percentage of DIF, balance of DIF, and anchor method

Percentage of DIF	Anchor Method	Balanced DIF			One-Sided DIF		
		Sample Size Per Group					
		500	750	1000	500	750	1000
	C4-SA(MPT)	.32	.27	.25	--	--	--
	MS[C4-SA(MPT)]	.29	.25	.22	--	--	--
	IF-DIF-free	.18	.18	.16	--	--	--
	IF-SA(MTT)	.37	.32	.30	--	--	--
	MS[IF-SA(MTT)]	.35	.30	.29	--	--	--
10	C4-DIF-free	.29	.28	.26	.30	.29	.26
	C4-SA(MPT)	.32	.26	.25	.34	.26	.23
	MS[C4-SA(MPT)]	.33	.27	.26	.34	.26	.25
	IF-DIF-free	.16	.18	.15	.16	.17	.15
	IF-SA(MTT)	.36	.29	.27	.36	.29	.28
	MS[IF-SA(MTT)]	.36	.29	.29	.36	.29	.28
20	C4-DIF-free	.27	.25	.22	.28	.26	.22
	C4-SA(MPT)	.31	.23	.20	.29	.25	.22
	MS[C4-SA(MPT)]	.32	.25	.23	.31	.25	.21
	IF-DIF-free	.15	.17	.14	.15	.17	.14
	IF-SA(MTT)	.34	.25	.23	.33	.26	.24
	MS[IF-SA(MTT)]	.33	.28	.27	.32	.27	.28
40	C4-DIF-free	.20	.17	.15	.20	.17	.15
	C4-SA(MPT)	.27	.19	.18	.36	.29	.28
	MS[C4-SA(MPT)]	.30	.20	.20	.35	.27	.24
	IF-DIF-free	.12	.11	.11	.12	.11	.12
	IF-SA(MTT)	.25	.20	.19	.43	.37	.36
	MS[IF-SA(MTT)]	.28	.23	.22	.42	.30	.29

Familywise False Positive Rates

Familywise false positive rates, shown in Table 11, ranged from .11 to .43 across all conditions. Familywise false positive rates were generally lower as the percentage of DIF increased, lower for IF-DIF-free compared to all other anchor methods, and lower as sample size increased. There were no clear differences in familywise false positive rates between the multistage and non-multistage methods. Familywise false positive rates were highest under the 40% one-sided DIF condition.

Table 12

Anchor contamination by sample size, percentage of DIF, balance of DIF, and anchor method

			Balanced DIF			One-Sided DIF		
			Sample Size Per Group					
Percentage of DIF		Anchor Method	500	750	1000	500	750	1000
	10	C4-SA(MPT)	.016	.007	.003	.019	.006	.001
		MS[C4-SA(MPT)]	.016	.008	.003	.017	.008	.001
		IF-SA(MTT)	.016	.005	.002	.017	.006	.002
		MS[IF-SA(MTT)]	.016	.006	.002	.017	.006	.002
	20	C4-SA(MPT)	.031	.016	.003	.041	.016	.006
		MS[C4-SA(MPT)]	.031	.014	.002	.042	.018	.006
		IF-SA(MTT)	.030	.010	.005	.043	.014	.009
		MS[IF-SA(MTT)]	.032	.012	.004	.040	.018	.006
	40	C4-SA(MPT)	.087	.041	.014	.244	.132	.082
		MS[C4-SA(MPT)]	.093	.039	.013	.230	.115	.069
		IF-SA(MTT)	.099	.047	.029	.251	.179	.128
		MS[IF-SA(MTT)]	.102	.041	.015	.224	.128	.062

Anchor Contamination Rates

The percentage of contaminated anchors, shown in Table 12, ranged from .001 to .251, decreased as sample size increased, and was highest under the one-sided 40% DIF condition. There was generally not a difference in anchor contamination between anchor selection methods,

except when DIF was set to 40%. With 40% DIF and sample sizes of 750 or 1000 per group, the multistage IF method performed better than the non-multistage IF method. Under the one-sided, 40% DIF condition, all multistage methods performed better than the non-multistage methods.

Familywise Anchor Contamination Rates

Familywise anchor contamination rates ranged from .005 to .945. Contamination decreased with increased sample size, was higher for larger percentages of DIF, and was generally higher for the one-sided DIF condition compared to the balanced DIF condition. Familywise anchor contamination was largest under the one-sided, 40% DIF condition. Under that condition, the multistage methods produced lower familywise anchor contamination than the non-multistage methods. Under all conditions, the C4 methods had lower familywise anchor contamination rates than the IF methods.

Table 13
Familywise anchor contamination by sample size, percentage of DIF, balance of DIF, and anchor method

			Balanced DIF			One-Sided DIF		
			Sample Size Per Group					
Percentage of DIF		Anchor Method	500	750	1000	500	750	1000
	10	C4-SA(MPT)	.063	.028	.010	.075	.025	.005
		MS[C4-SA(MPT)]	.063	.033	.010	.065	.030	.005
		IF-SA(MTT)	.150	.045	.018	.160	.060	.018
		MS[IF-SA(MTT)]	.158	.055	.020	.158	.055	.015
	20	C4-SA(MPT)	.120	.063	.010	.158	.063	.020
		MS[C4-SA(MPT)]	.123	.053	.008	.165	.073	.025
		IF-SA(MTT)	.258	.093	.048	.358	.123	.075
		MS[IF-SA(MTT)]	.263	.103	.033	.323	.153	.053
	40	C4-SA(MPT)	.320	.160	.053	.663	.400	.245
		MS[C4-SA(MPT)]	.320	.153	.048	.568	.300	.158
		IF-SA(MTT)	.630	.330	.198	.945	.810	.650
		MS[IF-SA(MTT)]	.608	.275	.100	.778	.465	.215

Observed Anchor Lengths for IF Selection Methods

The mean anchor lengths for the IF anchor selection methods are reported in Table 14 and the minimum anchor lengths are reported in Table 15. The maximum anchor length was 10 under all conditions. The mean anchor lengths were close to the anchor lengths used for the IF-DIF-free anchors and the minimum anchor lengths were within two items of the IF-DIF-free anchor lengths. The consistent maximum anchor length of 10 for all conditions was expected because the stopping criteria used in this study limited the maximum number of items to half of the number of items not displaying statistically significant DIF, rounded up the nearest whole number. Therefore, any replications identifying zero or one item with statistically significant DIF, leaving 20 or 19 items without statistically DIF, would have had a 10-item anchor.

Table 14

Mean anchor length by sample size, percentage of DIF, balance of DIF, and anchor method

			Balanced DIF			One-Sided DIF		
			Sample Size Per Group					
			500	750	1000	500	750	1000
Percentage of DIF		Anchor Method						
	0	IF-SA(MTT)	9.95	9.96	9.95	--	--	--
		MS[IF-SA(MTT)]	9.94	9.94	9.93	--	--	--
	10	IF-SA(MTT)	9.56	9.34	9.22	9.63	9.38	9.21
		MS[IF-SA(MTT)]	9.50	9.32	9.16	9.58	9.34	9.17
	20	IF-SA(MTT)	9.04	8.72	8.44	9.18	8.79	8.50
		MS[IF-SA(MTT)]	9.00	8.63	8.36	9.09	8.71	8.39
	40	IF-SA(MTT)	8.14	7.50	6.98	8.83	8.14	7.53
		MS[IF-SA(MTT)]	7.96	7.27	6.77	8.47	7.66	6.95

Table 15

Minimum anchor length by sample size, percentage of DIF, balance of DIF, and anchor method

			Balanced DIF			One-Sided DIF		
			Sample Size Per Group					
			500	750	1000	500	750	1000
Percentage of DIF		Anchor Method						
	0	IF-SA(MTT)	8	9	9	--	--	--
		MS[IF-SA(MTT)]	9	9	8	--	--	--
	10	IF-SA(MTT)	8	8	8	8	8	8
		MS[IF-SA(MTT)]	8	8	7	8	8	7
	20	IF-SA(MTT)	7	7	7	7	7	7
		MS[IF-SA(MTT)]	7	7	6	7	7	6
	40	IF-SA(MTT)	6	6	5	6	6	6
		MS[IF-SA(MTT)]	6	5	5	5	6	5

CHAPTER 5:

DISCUSSION AND CONCLUSIONS

Discussion

The purpose of this study was to determine if the proposed multistage methods would perform better than the methods proposed by Kopf et al. (2015b) in terms of true positive rates, false positive rates, familywise false positive rates, anchor contamination, and familywise anchor contamination. In total, four anchor selection methods were tested: C4-SA(MPT), MS[C4-SA(MPT)], IF-SA(MTT), and MS[IF-SA(MTT)]. For comparison purposes C4-DIF-free and IF-DIF-free anchors were also used. The study's hypotheses and research questions are displayed below and are followed by a discussion for each hypothesis and question.

Hypotheses

1. The multistage anchor selection methods will have higher true positive rates, lower false positive rates, lower familywise false positive rates, lower anchor contamination rates, and lower familywise anchor contamination than the non-multistage methods.
2. The anchor selection methods using IF will have higher true positive rates but also higher false positive rates than anchor selection methods using C4.
3. Familywise false positive rates will be greater than .05 for most, or all, conditions.

Questions

1. Will any of the studied methods result in DIF detection rates equal to the DIF detection rates for the DIF-free anchors for all conditions?

2. Will there be a difference in the anchor contamination rates between the IF and C4 methods?

Hypothesis 1

The multistage anchor selection methods will have higher true positive rates, lower false positive rates, lower familywise false positive rates, lower anchor contamination rates, and lower familywise anchor contamination rates than the non-multistage methods.

The first hypothesis was generally incorrect. The multistage methods only had higher true positive rates under the one-sided 40% DIF condition. The multistage methods had no impact on false positive rates, and only a small impact on familywise false positive rates under limited conditions. The multistage methods had lower anchor contamination and familywise anchor contamination rates under the one-sided 40% DIF condition. However, these differences were not large enough to lower false positive rates and only had a small impact on true positive and familywise false positive rates.

Hypothesis 2

The anchor selection methods using IF will have higher true positive rates but also higher false positive rates than anchor selection methods using C4.

The second hypothesis was partially correct. As expected, IF methods had higher true positive rates than C4 methods, although these differences were rather small. However, it was also expected that IF methods would have higher false positive rates. This was not observed. There were no differences in the false positive rates between IF methods and C4 methods.

The lack of difference in false positive rates between the IF anchor selection methods and C4 anchor selection methods was surprising because other studies had shown that longer anchors tend to produce higher false positive rates (Kopf et al., 2015b; Woods, 2009). However, in those

studies, the differences in false positive rates due to longer anchors was relatively small. Additionally, there are some differences between those studies and this study which make direct comparisons difficult.

Woods (2009) used a graded response model and a sample size of 1,500 for the reference group and 500 for the focal group. Kopf et al. (2015b) used stopping criteria for IF which resulted in a much longer anchor than the stopping criteria used in this study, and tested the items within the anchor for DIF, which was not done in this study. Woods used a program written in C++ to conduct her simulation study, while Kopf et al. (2015b) used R. Neither study reported details of their software such as the estimators used, so it is not possible to determine if the estimators used by PROC IRT were different than the estimators used in these other two studies, nor if any possible differences could have led to slight variations in the results. Given all the differences between those studies and this study and the small changes in false positive rates for IF methods that were observed in those studies, there was not concern for the lack of an observed difference in false positive rates between IF anchor selection methods and C4 anchor selection methods in this study.

Hypothesis 3

Familywise false positive rates will be greater than .05 for most, or all, conditions.

The third hypothesis was correct. Familywise false positive rates were well above .05 for all conditions. Interestingly, familywise false positive rates are not a huge area of research within the published DIF research. There are several applied DIF studies that use methods to control for multiple comparisons (Ballert, Post, Brinkhof & Reinhardt, 2015; Chen, Pan, Chung & Chen, 2015). However, there is a limited number of simulation studies exploring the impact of multiple comparisons on DIF detection (Kim, 2010; Kim & Oshima, 2012). One reason for this may be

that efforts to control for multiple comparisons often reduce true positive rates (Kim & Oshima, 2012). Due to the negative impact DIF can have on test score interpretations, it may be more acceptable for some researchers to over identify DIF rather than to under identify DIF. However, unnecessarily removing items can be expensive due to the cost to develop items, and it can lower test validity and reliability. A method for keeping familywise error rates within an acceptable range while maintaining a high level of true positive rates would be ideal and should be the subject of future research.

Question 1

Will any of the studied methods result in DIF detection rates equal to the DIF detection rates for the DIF-free anchors for all conditions?

Under most conditions all the anchor selection methods worked as well as the DIF-free anchors. However, the DIF-free anchors had higher true positive rates and lower false positive rates under the one-sided 40% DIF condition. Also, IF-DIF-free had slightly lower false positive rates and familywise false positive rates than other methods.

The similarities in anchor detection rates between the DIF-free anchors and the anchor selection methods under most conditions was similar to the results in Kopf et al.'s (2015b) study. Kopf et al. (2015b) found the largest differences between the DIF-free anchor and the anchor selection methods under the 40% one-sided DIF condition. Under all other conditions, Kopf et al. found small or no differences.

The lower familywise false positive rates for IF-DIF-free were likely due to the lower false positive rates for IF-DIF-free. However, the lower false positive rates for IF-DIF-free were not observed by Kopf et al. (2015b), and were somewhat concerning since they were not expected, although they were only about .01 lower than the false positive rates for other methods.

It is possible that the longer anchors for IF-DIF-free led to lower false positive rates when compared to the C4 methods. Additionally, IF-DIF-free always had uncontaminated anchors which may explain why IF-DIF-free had lower false positive rates than the IF anchors selection methods in the 10%, 20%, and 40% DIF conditions. However, the issue of anchor contamination does not explain why IF-DIF-free had lower false positive rates than the IF anchor selection methods in the 0% DIF condition.

It is possible that the reason for this observation may be explained by which items were selected for the anchor. IF-DIF-free randomly selected 10 items, while the IF anchor selection methods used an iterative procedure which selected the 10 items presumed to be the most likely to be DIF-free. However, during the data generation process, even items which were simulated to be DIF-free may occasionally have displayed some amount of DIF due to random variations in data generation. Using IF-DIF-free, those unintended DIF items had the same probability of being in the anchor as any other item. Once these items were in the anchor they were not being tested for DIF. However, using the IF anchor selection methods, those items should have been less likely to be included in the anchor which would mean they would have been tested for DIF and possibly shown statistically significant DIF. If this hypothesis was true, there should have been a decrease in the difference in familywise false positive rates between IF-DIF-free and the IF anchor selection methods as sample size increased because larger sample sizes would lead to item parameters generated closer to the intended parameters. Under the 0% DIF condition the difference in familywise false positive rates between IF-DIF-free and IF-SA(MTT) was .19 when sample sizes were 500 per group and decreased to .14 when sample sizes were 1000 per group. A similar trend was observed under all conditions. While these data do support the hypothesis explaining the lower false positive rates observed using the IF-DIF-free anchor, and the

hypothesis does provide a reasonable explanation for this observation, further study would be needed to support or refute this hypothesis.

It should also be noted that Kopf et al. (2015b) found differences in false positive rates between the DIF-free anchor and the anchor selection methods they tested. However, unlike this study, Kopf et al. (2015b) found differences in the C4 methods, not the IF methods. Under the 0% DIF condition, the DIF-free C4 method they tested had higher false positive rates than all the anchor selection methods they tested. The false positive rates for DIF-free C4 were about .05, and the false positive rates for all the anchor selection methods were about .03. Since all these anchor methods had equal anchor lengths and DIF-free anchors, the differences in false positive rates must be due to some other factor. However, like this study, what that factor is isn't clear, and further study would be needed to better understand these small differences.

Question 2

Will there be a difference in the anchor contamination rates between the IF and C4 methods?

There was generally not a difference in anchor contamination rates between anchor selection methods. When there were slight differences, as were observed in the one-sided 40% DIF condition, these differences did not appear to be large enough to impact false positive rates, and only had a slight impact on true positive rates and familywise false positive rates. However, it may also be true that the differences in anchor contamination rates were large enough to impact false positive rates, but that this study design did not provide enough power to observe those differences.

Limitations

This study is limited by the conditions of the simulation design. A limited number of sample sizes, percentages of DIF, and balances of DIF were used. Additionally, the test length,

model used to generate the data, model used to analyze the data, and software were all fixed. Changes to any of these conditions may produce different results; therefore, the conclusions and recommendations within this study are limited to the study conditions.

Conclusions

Overall, the application of the multistage anchor selection methods did not produce better results in DIF detection rates than the methods proposed by Kopf et al. (2015b). The multistage methods did reduce anchor contamination and slightly improve true positive rates in some conditions. However, these improvements were not large enough to be practically significant. Also, the multistage methods were more complicated to apply than the non-multistage methods. Given these two factors, the multistage methods are not recommended for use in DIF detection under the conditions simulated in this study.

The methods proposed by Kopf et al. (2015b) performed well overall. The DIF-free anchors only had better DIF detection rates than the anchor selection methods under the one-sided 40% DIF condition. However, even under this condition, the anchor selection methods performed reasonably well. Because there were not practical differences in the DIF detection rates between methods and C4-SA(MPT) is the easiest method to implement, C4-SA(MPT) is recommended for DIF detection. However, this recommendation is only applicable to the conditions simulated in this study which are limited. These conditions included data that fit the Rasch model, equal sample sizes between the reference and focal group, as well as a moderate amount of DIF of .4. Deviations from these conditions may produce better or worse DIF detection rates; therefore, practitioners should use caution when applying these techniques to datasets which have different parameters than the conditions simulated in this study.

Recommendations for Further Research

There are three recommendations related to this study that need further research. Additionally, a fourth recommendation about the reporting of simulation studies is included that could help future researchers. First, as Kopf et al. (2015b) found in their study, the DIF detection methods examined in this study had lower true positive rates than the DIF-free anchors under the most extreme DIF conditions. Also, the anchor selection methods had higher familywise false positive rates. Therefore, DIF detection rates could be improved if methods to better select DIF-free anchor items were developed. Second, it is unclear how much anchor contamination is needed to see significant changes in DIF detection rates. A simulation study exploring this issue would be beneficial. Third, familywise false positive rates were high for all anchor methods in this study. Identifying DIF detection methods that would control familywise false positive rates while maintaining high true positive rates would be useful. Fourth, when reporting simulation studies, it is recommended that researchers provide clear justifications for their study design decisions as well as clearly providing all details of their study design. As further explained below, several study design details were not reported in the published literature which would have been useful when designing this study.

Many simulation studies were reviewed while designing this project. However, justifications for the design of these studies were not always provided. For example, the number of replications in the reviewed simulation studies ranged from 100 to 2,000 with 100 being the most common number of replications. However, none of these studies provided justifications for the number of replications that were used. Providing justifications for study all design decisions, not just the number of replications, would be helpful both to other researchers designing

simulation studies as well as researchers evaluating the results of these published simulation studies.

Besides justifying study design decisions, it is also recommended that researchers clearly describe all aspects of their study design. Kopf et al. (2015b) stated that they used the software R to conduct their study. However, details such as the R package or the estimators used were not provided. Such details may be of use to other researchers who want to build off the work of a study as well as researchers who simply want to evaluate the study.

REFERENCES

AERA, APA, & NCME. (1999). *Standards for Educational and Psychological Testing*. Washington, DC.

Allalouf, A., (2003). Revising translated differential item functioning as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*(1), 55-73.

Apinyapibal, S., Lawthong, N., & Kanjanawasee, S. (2015). A comparative analysis of the efficacy of differential item functioning detection for dichotomously scored items among logistic regression, SIBTEST and Raschtree methods. *Procedia - Social and Behavioral Sciences, 191*, 21–25. http://doi.org/10.1016/j.sbspro.2015.04.664

Atar, B., & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *Hacettepe University Journal of Education*, (41), 36–47.

Ballert, C. S., Post, M. W., Brinkhof, M. W., & Reinhardt, J. D. (2015). Psychometric properties of the Nottwil Environmental Factors Inventory short form. *Archives of Physical Medicine and Rehabilitation, 96*. 233-240. http://dx.doi.org/10.1016/j.apmr.2014.09.004

Balluerka, N., Plewis, I., Gorostiaga, A., & Padilla, J. (2014). Examining sources of DIF in psychological and educational assessment using multilevel logistic regression. *Methodology, 10*(2), 71-79. http://dx.doi.org/10.1027/1614-2241/a000076

Benitez, I., Padilla, J., Montesisnos, M. D. H., & Sireci, S. G. (2016). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education, 29*(1), 1-16. http://dx.doi.org/10.1080/08957347.2015.1102915

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

Chen, Y. L., Pan, A. W., Chung, L., & Chen, T. J. (2015). Examining the validity and reliability of the Taita symptom checklist using Rasch analysis. *Journal of the Formosan Medical Association 114*. 221-230. http://dx.doi.org/10.1016/j.jfma.2013.10.004

Cho, S. Suh, Y. & Lee, W. (2016). After differential item functioning is detected: IRT item calibration and scoring in the presence of DIF. *Applied Psychological Measurement, 40*(8), 573-591. http://dx.doi.org/10.1177/0146621616664304

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335–350. http://doi.org/10.1177/014662169301700402

Croft, S. J., Roberts, M. A., & Stenhouse, V. L. (2016). The perfect storm of education reform: High-stakes testing and teacher evaluation. *Social Justice*, (1), 70.

Darling-Hammond, L., Adamson, F. (2013). Developing assessments of deeper learning: The costs and benefits of using tests that help students learn. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Fidalgo, A. M., Alavi, S. M., & Amirian, S. M. R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing, 31*(4), 433–451. http://doi.org/10.1177/0265532214526748

Fieo, R., Mukherjee, S., Dmitrieva, N. O., Fyffe, D. C., Gross, A. L., Sanders, E. R., … Gibbons, L. E. (2015). Differential item functioning due to cognitive status does not impact depressive symptom measures in four heterogeneous samples of older adults.

International Journal of Geriatric Psychiatry, *30*(9), 911–918. http://doi.org/10.1002/gps.4234

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. http://doi.org/10.1177/0146621605275728

Florida Department of Education (2011). Request for proposal (RFP) for discretionary, competitive project. Retrieved from: http://www.fldoe.org/core/fileparse.php/5648/urlt/0073522-doe905rttt-5-3-11.pdf

Florida Department of Education. (2016). Florida standards assessments 2015-2016: Volume 2 test development. Retrieved from http://fsassessments.org/wp-content/uploads/2016/04/V2_FSA-Technical-Report-Year-2015-2016_FINAL.pdf

González-Betanzos, F., & Abad, F. J. (2012). The effects of purification and the evaluation of differential item functioning with the likelihood ratio test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *8*(4), 134–145. http://doi.org/10.1027/1614-2241/a000046

Guilera, G., Gomez-Benito, J., Hidalgo, M., & Sanchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods*, *18*(4), 553–571.

Hidalgo, M. D., Galindo-Garre, F., & Gómez-Benito, J. (2015). Differential item functioning and cut-off scores: Implications for test score interpretation. *Anuario de Psicología*, *45*(1), 55–69.

Holland, P. W., & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Retrieved from http://eric.ed.gov/?id=ED272577

Huang, J. & Sheeran, T. J. (2011). Identifying causes of English-Chinese translation differential item functioning. *International Journal of Applied Educational Studies, 12*(1), 16-32.

Jarl, G., Heinemann, A. W., Lindner, H. Y., & Hermansson, L. M. N. (2015). Cross-cultural validity and differential item functioning of the orthotics and prosthetics users' survey with Swedish and United States users of lower-limb prosthesis. *Archives of Physical Medicine and Rehabilitation, 96*(9), 1615–1626. http://doi.org/10.1016/j.apmr.2015.03.003

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351), 631–639. http://doi.org/10.2307/2285946

Kabasakal, K. A., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Comparing performances (type I error and power) of IRT likelihood ratio SIBTEST and Mantel-Haenszel methods in the determination of differential item functioning. *Educational Sciences: Theory & Practice, 14*(6), 2186–2193. http://doi.org/10.12738/estp.2014.6.2165

Khalid, M. N., & Glas, C. A. W. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement, 50*, 186–197. http://doi.org/10.1016/j.measurement.2013.12.019

Kim, H. (2010). Controlling Type I error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing (dissertation). Georgia State University, Atlanta, GA.

Kim, J. & Oshima, T. C. (2012). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 73*(3), 458-470. http://dx.doi.org/10.1177/0013164412467033

Kopf, J., Zeileis, A., & Strobl, C. (2015a). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83–103. http://doi.org/10.1177/0146621614544195

Kopf, J., Zeileis, A., & Strobl, C. (2015b). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. http://doi.org/10.1177/0013164414529792

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica*, *30*(2), 343+.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology*, *96*(5), 927–940. http://doi.org/10.1037/a0023496

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031. http://doi.org/10.1037/a0027934

Michaelides, M. P. (2008). An Illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment, Research & Evaluation*, *13*, 1–16.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297–334. http://doi.org/10.1177/014662169301700401

Morris, T.L., Lee, Shing-Huei, L., & Barnes, L.L.B. (2008). The development and use of an instrument to measure willingness to seek help from peers and teachers when studying college mathematics. *Learning Environments Research, 11*, 227-243.

http://doi.org/10.1007/s10984-008-9046-3

Murray, A. L., Booth, T., & McKenzie, K. (2015). An analysis of differential item functioning by gender in the Learning Disability Screening Questionnaire (LDSQ). *Research in Developmental Disabilities*, *39*, 76–82. http://doi.org/10.1016/j.ridd.2014.12.006

Özdemir, B. (2015). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia - Social and Behavioral Sciences*, *174*, 2075–2083. http://doi.org/10.1016/j.sbspro.2015.02.004

Pae, T.-I., & Park, G.-P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, *23*(4), 475–496. http://doi.org/10.1191/0265532206lt338oa

Parslow, R. A., Christensen, H., Griffiths, K.M., & Groves, C. (2006). The Warpy Thoughts Scale: A new 20-item instrument to measure dysfunctional attitudes. *Cognitive Behavior Therapy*, *35*(2), 106-116. http://doi.org/10.1080/16506070500372279

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement*, *34*(6), 453–456. http://doi.org/10.1177/0146621610367789

Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In C. R. R. and S. Sinharay (Ed.), *Handbook of Statistics* (Vol. 26, pp. 125–167). Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/S016971610626005X

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502.

Ready, R. E., & Veague, H. B. (2014). Training in psychological assessment: Current practices of clinical psychology programs. *Professional Psychology: Research and Practice*, *45*(4), 278–282. http://doi.org/10.1037/a0037439

Rivas, G. E. L., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251–265. http://doi.org/10.1177/0146621608321760

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116. http://doi.org/10.1177/014662169301700201

Roth, D. L., Dilworth-Anderson, P., Jin Huang, Gross, A. L., & Gitlin, L. N. (2015). Positive aspects of family caregiving for dementia: Differential item functioning by race. *Journals of Gerontology: Series B: Psychological Sciences and Social Sciences*, *70*(6), 813–819.

Rudner, L. M. (2007). Implementing the Graduate Management Admission Test® computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*

Sekercioglu, G., Bayat, N., & Bakir, S. (2014). Psychometric properties of Science Items Comprehension Test. *Education and Science, 39*(176), 447-455. http://doi.org/10.15390/EB.2014.3692

Shaffer, C., & McCabe, S. (2013). Evaluating the predictive validity of preadmission academic criteria: High-stakes assessment. *Teaching and Learning in Nursing*, *8*(4), 157–161. http://doi.org/10.1016/j.teln.2013.07.005

Shih, C. L., Liu, T. H., & Wang, W. C. (2014). Controlling type I error rates in assessing DIF for logistic regression method combined with SIBTEST regression correction procedure and DIF-free-then-DIF strategy. *Educational and Psychological Measurement, 74*(6), 1018–1048. http://doi.org/10.1177/0013164413520545

Shih, C. L., & Wang, W. C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184–199. http://doi.org/10.1177/0146621608321758

Steinmayr, R., Bergold, S., Margraf-Stiksrud, J., & Freund, P. A. (2015). Gender differences on general knowledge tests: Are they due to differential item functioning? *Intelligence*, *50*, 164–174. http://doi.org/10.1016/j.intell.2015.04.001

Takane, Y., & Leeuw, J. de. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408. http://doi.org/10.1007/BF02294363

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. http://doi.org/10.1177/1094428114553062

U. S. Department of Education. (2014). State scope of work - Florida. Retrieved from: https://www2.ed.gov/programs/racetothetop/state-scope-of-work/florida.pdf

Wang, W. C. (2004). Effects of anchor items methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education, 72*(3), 221-261.

Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*(3), 166–180. http://doi.org/10.1177/0146621609355279

Wang, W. C., Shih, C. L., & Sun, G. W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*(4), 687–708. http://doi.org/10.1177/0013164411426157

Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*(5), 713–731. http://doi.org/10.1177/0013164409332228

Wang, W. C., & Su, Y. H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF Detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113–144. http://doi.org/10.1207/s15324818ame1702_2

Warne, R. T., Yoon, M., & Price, C. J. (2014). Exploring the various interpretations of "test bias." *Cultural Diversity and Ethnic Minority Psychology*, *20*(4), 570–582. http://doi.org/10.1037/a0036503

Woods, C. M. (2009). Empirical Selection of Anchors for Tests of Differential Item Functioning. *Applied Psychological Measurement*, *33*(1), 42–57. http://doi.org/10.1177/0146621607314044

Zhang, Y., Dorans, N. J., & Matthews-López, J. L. (2005). *Using DIF Dissection Method to Assess Effects of Item Deletion. Research Report No. 2005-10. ETS RR-05-23*. College Board. Retrieved from http://eric.ed.gov/?id=ED563094

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233–251.

APPENDIX A: CODE FOR DATA GENERATION

```
LIBNAME LIB '/folders/myfolders/Data';

data B_Temp_1;
      input B;
datalines;
-2.522
-1.902
-1.351
-1.092
-0.234
-0.317
 0.037
 0.268
-0.571
 0.317
 0.295
 0.778
 1.514
 1.744
 1.951
-1.152
-0.526
 1.104
 0.961
 1.314
-2.198
-1.621
-0.761
-1.179
-0.610
```

```
-0.291
 0.067
 0.706
-2.713
 0.213
 0.116
 0.273
 0.840
 0.745
 1.485
-1.208
 0.189
 0.345
 0.962
 1.592
;
run;
%macro DataGen(N_Ref, N_Foc, N_DIF, Balance);
%do Rep=1 %to 400;
data temp1;
      do Person=1 to &N_Ref+&N_Foc;
      if Person <= &N_Ref then Focal=0;
      if Person  > &N_Ref then Focal=1;
      if Focal=0 Then Theta= 0 + 1*rannor(&Rep);
      if Focal=1 Then Theta=-1 + 1*rannor(&Rep);
            do    Item=1 to 20;
            output;
            end;
      end;
run;

data B_Temp_2;
      set B_Temp_1;
      Rand=ranuni(&rep);
run;

proc sort data=B_Temp_2 out=B_Temp_3;
      by Rand;
run;
```

```
Data B_Temp_4;
      set B_Temp_3;
      Item=_N_;
      drop Rand;
      if Item < 21;
run;

proc sort data=temp1 out=temp2;
      by item;
run;

data temp3;
      merge temp2 B_Temp_4;
      by Item;
run;

data temp4;
      set temp3;
            DIF=0;
            if Focal=1 AND &Balance='O' AND Item <= &N_DIF then DIF=(0.4);
            if Focal=1 AND &Balance='B' AND Item <= &N_DIF AND MOD(Item,2)=1 then DIF=(-0.4);
            if Focal=1 AND &Balance='B' AND Item <= &N_DIF AND MOD(Item,2)=0 then DIF=(0.4);
            U=ranuni(&Rep);
            B_Parameter=B+DIF;
            Prob=exp(Theta-B_Parameter)/(1+exp(Theta-B_Parameter));
            R=0;
            if U < Prob THEN R=1;
run;

proc sort data=temp4 out=temp5;
      by person item;
run;

%let R=R;
%let F=F;
%let D=D;
%let B= %sysfunc(dequote(&Balance));
%let File=DataGen_;

data lib.&File.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&Rep.;
```

```
      set temp5;
run;

proc sort data=temp5 out=temp6;
      by Focal Person;
run;

proc transpose data=temp6 out=temp7 prefix=R;
      by Focal person;
      id item;
      var R;
run;

proc sort data=temp7 out=temp8;
      by person;
run;

data lib.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&Rep.;
      set temp8;
      drop _name_;
run;

%end;

%mend DataGen;

%DataGen(500, 500, 0, 'B');
%DataGen(750, 750, 0, 'B');
%DataGen(1000, 1000, 0, 'B');

%DataGen(500, 500, 2, 'B');
%DataGen(750, 750, 2, 'B');
%DataGen(1000, 1000, 2, 'B');
%DataGen(500, 500, 2, 'O');
%DataGen(750, 750, 2, 'O');
%DataGen(1000, 1000, 2, 'O');

%DataGen(500, 500, 4, 'B');
%DataGen(750, 750, 4, 'B');
%DataGen(1000, 1000, 4, 'B');
```

```
%DataGen(500, 500, 4, 'O');
%DataGen(750, 750, 4, 'O');
%DataGen(1000, 1000, 4, 'O');

%DataGen(500, 500, 8, 'B');
%DataGen(750, 750, 8, 'B');
%DataGen(1000, 1000, 8, 'B');
%DataGen(500, 500, 8, 'O');
%DataGen(750, 750, 8, 'O');
%DataGen(1000, 1000, 8, 'O');
```

APPENDIX B: CODE TO APPLY ANCHOR SELECTION METHODS

```
LIBNAME LIB '/folders/myfolders/Data';
LIBNAME LIBOUT '/folders/myfolders/Output';
options nonotes nosource nosource2 errors=1;
ods exclude all;
*****************************************************************************************;
*****************************************************************************************;
*This macro runs proc IRT using "Anc" anchors. The macro then splits the file by focal group
membership, merges the parameter estimates and standard errors, then calculates the Wald test
and p-value. The results from multiple iterations are saved to a file;
*****************************************************************************************;
*****************************************************************************************;
%macro IRT();
*Running IRT based on Anc anchors;
ods output ParameterEstimates=IRT_Out;
proc irt data=Data resfunc=OneP;
   var R1-R20;
   group focal;
   factor f1 -> R1-R20 = 20*1;
   mean f1;
   equality &Anc;
   fixvalue &Anc1/parm=[INTERCEPT] value=0;
run;
ods output close;
*Selecting b parameter estimates when Focal=0;
data IRT_Out_0;
      set IRT_Out;
      if Focal=0;
      if parameter="Difficulty";
      drop Probt Parameter Focal;
      rename Estimate=Estimate0_&i
```

```
                StdErr=StdErr0_&i;
run;
*Selecting b parameter estimates when Focal=1;
data IRT_Out_1;
      set IRT_Out;
      if Focal=1;
      if parameter="Difficulty";
      drop Probt Parameter Focal;
      rename Estimate=Estimate1_&i
                StdErr=StdErr1_&i;
run;
*Merging group estimates and completing Wald test;
data IRT_Out_All;
      MERGE IRT_Out_0
              IRT_Out_1;
      wald_&i = abs(Estimate1_&i-Estimate0_&i)/sqrt((StdErr0_&i*StdErr0_&i + StdErr1_&i*StdErr1_&i));
      pvalue_&i = 1 - probnorm(wald_&i);
      if Item="&AncID." then Wald_&i=.;
      if Item="&AncID." then pvalue_&i=.;
run;
*Sorting data so I can merge;
proc sort data=IRT_Out_All;
by item;
run;
*Merging data with other iterations;
data IRT_Temp;
      Merge IRT_Final IRT_Out_All;
      by Item;
run;
*Renaming results as Final so it can be merged with the next iteration;
data IRT_Final;
      set IRT_Temp;
run;
%mend;
*************************************************************************************;
*************************************************************************************;
*This macro calls the IRT macro and using each item as a SA.;
*************************************************************************************;
*************************************************************************************;
%macro SA(N_Ref, N_Foc, N_DIF, Balance, Pre, Rep);
```

```
*************************************************************************************************;
*DEFINING MACRO VARIABLES;
data null_;
      set Anchors;
      cnt = left(put(_n_,6.));
      call symput('Item'||cnt,Item);
      call symput('Count',cnt);
run;

proc sort data=Anchors;
      by Item;
run;
*************************************************************************************************;
*CREATING DATASET TO MERGE WALD AND PVALUES TO;
data IRT_Final;
      set Anchors;
      if Flag=1;
      Rand=ranuni(&Rep);
run;
*Sorting items so I can merge;
proc sort data=IRT_Final;
      by Item;
run;
*************************************************************************************************;
*RUNNING IRT MACRO FOR EACH ANCID
*Opening generated data;
data Data;
      set lib.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&Rep.;
run;
*Running macro;
%do i = 1 %to &COUNT;
      %let Anc = &&ITEM&i;
      %let AncID = &&ITEM&i;
      %let Anc1 = &&ITEM&i;
      %IRT;
%end;
*************************************************************************************************;
*CALCULATING MPT AND MTT RANKS;
*Calculating item-level mean pvalue and Wald test;
data SA_Temp;
```

```
      set IRT_Final;
      if Flag=1;
      PValue_Mean=mean(of pvalue_1-pvalue_&COUNT);
      Wald_Mean=mean(of Wald_1-Wald_&COUNT);
      match=1;
run;
*Ranking item level means;
Proc Rank data=SA_Temp Out=SA_Rank;
      var PValue_Mean Wald_Mean;
      ranks PValue_Mean_Rank Wald_Mean_Rank;
run;
*Selecting .5*K item level means to be MPT/MTT;
data SA_MPT;
      set SA_Rank;
      if PValue_Mean_Rank=round(.5*&Count);
      MPT=PValue_Mean;
      keep MPT match;
      match=1;
run;

data SA_MTT;
      set SA_Rank;
      if Wald_Mean_Rank=round(.5*&Count);
      MTT=Wald_Mean;
      keep MTT match;
      match=1;
run;
*Merging MPT/MTT to Final dataset;
data SA_Temp2;
      Merge SA_Temp SA_MTT SA_MPT;
      by Match;
run;
*Counting times above/below MPP/MPT;
*Adding ranuni to count so I can rank ties;
data SA_Temp3;
      set SA_Temp2;

      array Wald_{*} Wald_1-Wald_&Count;
            MTT_Count=0;
            do _n_=1 to dim(Wald_);
```

```
            if 0<Wald_{_n_}<MTT then MTT_Count+1;
      end;

      array PValue_{*}  PValue_1-PValue_&Count;
            MPT_Count=0;
            do _n_=1 to dim(PValue_);
            if 0<PValue_{_n_}>MPT then MPT_Count+1;
      end;

      MTT_Count2=MTT_Count+Rand;
      MPT_Count2=MPT_Count+Rand;
run;
*Ranking items by MPT and MTT;
proc rank data=SA_Temp3 Out=&Pre. descending;
      var MTT_Count2 MPT_Count2;
      ranks MTT_Rank MPT_Rank;
run;
%mend SA;
*************************************************************************************;
*************************************************************************************;
***This macro runs IF, MS[IF], C4, and MS[C4] anchor methods***;
*************************************************************************************;
*************************************************************************************;
%Macro IFC4(N_Ref, N_Foc, N_DIF, Balance, Pre);
%let R=R;
%let F=F;
%let D=D;
%let B=%sysfunc(dequote(&Balance));
*************************************************************************************;
      *Looping through simulated datasets with IF and then C4;
      %do RepLoop=1 %to 400;
            %let Stage=0;
            %let AncID_Prior=NONE;
            data AncID_IF_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
            run;
            *Creating initial anchor dataset since all items are used as SA.;
            data Anchors;
                  length Item $3;
                  do    Number=1 to 20;
                        Item = cat('R',Number);
```

```
                        Flag=1;
                        output;
                  end;
                  keep Item Flag;
            run;
            *Running Prelimnary SA;
            %SA(&N_Ref, &N_Foc, &N_DIF, &Balance, SA_0, &RepLoop);
            *Creating datasets to save ranks to;
            data Ranks_IF;
                  set SA_0;
                  MTT_Rank_0 = MTT_Rank;
                  keep Item MTT_Rank_0;
            run;
            data Ranks_C4;
                  set SA_0;
                  MPT_Rank_0 = MPT_Rank;
                  keep Item MPT_Rank_0;
            run;
*************************************************************************************;
            *Multi Stage IF Loop - repeates until AncID_Prior=AncID or 10 Stages;
            %do %until (&Stage=10);
*************************************************************************************;
                  *IF Loop;
                  %Let DIF_Free_Count=40;
                  %Let IF_Loop=1;
                  %Let i=1;
                  %do %until (&DIF_Free_Count <= (%eval(2*(&IF_Loop-1))));
*************************************************************************************;
                        *DEFINING ANCHOR;
                        *Selecting top IF_Loop ranked items for anchor;
                        data AnchorItem;
                              set SA_&Stage.;
                              if MTT_Rank <= &IF_Loop;
                              Keep Item;
                        run;
                        *Restructuring data;
                        proc transpose data=AnchorItem out=AnchorItem2;
                              var Item;
                        run;
                        *Creating variables Anchor and AnchorID which I will turn into macro variables;
```

```
data AnchorItem3;
      length AnchorID $120;
      set AnchorItem2;
      Anchor = catx(" ",of COL:);
      AnchorID = compress(cat(of COL:));
      Anchor1 = COL1;
run;
*Creating Anchor and AnchorID Macro Variables;
data AnchorItem4;
      set AnchorItem3;
      call symput('Anc',Anchor);
      call symput('AncID',AnchorID);
      call symput('Anc1',Anchor1);
run;
*************************************************************************************;
*CREATING DATASET TO TRACK IF ANCHORS;
data AnchorItem5;
      set AnchorItem4;
      Stage=%eval(&Stage.+1);
      IF_Loop=&IF_Loop;
      keep AnchorID Stage IF_Loop;
run;
*Merging new anchor id with final anchor id file;
data AnchorItem6;
      set AnchorItem5
      AncID_IF_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
run;
*Renaming final anchor id file which is needed for merge;
data AncID_IF_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
      set AnchorItem6;
run;
*************************************************************************************;
*RUNNING IRT MACRO WITH NEW ANCHOR;
*Creating Final dataset for IRT macro;
data IRT_Final;
      length Item $3;
      do    Number=1 to 20;
            Item = cat('R',Number);
            output;
      end;
```

```
                        keep Item;
                    run;

                    proc sort data=IRT_Final;
                        by Item;
                    run;

                    %IRT;

                    Data &Pre.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        set IRT_Final;
                        DIF=0;
                        if 0 <= pvalue_&i < .025 then DIF=1;
                    run;
*************************************************************************************;
                    *COUNTING NUMBER OF DIF ITEMS AND CREATING A MACRO VARIABLE;
                    data DIF_Free_Count;
                        merge Anchors &Pre.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        by item;
                        if DIF=0;
                        if flag=1;
                    run;

                    data DIF_Free_Count2;
                        set DIF_Free_Count;
                        cnt = left(put(_n_,6.));
                        call symput('DIF_Free_Count',cnt);
                    run;
*************************************************************************************;
                    *Calculating IF_Loop;
                    %let IF_Loop = %eval(&IF_Loop+1);
                %end;
*************************************************************************************;
                *SELECTING NON-DIF ITEMS FOR NEXT IF STAGE;
                data Anchors;
                    set A_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                    if DIF=0;
                    keep Item Flag;
                    Flag=1;
                run;
```

```
*************************************************************************************;
                  *SAVING DIF Test FOR IF-SA(MTT);
                  %if &Stage=0 %then %do;
                        Data LIBOUT.IF_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                              set A_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        run;
                  %end;
*************************************************************************************;
                  *Calculating Stage so that if AncID_Prior=AncID Stage=10 and the loop stops;
                  %if &AncID_Prior=&ANCID %then %let Stage = 10;
                  %else %let Stage = %eval(&Stage+1);
*************************************************************************************;
                  *RUNNING SA WITH NON_DIF ITEMS IF Stage < 10;
                  %if &Stage < 10 %then %do;
                        %let AncID_Prior=&ANCID;
                        %SA(&N_Ref, &N_Foc, &N_DIF, &Balance, SA_&Stage, &RepLoop);
*************************************************************************************;
                        *Saving ranks for each stage;
                        data Ranks_IF_2;
                              set SA_&Stage;
                              MTT_Rank_&Stage.=MTT_Rank;
                              keep Item MTT_Rank_&Stage.;
                        run;
                        data Ranks_IF_3;
                              merge Ranks_IF_2 Ranks_IF;
                              by Item;
                        run;
                        data Ranks_IF;
                              set Ranks_IF_3;
                        run;
                  %end;
*************************************************************************************;
                  *SAVING DIF Test FOR MS[C4-SA(MPT)];
                  %if &Stage=10 %then %do;
                        Data LIBOUT.MSIF_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                              set A_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        run;
                        data LIBOUT.AncID_IF_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                              set AncID_IF_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        run;
```

```
                        data LIBOUT.Ranks_IF_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                              set Ranks_IF;
                        run;
                  %end;
            %end;
*************************************************************************************;
*************************************************************************************;
*************************************************************************************;
            *Defining variables for Multistage C4 Loop;
            %let n=0;
            %let AncID_Prior=NONE;
            data AncID_C4_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
            run;
*************************************************************************************;
            *Running Multistage C4 loop;
            %do %until (&n=10);
*************************************************************************************;
                  *DEFINING ANCHOR;
                  *Selecting top 4 ranked items for anchor;
                  data AnchorItem;
                        set SA_&N.;
                        if MPT_Rank < 5;
                        Keep Item;
                  run;
                  *Restructuring data;
                  proc transpose data=AnchorItem out=AnchorItem2;
                        var Item;
                  run;
                  *Creating variables Anchor and AnchorID which I will turn into macro variables;
                  data AnchorItem3;
                        set AnchorItem2;
                        Anchor = catx(" ",of COL:);
                        AnchorID = compress(cat(of COL:));
                        Anchor1 = COL1;
                  run;
                  *Creating Anchor and AnchorID Macro Varaibles;
                  data AnchorItem4;
                        set AnchorItem3;
                        call symput('Anc',Anchor);
                        call symput('AncID',AnchorID);
```

```
                  call symput('Anc1',Anchor1);
            run;
*************************************************************************************;
            *CREATING DATASET TO TRACK C4 ANCHORS;
            data AnchorItem5;
                  set AnchorItem4;
                  Stage=%eval(&N.+1);
                  keep AnchorID Stage;
            run;
            *Merging new anchor id with final anchor id file;
            data AnchorItem6;
                  set AnchorItem5
                  AncID_C4_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
            run;
            *Renaming final anchor id file which is needed for merge;
            data AncID_C4_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                  set AnchorItem6;
            run;
*************************************************************************************;
            *RUNNING IRT MACRO WITH NEW ANCHOR;
            *Creating Final dataset for IRT macro;
            data IRT_Final;
                  length Item $3;
                  do    Number=1 to 20;
                        Item = cat('R',Number);
                        output;
                  end;
                  keep Item;
            run;

            proc sort data=IRT_Final;
                  by Item;
            run;

            %Let i=1;
            %IRT;

            Data &Pre.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                  set IRT_Final;
                  DIF=0;
```

```
                        if 0 <= pvalue_&i < .025 then DIF=1;
                  run;
*************************************************************************************;
                  *SELECTING NON-DIF ITEMS FOR NEXT STAGE;
                  data Anchors;
                        set A_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        if DIF=0;
                        keep Item Flag;
                        Flag=1;
                  run;
*************************************************************************************;
                  *SAVING DIF Test FOR C4-SA(MPT);
                  %if &n=0 %then %do;
                        Data LIBOUT.C4_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                              set A_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        run;
                  %end;
*************************************************************************************;
                  *Reculating Stage so that if AncID_Prior=AncID Stage=10 and the loop stops;
                  %if &AncID_Prior=&ANCID %then %let n=10;
                  %else %let n=%eval(&n+1);
*************************************************************************************;
                  *RUNNING SA WITH NON_DIF ITEMS IF ANCID_PRIOR NE ANCID;
                  %if &n < 10 %then %do;
                        %let AncID_Prior=&ANCID;
                        %SA(&N_Ref, &N_Foc, &N_DIF, &Balance, SA_&n, &RepLoop)
*************************************************************************************;
                        *Saving ranks for each stage;
                        data Ranks_C4_2;
                              set SA_&n;
                              MPT_Rank_&n.=MPT_Rank;
                              keep Item MPT_Rank_&n.;
                        run;
                        data Ranks_C4_3;
                              merge Ranks_C4_2 Ranks_C4;
                              by Item;
                        run;
                        data Ranks_C4;
                              set Ranks_C4_3;
                        run;
```

```
                  %end;
*************************************************************************************;
                  *SAVING DIF Test FOR MS[C4-SA(MPT)];
*************************************************************************************;
                  %if &n=10 %then %do;
                        Data LIBOUT.MSC4_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                              set A_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        run;
                        data LIBOUT.AncID_C4_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                              set AncID_C4_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                        run;
                        data LIBOUT.Ranks_C4_&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&RepLoop.;
                              set Ranks_C4;
                        run;
                  %end;
            %end;
            proc datasets library=work kill;
            run;
            quit;
      %end;
%Mend;

%IFC4(500, 500, 0, 'B', A_);
%IFC4(750, 750, 0, 'B', A_);
%IFC4(1000, 1000, 0, 'B', A_);

%IFC4(500, 500, 2, 'B', A_);
%IFC4(750, 750, 2, 'B', A_);
%IFC4(1000, 1000, 2, 'B', A_);
%IFC4(500, 500, 2, 'O', A_);
%IFC4(750, 750, 2, 'O', A_);
%IFC4(1000, 1000, 2, 'O', A_);

%IFC4(500, 500, 4, 'B', A_);
%IFC4(750, 750, 4, 'B', A_);
%IFC4(1000, 1000, 4, 'B', A_);
%IFC4(500, 500, 4, 'O', A_);
%IFC4(750, 750, 4, 'O', A_);
%IFC4(1000, 1000, 4, 'O', A_);
```

```
%IFC4(500, 500, 8, 'B', A_);
%IFC4(750, 750, 8, 'B', A_);
%IFC4(1000, 1000, 8, 'B', A_);
%IFC4(500, 500, 8, 'O', A_);
%IFC4(750, 750, 8, 'O', A_);
%IFC4(1000, 1000, 8, 'O', A_);
```

APPENDIX C: CODE TO APPLY DIF-FREE ANCHORS

```
options nonotes nosource nosource2 errors=1;
*Options macrogen mlogic symbolgen notes;
ods exclude all;
LIBNAME LIB '/folders/myfolders/Data';
LIBNAME LIBOUT '/folders/myfolders/Output';
*************************************************************************************;
*************************************************************************************;
%macro Perfect();
*Running IRT based on Anc anchors;
ods output ParameterEstimates=IRT_Out;
proc irt data=Data resfunc=OneP;
   var R1-R20;
   group focal;
   factor f1 -> R1-R20 = 20*1;
   mean f1;
   equality &Anc1-R20;
   fixvalue R20/parm=[INTERCEPT] value=0;
run;
ods output close;
*Selecting b parameter estimates when Focal=0;
data IRT_Out_0;
      set IRT_Out;
      if Focal=0;
      if parameter="Difficulty";
      drop Probt Parameter Focal;
      rename Estimate=Estimate0
             StdErr=StdErr0;
run;
*Selecting b parameter estimates when Focal=1;
data IRT_Out_1;
```

```
      set IRT_Out;
      if Focal=1;
      if parameter="Difficulty";
      drop Probt Parameter Focal;
      rename Estimate=Estimate1
               StdErr=StdErr1;
run;
*Merging group estimates and completing Wald test;
data IRT_Out_All;
      MERGE IRT_Out_0
              IRT_Out_1;
      wald = abs(Estimate1-Estimate0)/sqrt((StdErr0*StdErr0 + StdErr1*StdErr1));
      pvalue = 1 - probnorm(wald);
      DIF=0;
      if 0 <= pvalue < .025 then DIF=1;
run;
*Saving output;
data LIBOUT.&Pre.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&Rep.;
      set IRT_Out_All;
run;
%mend Perfect;
*************************************************************************************;

%macro RunPerfect(N_Ref, N_Foc, N_DIF, Balance);
%let R=R;
%let F=F;
%let D=D;
%let B=%sysfunc(dequote(&Balance));
%do Rep=1 %to 400;
%let Anc1=R17;
%let Pre=PerC4_;
data Data;
      set LIB.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&Rep.;
run;
%Perfect;
%end;
************************************************;
*Manually change Anc1 depending on percentage of DIF;
************************************************;
%do Rep=1 %to 400;
```

```
%let Anc1=R13;
%let Pre=PerIF_;
data Data;
      set LIB.&R.&N_Ref.&F.&N_Foc.&D.&N_DIF.&B.&Rep.;
run;
%Perfect;
%end;
%Mend RunPerfect;

%RunPerfect(500, 500, 0, 'B');
%RunPerfect(750, 750, 0, 'B');
%RunPerfect(1000, 1000, 0, 'B');

%RunPerfect(500, 500, 2, 'O');
%RunPerfect(750, 750, 2, 'O');
%RunPerfect(1000, 1000, 2, 'O');
%RunPerfect(500, 500, 2, 'B');
%RunPerfect(750, 750, 2, 'B');
%RunPerfect(1000, 1000, 2, 'B');

%RunPerfect(500, 500, 4, 'O');
%RunPerfect(750, 750, 4, 'O');
%RunPerfect(1000, 1000, 4, 'O');
%RunPerfect(500, 500, 4, 'B');
%RunPerfect(750, 750, 4, 'B');
%RunPerfect(1000, 1000, 4, 'B');

%RunPerfect(500, 500, 8, 'O');
%RunPerfect(750, 750, 8, 'O');
%RunPerfect(1000, 1000, 8, 'O');
%RunPerfect(500, 500, 8, 'B');
%RunPerfect(750, 750, 8, 'B');
%RunPerfect(1000, 1000, 8, 'B');
```

APPENDIX D: IRB EXEMPTION LETTER

USF
UNIVERSITY OF
SOUTH FLORIDA

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-7091

4/5/2017

Brandon Craig
L-CACHE - Leadership, Counseling, Adult, Career & Higher Education
Tampa, FL 33612

RE: **Not Human Subjects Research Determination**
IRB#: Pro00029209
Title: The Empirical Selection of Anchor Items Using a Multistage Approach

Dear Mr. Craig:

The Institutional Review Board (IRB) has reviewed your application and determined the activities do not meet the definition of human subjects research. Therefore, this project is not under the purview of the USF IRB and approval is not required. If the scope of your project changes in the future, please contact the IRB for further guidance.

All research activities, regardless of the level of IRB oversight, must be conducted in a manner that is consistent with the ethical principles of your profession. Please note that there may be requirements under the HIPAA Privacy Rule that apply to the information/data you will utilize. For further information, please contact a HIPAA Program administrator at 813-974-5638.

We appreciate your dedication to the ethical conduct of research at the University of South Florida. If you have any questions regarding this matter, please call 813-974-5638. Sincerely,

John A Schinka, Ph.D.

John Schinka, Ph.D., Chairperson
USF Institutional Review Board

ABOUT THE AUTHOR

Brandon D. Craig has earned a B.A. in Interdisciplinary Social Science from the University of South Florida, a M.F.A in Children's Literature from Simmons College, a M.P.H. in Biostatistics from the University of South Florida, and a Ph.D. in Educational Measurement and Research from the University of South Florida. Previously, he taught fourth grade in Baltimore and was a child welfare case manager in Florida. For the past five years he has worked in the assessment and accountability offices of two large school districts, working first for Hillsborough County Public Schools and then for Polk County Public Schools.