December 2014

Dissecting the Impact of DIF/DBF on Ability Estimation and Person Fit

Kevin Cappaert
University of Wisconsin-Milwaukee

DISSECTING THE IMPACT OF DIF/DBF ON ABILITY ESTIMATION AND PERSON FIT

by

Kevin J Cappaert

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Educational Psychology

at

The University of Wisconsin-Milwaukee

December 2014

ABSTRACT

DISSECTING THE IMPACT OF DIF/DBF ON
ABILITY ESTIMATION AND PERSON FIT

by

Kevin J Cappaert

The University of Wisconsin-Milwaukee, 2014
Under the Supervision of Professor Cindy Walker

Prior research has shown that differential item functioning (DIF) and differential bundle functioning (DBF) can influence ability estimation in unidimensional item repose theory (IRT); however, the relationship between ability estimation and uniform and non-uniform DIF/DBF has not been thoroughly investigated. Therefore, a simulation study was conducted to more thoroughly investigate how DIF/DBF and other related factors influence ability estimation in IRT. The factors examined included bundle size, the sum of uniform DIF in a bundle, magnitude of non-uniform DIF in each item in a bundle, balance of reference and focal group examinees, test length, and impact. Results indicated that an increase in uniform DIF/DBF leads to positive ability estimation bias for reference group examinees ability estimates. The magnitude of non-uniform DIF/DBF was found to influence the root mean squared error (RMSE) of ability estimates and standard error of the estimates. Specifically, lower RMSE and lower standard errors were obtained when items were simulated to be more discriminating for the reference group. Rank order correlations between true and estimated ability were found to be highly consistent regardless of the magnitude of uniform and non-uniform DIF/DBF in the

bundle. Finally, Crossing SIBTEST was found to provide acceptable type-I error rates and power when uniform DBF was simulated.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

CHAPTER 1

INTRODUCTION

Differential item functioning (DIF) has long been a useful tool to assess the degree of validity in a test item by determining if one group has a greater probability of answering an item correctly than another group, after conditioning on ability (Roussos & Stout, 1996). DIF is not to be confused with *impact*, or an occurrence when differences between mean group abilities are expected because there is a true difference between the abilities of both groups (Clauser & Mazor, 1998). DIF can manifest itself in both uniform and non-uniform (also referred to as crossing DIF) ways. In an IRT framework uniform DIF occurs when members of separate groups have the same level of a given trait ability, but the item difficulty is not equal for the members of each of those groups meaning the item is harder for one group than the other. Non-uniform DIF, on the other hand, occurs when the ability to discriminate between people with similar trait abilities is not the same for both groups.

Differential bundle functioning (DBF) is an extension of DIF in which multiple items work in concert to function differentially for different groups after controlling for the ability of interest (Douglas, Roussos, & Stout, 1996). To date, the understanding of the influence of DIF/DBF in the estimation of test performance and ability estimation has only been investigated in a limited number of studies (e.g. Roznowski & Reith, 1999; Takala & Kaftandjieva, 2000; Zumbo, 2003; Pae & Park, 2006; Walker, Zhang, Banks, & Cappaert; 2012). More importantly, there has only been one study (Walker et al., 2012) that has directly investigated the effect of DIF/DBF on ability estimation; however, this study did not investigate the influence of non-uniform DIF nor the commonly associated

influence of impact and unbalanced reference/focal group samples sizes on ability estimation. Since DIF/DBF is considered a threat to validity, it is paramount to determine what factors related to DIF/DBF actually lead to bias in ability estimation, and to what degree. This paper serves three purposes. First, many different factors related to DIF/DBF (such as the magnitude of DIF/DBF, impact, and bundle size) need to be investigated to determine which factors, or combination thereof, are more likely to result in ability estimation bias. Next, the influence of DIF/DBF on person-fit will be investigated. Lastly, consideration needs to be given to the method of DIF/DBF detection after determining which factors influence ability estimation.

There are many factors related to DIF that may impact ability estimation. DIF/DFB is expected to result in item parameter differences between groups of individuals because the incorrect model, in this case a unidimensional IRT model, is used to estimate the data. Though the influence of uniform DIF/DBF on ability estimation in a unidimensional IRT model has been investigated, many other associated DIF/DBF concepts have not been studied. One such factor is impact. When testing for DIF/DBF it is often found that impact is present in the data. The relationship between the presence of impact and ability estimation needs to be thoroughly investigated because impact results in two unique distributions. Specifically, impact results in different underlying ability distributions for the reference and focal group (Ackerman, 1992). It is unknown what influence these distinct reference and focal group distributions have on ability estimation in the presence of DIF/DBF. Secondly, careful consideration needs to be given to the difference between a balanced sample size, or equal numbers of examinees in the reference and focal groups, and an unbalanced sample size, or an unequal number of

examinees in the reference and focal groups on ability estimation, when controlling for the total sample size. Also of consideration are the type of DIF, either uniform or non-uniform, and the magnitude of DIF in each item. Specifically, it is believed that non-uniform DIF will result in a uniquely detrimental influence on ability estimation because items with low discrimination contain less information than items with high levels of discrimination (Hambleton & Swaminathan, 1984). Therefore, less information would be obtained from the group for which the item is not discriminating as well, resulting in a greater standard error of the ability estimate. Lastly, total test length will be considered. In a previous study, Walker et al. (2012) found that a greater proportion of DIF/DBF coupled with a shorter test length resulted in greater ability estimation bias than a longer test with a lesser proportion of DIF/DBF. Since there has been such little research in this topic, replication of these findings would help to cement the conclusions drawn by Walker et al. (2012).

This investigation will also consider design factors that are specific to DBF because of the complex nature of bundles in relation to single item DIF. For instance, bundles require the consideration of bundle size. Walker et al. (2012) investigated the influence of bundle size with 3 and 5 item bundles and found that as the sum of DBF increased, the differences in ability estimates was greater with a 5 item bundle, as compared to a 3 item bundle. However, further research is needed to determine the degree to which bias is related to the proportion of items on the overall test, as opposed to the number of items in a bundle. The current study will propose a method to manipulate both test length and bundle size to make this determination.

The relationship between ability estimation and DBF requires careful consideration because bundles can function solely against the focal group or against both the focal and reference group. Nandakumar (1993) demonstrated that the presence of amplification, which occurs when multiple items in a bundle all function against the same group, leads to increased power to detect DIF, and the presence of cancellation, which occurs when individual items function against both the focal and reference groups at equal levels, cancels out the overall DIF effect at the test level. It wasn't until recently, when Walker et al. investigated the influence of amplification and cancellation on ability estimation, that it was determined that a greater sum of DIF results in positive ability estimation bias for reference group examinees and negative estimation bias for the focal group examinees. Though investigations by Walker et al. gave some insight into this relationship, the influence of amplification needs to be examined from a broader perspective, and, at greater magnitudes than previously investigated, for a better understanding.

Thirdly, the influence of DIF/DBF on item person fit will be investigated to better explain the relationship between DIF/DBF and ability estimation. Person-fit-statistics have been developed to identify examinees with aberrant item response patterns which lead to spuriously high or spuriously low ability estimates due to factors such as cheating, careless responding, guessing, creative responding, and random responding (Karabastos, 2003). Person-fit-statistics measure how well a response pattern matches the model being used to estimate ability, so poor person-fit results in inaccurate estimation for that individual. Since DIF involves differential difficulty and/or discrimination parameters for reference and focal group examinees, estimated ability is believed to lead to person-

misfit. The standardized log-likelihood index (l_z: Drasgow, Levine, & Williams, 1985) will be used to assess person fit as it has been regarded as one of the most commonly used person fit statistics and can be used with the 2-PL and 3-PL models (Meijer & Sijtsma, 2001, Rupp, 2013).

Lastly, there are only a few procedures which assess both DIF and DBF. The most commonly used procedures are SIBTEST (Shealy & Stout, 1993), the multiple indicator multiple cause model (MIMIC) (Muthen & Lehman, 1985; Finch, 2005; Woods and Grimm, 2011), and DFIT (Oshima, Raju, Flowers, & Slinde, 1998; Oshima, Raju, & Nanda, 2006). These three methods conduct a prioi hypothesis tests for items and/or bundles to detect uniform DIF/DBF. SIBTEST and the MIMIC model have been used to detect uniform DIF/DBF, and the MIMIC model technique has been expanded to detect crossing DIF (Woods and Grimm, 2011), while Crossing SIBTEST (Li and Stout, 1996) has been used to detect crossing DIF/DBF. Though MIMIC model invariance techniques have been shown to be promising for testing DBF, especially in instances of impact (Finch, 2012), SIBTEST will be used in the current study to detect DIF/DBF because research has not been conducted to formally test the capability of detection using a MIMIC model for crossing DBF. In addition results from Woods and Grimm (2011) found high type I error with single item non-uniform DIF detection using the MIMIC model. Comparison of DFIT to SIBTEST and the MIMIC model will not be completed for this investigation because of findings by Russell (2005) that showed type-I error inflation with the presence of impact with DFIT as well as lower power. Therefore, the relationship between the magnitude of the DIF/DBF effect and estimation bias estimates will be investigated for SIBTEST and Crossing SIBTEST only.

Currently when DIF is detected, even if an effect size measure is available, it is not evident when detectable DIF will actually influence the estimation of ability to a practical degree. By investigating many of the variations of DIF/DBF (e.g. varying bundle size, test length, etc.) in relation to the amount of ability estimation bias resulting from those variations, it should be possible to outline effect size guidelines which can be used by practitioners.

CHAPTER 2

LITERATURE REVIEW

The IRT Model and IRT Assumptions

Two commonly used approaches in educational measurement are classical test theory (CTT) and item response theory (IRT). CTT models the performance of an individual test taker as a function of their true score plus error, or $X = T + E$. On the other hand, IRT is a measurement paradigm that focuses on item performance, rather than overall test performance. The IRT paradigm positions persons and items on the same latent trait, or ability, continuum. IRT models both item and person parameters on the same scale by estimating a probability function of a correct response to an item given an ability level and item characteristics. Therefore, the model for IRT assesses the following:

$$\boldsymbol{P_i(\theta) = P_i(X_i = 1|\theta, a, b, c)} \quad (1)$$

which states that the probability of an examinee responding correctly to item X_i depends on both the examinee's ability level and the item parameters, in this case a, b, c (Embretson & Reise, 2000). To give an example, in an exam estimating mathematics knowledge, responses to an individual mathematics item rely on both the item characteristics (e.g. how difficult/ discriminating an item is) and the ability level of an individual.

Item response function.

Taking this one step further, IRT provides individual item response functions (IRF) for each item. An example of this is depicted in Figure 1. The function for the 3 parameter logistic model (3 PL) is:

$$\boldsymbol{P_i(X_i = 1|\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{-Da_i(\theta - b_i)}}} \quad (2)$$

where b_i is the item difficulty parameter, or the point at which an individual has an equal probability of getting an item correct or incorrect, which is also known as the point of inflection (Hambleton & Swaminathan, 1984). In Figure 1, the inflection point is at 0 indicating a person with ability, or θ, equal to 0 has a 50 percent chance of answering the item correctly, and consequently a 50 percent chance of answering the item incorrectly. The item discrimination parameter, represented by a_i , is the slope of the function at the point of inflection. The discrimination parameter gives an indication of how well an item can differentiate between different ability levels across the continuum (deAyala, 2009). As the slope steepens at the inflection point, the ability to discriminate between two individuals with different trait abilities also increases. The c_i parameter is a lower bound for the function, often referred to as the pseudo-guessing parameter or lower asymptote. This parameter gives the chance probability of getting an item correct for an individual with low ability; however, this chance is often lower than chance alone, because item writers develop distractors which vary in their degree of attractiveness to people at different ability levels along the continuum (De Ayala, 2009; Hambleton & Swaminathan, 1984; Lord, 1974). The pseudo-guessing parameter was set at 0 in Figure 1; however, when the pseudo-guessing parameter is greater than 0, the point of inflection increases to (1 + c)/2 (Hambleton & Swaminathan, 1984). Lastly, *D* is simply a correction factor to minimize the difference between the normal ogive and logistic functions.

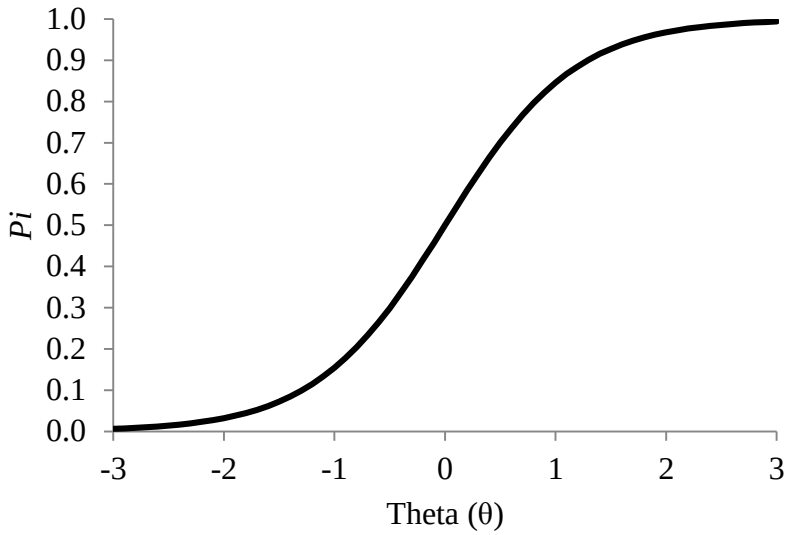

Figure 1: The probability of answering item i correctly given θ (a given ability level).

When a pseudo guessing parameter is not included in the model, and thus c is replaced with 0, the model reduces to the 2 PL which is depicted in Equation 3:

$$\boldsymbol{P_i(X_i = 1|\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{-Da_i(\theta - b_i)}}} \quad (3)$$

where *a*, *b,* and *D* are the same as in Equation 2. Taken one step further, if all of the discrimination parameters are equivalent then the model reduces further into the 1 PL which appears similar to equation 3 except the discrimination parameter (*a*) is fixed, so that the subscript is dropped from the equation. As a special case of the 1 PL, if the discrimination parameters are all equal to 1 the model reduces to the Rasch model which is depicted in Equation 4:

$$P_i(X_i = 1|\theta) = \frac{e^{D(\theta - b_i)}}{1+e^{(\theta - b_i)}} \quad (4)$$

where *b* and *D* are the same as in Equation 2.

Descriptions of the item information, test information function, and standard error of measurement will be based on the 2-PL model due to the direct link between the

parameters of these functions and the two types of DIF considered in this study, which will be discussed later.

Item and test information.

Item information is directly related to the accuracy of ability estimation, where greater item information yields greater accuracy. The amount of information can be calculated for an individual item as well as for the entire test. The item information functions for the three logistic item response models for dichotomous items are depicted in Table 1:

Table 1

Item Information Functions for 1, 2, and 3 PL IRT Models

Model	Item Information
1 PL	$P_i(\theta)Q_i(\theta)$
2 PL	$a_i^2 P_i(\theta)Q_i(\theta)$
3 PL	$\left[a_i^2 \frac{1-P_i(\theta)}{P_i(\theta)}\right]\left[\frac{(P_i(\theta)-c_i)^2}{(1-c_i)^2}\right]$

where P_i is the conditional probability of a correct response along the continuum, and Q_i is the conditional probability of an incorrect response (Embretson & Reise, 2000). Because the upper bound of a probability is 1, Q_i is equal to (1 - P_i). All three functions calculate information along the θ continuum. Of importance to ability estimation is the change from the 1PL to the 2PL because this change introduces a squared discrimination parameter. The addition of the discrimination parameter to item information becomes increasingly influential as the discrimination parameter gets larger. Thus, items with discrimination parameters greater than 1 contain much more information at a given

location close to the item difficulty parameter than an item with a discrimination parameter less than 1.

The test information function is equal to the sum of the item information functions (Equation 5).

$$TI(\theta) = \sum_{i=1}^{I} I(\theta) \quad (5)$$

The more information at a given θ, the more accurate the estimates of one's θ will be (de Ayala, 2009); therefore, a large threat to the accuracy of ability estimates, according to the information in an item, is the discrimination parameter. As will be discussed later, it can then be expected that biased estimates of θ should result if two groups are estimated with different discrimination parameters.

The standard error of measurement (SE) is directly related to the test information and is represented in Equation 6:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}}. \quad (6)$$

The function for standard error indicates that the greater the test information, the lower the SE is, given a location on the θ distribution. Because the function is conditional on θ, different θ locations on the distribution will result in different SEs. The magnitude of the standard error depends on (1) the number of test items, where a greater number of test items yields a lower standard error, (2) the quality of test items, where more discriminating items lead to smaller standard errors, and (3) the match between item difficulty and examinee ability, where smaller standard errors are found in instances where difficulty parameters are equal to one's respective ability (Hambleton, Swaminathan, & Rogers, 1991).

Assumptions.

The IRT model relies on certain assumptions. The assumptions in IRT are: unidimensionality, local independence, and monotonicity of the IRT function (de Ayala, 2009). The first assumption is unidimensionality which means that the trait being measured falls on only one continuum, or one dimension. This means that the response on an item is solely due to the level of the single trait being measured. A unidimensional trait should also display properties of parameter invariance, which means that for any sample of individuals, the IRF will be the same, up to a linear transformation. Unlike in CTT, where the item difficulty and item discrimination parameters change depending on which sample is being used, item parameters in IRT are invariant of the sample used to estimate them, up to a linear transformation in a given IRT model (Rupp & Zumbo, 2006). Put another way it should not matter which sample is being used to generate the item parameters, as long as the samples are equated. Even though individual samples of respondents may have different ability distributions, the item parameters are unaffected by the sample itself. Therefore, one big advantage of IRT, over CTT, is that once item parameters are known they can be used in subsequent tests without re-estimation.

The assumption of unidimensionality is important to IRT ability estimation because a violation of the unidimensionality assumption indicates that an item is no longer measuring only the primary trait of interest, but rather multiple traits, which cannot be modeled with a unidimensional IRT model.

The second assumption is that of local independence (LI), which is actually a sub-assumption under unidimensionality. IRT is a probabilistic function of a response to an item that is dependent on one's ability and the item characteristics. The assumption of

local independence states that the response to one item in a measure does not influence the response on any other item in the measure, conditioned on the latent trait(s) being measured, so conditional item responses are, therefore, statistically independent (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1984). This means that item responses are independent of each other when they are conditioned on the latent trait being measured, so a violation of this assumption, or local dependence, would be indicative that the response to one item depends on the response to another item. Local dependence results in biased parameter estimation because it implies that an additional dimension is being measured by a test (Wainer, Bradlow, & Wang, 2007). A testlet is defined as group of related items to a single content area which is developed as a unit in order to measure a content areas such as reading comprehension (Wainer & Kiely, 1987). Continuing with the same example, a testlet with multiple items relating to a single reading passage may lead to a violation of local independence because items from a common passage may not be independent of each other; they are actually directly related to each other because they require the reading of the same passage and thus the same content domain. Depending on the content domain of the testlet passage, some individuals may have greater interest or prior knowledge in the topic which may lead to a more understandable passage for some examinees in relation to others. It has been recommended that the influence of a testlet can be controlled through balancing within a content area and across an entire test (Wainer, Sireci, & Thissen, 1991).

Another assumption is that the IRF is the general nature of the function. The IRF is a probabilistic monotonically increasing logistic function in which individuals with more of the trait ability being measured have a greater probability of endorsing the

correct response to an item (Hambleton, Swaminathan, & Rogers, 1991). Therefore, items that have discrimination parameters that are less than or equal to 0 result in a direct violation of this assumption.

Differential item functioning (DIF) can be thought of as a violation of the unidimensionality assumption in IRT. Theoretically, DIF occurs when one fits a unidimensional IRT model to test data when, in fact, multiple dimensions are being measured by test items and the two groups differ in their underlying ability distributions on the secondary dimensions.

Item Bias and DIF

Shealy and Stout (1993) have postulated that DIF is a direct violation of the unidimensionality assumption in IRT (meaning the cause for DIF is multidimensionality). Since commonly used unidimensional IRT models do not account for multiple dimensions, DIF results in a direct violation of the unidimensionality assumption when a unidimensional model is used for estimation.

Secondary dimensions that are being measured by a test can be either intentional or unintentional. According to Shealy and Stout's (1993) depiction of DIF, DIF can result from an unintentional dimension being assessed, or a *nuisance dimension*, in addition to the ability of interest, or the *target ability* (Ackerman, 1992). The introduction of a nuisance dimension creates an unfair test, where one group of individuals has an unfair advantage compared to another group, because their knowledge on the secondary dimension is greater than the other group. Since the second dimension is not explicitly measured in unidimensional IRT models, this advantage is unknown until DIF analyses are conducted. DIF can be labeled either benign or adverse (Douglas et al.,

(1996). Adverse DIF occurs when the secondary dimension is not relevant to the target ability such that it has an adverse effect on the differences between groups. Conversely, benign DIF occurs when the secondary dimension is auxiliary to the primary target dimension generally indicating there are cultural difference between groups that explain the difference on the primary target ability.

Shealy and Stout (1993) defined test bias as "a test that is less valid for one group of examinees than for another group and hence acts unfairly in its attempt to assess the examinee differences in an intended to be measured trait" (p 159). Test bias leads to non-comparable test scores for two or more groups (Nandakumar, 1993). Item bias on the other hand is a single unfair item that results in measurement non-invariance among two or more groups. This conceptualization of item bias directly extends to the notion of test bias; however, group differences do not necessarily indicate DIF. It is possible to have group differences that are not caused by item or test bias. Ackerman (1992) explained that group differences do not indicate bias if items are measuring only the valid skills or constructs of interest. *Impact* is the term used to describe the manifestation of true ability differences between two groups of individuals (Ackerman, 1992). Impact is caused by a between-group difference in performance caused by differences on a valid skill; when performance differs between groups, bias cannot be assumed (Clauser & Mazor, 1998). For instance, on typical SAT-Mathematics tests Asian Americans generally score higher than Whites, and junior and seniors score higher than middle school students (Dorans & Holland, 1993). This indicates impact and not DIF/DBF because the differences are stable and consistent differences among different test taking groups. Another example of impact is reading test performance between those who speak a language other than

English at home and those who speak only or mostly English at home. Those who speak another language could be thought of as the focal group while the mostly English speakers would be the reference group (Kim & Jang, 2009). It might be hypothesized that those who do not speak English as a first language would not score as well because reading performance most likely shares a direct relationship with one's native language.

DIF occurs when an item on a test results in an unfair probability of success for one group of individuals over another, only after matching the groups by their target ability level (Zumbo, 1999). DIF is a necessary, but not sufficient condition for item bias, such that if DIF is not present in an item, the item cannot be biased, but if DIF is present in an item, item bias is not necessarily extant (Zumbo, 1999). If an item contains DIF, the item may simply be multidimensional; this does not necessarily indicate test bias but rather just the potential for bias (Ackerman, 1992; Shealy & Stout, 1993). Item bias only results when the secondary dimension has an influence on item parameters but is not associated to what the researcher wants to measure. Item bias then, does not occur when the secondary dimension is directly linked to what the researcher wants to measure (the primary trait). Zumbo (1999) posits that follow up tests need to be conducted to determine if bias is present in individual items (such as a content analysis or a statistical analysis) because bias is not present unless the researcher is able to determine the explanation, or reasoning, of the bias in the item. This means that item bias is seen as a special case of DIF, and that the main difference between DIF and item bias is that the occurrence of DIF has no explicit validity claims (Nandakumar, 1993).

As mentioned previously DIF can be conceptualized in IRT as different item parameters for distinct groups of individuals. DIF analyses are typically framed around

the comparison of the performance of a *focal group* to a *reference group*. The focal group is typically the group believed to be disadvantaged by the multidimensional nature of the item, and this group's performance is compared to the reference group, or the group believed to be advantaged. The focus of a DIF analysis is on the focal group because they are the group believed to be disadvantaged by unfair items on a test. As a concrete example of DIF, imagine an item that assesses reading comprehension skills in a passage involving sports, which is a stereotypical male interest. A DIF analysis might be conducted between males and females to see if this item unfairly disadvantages females, after conditioning on ability. DIF might be present in the item due to males' greater interest in the topic as a whole, or male's prior knowledge on the topic.

DIF then results in different probabilities of answering an item correctly after conditioning on ability. This can be visually represented in IRT by plotting IRFs for each group in the same plot. If the IRFs are different for two groups, DIF is present in the item. Since IRFs plot the probability of a correct response at a given ability level, DIF can be directly observed by overlaying IRFs for each group of interest.

DIF can manifest itself in one of two ways: uniform or non-uniform DIF. Uniform DIF can be conceptualized as differences in the difficulty parameter in IRT. Uniform DIF is depicted in the IRFs in Figure 2. In this example the difficulty parameter for Group A (solid line) is 0, the difficulty parameter for Group B (dotted line) is 1, and the discrimination parameter was held constant at 1 and the pseudo-guessing parameter held at 0. In this example, uniform DIF manifests itself because individuals from Group B have a lower probability of answering an item correctly at any given ability level on the continuum than individuals from Group A.

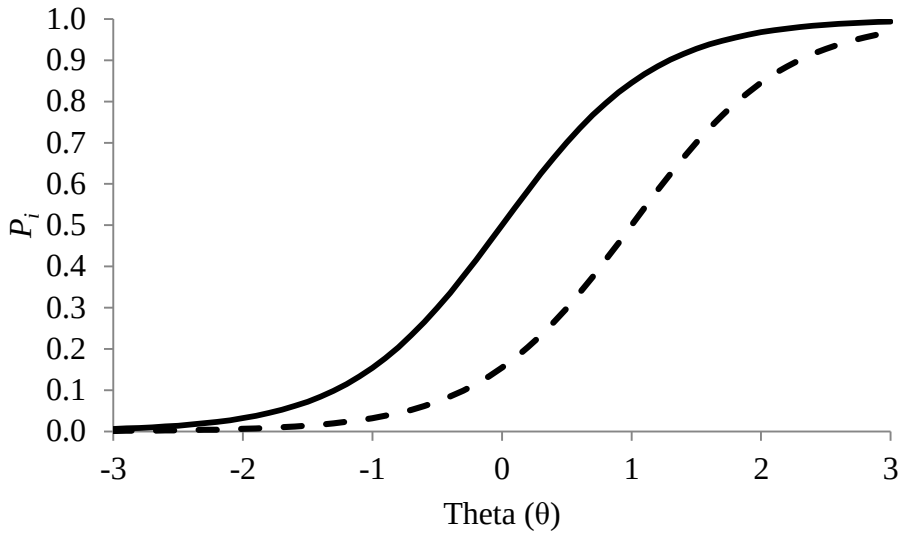

Figure 2. Uniform DIF in a single item.

DIF can also be conceptualized as a difference in the discrimination parameter, also known as *crossing DIF* (CDIF), or *non-uniform DIF*. In this case DIF can be conceptualized as differences in the discrimination parameter. Crossing DIF can also occur if there are differences between reference and focal groups on the difficulty parameter, but this is not required. Figure 3 gives a visual representation of crossing DIF (when holding the difficulty parameter at 0 and pseudo-guessing parameter at 0) where Group B has a greater probability of answering the item correctly than Group A with ability levels below 0, while Group A has a greater probability of answering the item correctly than Group B with ability levels greater than 0. Crossing DIF occurs because the item is also much more discriminating for Group A (a = 1.5) than for Group B (a = 0.75) which can also have an influence on the ability estimation process. It is also possible for crossing DIF to manifest itself in items where there are not only different discrimination parameters but also difficulty parameters for reference and focal group examinees.

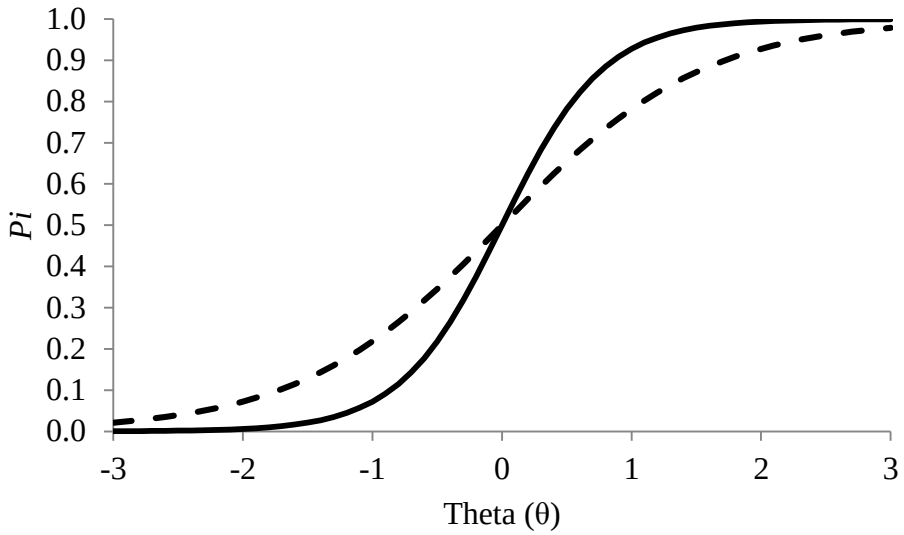

Figure 3. Non-uniform DIF in a single item.

As mentioned previously, more discriminating items are regarded as preferable items because they better differentiate between two points on the ability continuum (de Ayala, 2009). Therefore, varying discrimination parameters for two distinct groups may greatly decrease the accuracy of ability estimates for one group in relation to the other. Recalling the discussion earlier regarding item information, the addition of the discrimination parameter for the 2-PL model over the 1-PL model added an additional squared discrimination term to the item information function to account for the additional item parameter. Therefore, the resulting item information function, as previously referenced in Table 1, indicates that any item estimated to have a discrimination parameter greater than 1 will produce much more information than an item with a discrimination parameter less than 1. In general, the greater the discrimination parameter, the greater the maximum item information for estimating θ at a given point, which in turn decreases the uncertainty about a person's location along the ability continuum (de Ayala, 2009).

DBF

Amplification and cancellation.

Differential bundle functioning (DBF) can be thought of as an extension of DIF analysis. DBF is an instance of multiple DIF items which function together on the same secondary dimension conditioning on ability (Walker et al., 2012). The nature of DBF is much more complicated than DIF because of the influences individual items have on the overall effect of DBF. One such issue is that the bundling of items carries with it many unique combinations of directionality meaning that individual items in a bundle can favor the reference or focal group. Secondly, the magnitude of DIF of individual items within that bundle can vary. Thirdly, influences such as the size of the bundle and proportion of DIF items in relation to the overall measure can be influential.

Nandakumar (1993) performed a study which demonstrated the influence of amplification and cancellation. *Amplification* can be defined as occurring when a set of items in a bundle demonstrate DIF in favor of the same group in a collective manner, which amplifies the potential to detect DIF at the score level (Nandakumar, 1993). Amplification is an especially important concept in instances when assessment of single items do not yield a detectable amount of DIF. Amplification allows multiple items to act in concert to increase the ability to identify a combined DIF effect against the focal group (Douglas, Roussos, & Stout, 1996). Even if DIF is not detected for any single item, the aggregated effect of multiple items with small or moderate magnitudes of DIF can be quite large, which may suggest a large effect at the test level (Camilli & Penfield, 1997; Penfield & Algina, 2006). Amplification can be thought of as an increased sensitivity to detect a DIF effect in a group of items which all function against the same

group. To illustrate this concept consider reading comprehension passage with multiple items (for instance 4 items out of 20) involving sports topics. Even though the items measure reading comprehension, the items pertaining to sports may be easier for male test takers than female test takers because of their prior knowledge or interest. The items in the bundle require the same secondary knowledge of sports, even though the target ability being measured is reading comprehension.

Cancellation on the other hand can be thought of as an occurrence of items displaying DIF against both the focal and reference groups at *equal levels* in the same bundle, leading to the cancellation of any DIF effect at the test level (Nandakumar, 1993). It has been recommended that decisions to remove items due to DIF should not only be done at the item level but also at the test level (Nandakumar, 1993). Though test items may be found to function differentially, it is possible that those items are necessary to the test and, therefore, one may need to retain these items in the test (Zieky, 1993). By investigating a bundle effect, items can be added or removed in order to counterbalance the effect of the secondary dimension through cancellation. Nandakumar (1993) recommends that items should also be eliminated or added when investigating the effect of DIF at the test level. Even though cancellation eliminates the test level effect of DIF, the belief that cancellation does not yield biased estimates of ability has not yet been thoroughly tested. Even if cancellation can eliminate the ability to detect DIF statistically, it does not indicate what the practical significance is of the bundle on ability estimation.

Bundle formation.

The process of bundling items suspected of DBF is important because if DIF/DBF influences ability estimation to a large degree then the failure to detect such DBF would result in both biased measures and biased estimates of ability. There are four common methods of identifying bundles in DBF. These four methods utilize the test specification table, expert knowledge, statistical detection, and cognitive psychology research to develop bundles.

One such confirmatory method attempts to determine the underlying dimensions which are the root cause of DBF (Gierl, Bisanz, Bisanz, & Boughton, 2001; Gierl, Tan, & Wang, 2005). This method assumes that a test is developed to measure multiple content areas and therefore requires some form of a test specification table. A test specification table may imply that a multidimensional structure is present in the data due to different content and/or cognitive dimensions of the test. This test specification table can then be used to sample items from each content area and determine if there are multiple dimensions within different content areas. Oshima, Raju, Flowers, and Slinde (1998) completed such an investigation, comparing gender, using a test specification table from a reading comprehension measure to identify cognitive dimensions in a test that might result in DBF.

A second confirmatory method, suggested by Douglas, Roussos, and Stout (1996), recommend using expert review. This method employs content experts to identify groupings of items which are believed to measure a secondary dimension. This method involves experts reading the items to determine if there are any common themes or content similarities among the items, in order to create bundles to test for DBF.

O'Neill and McPeek (1993) found that reading passages involving science tended to favor males, while reading passages involving humanities tended to favor women. Douglas et al. (1996) recommend using this technique in instances when a test is overly dominated by one dimension and secondary dimensions are not detectable through dimensionality assessment tools. They indicate that this method has no way to identify how large the DIF in a bundle will be.

A third bundle formation method, which is exploratory in nature, was also suggested by Douglas et al. (1996). Douglas et al. (1996) recommend using a combination of agglomerative hierarchical cluster analysis (HCA) and DIMTEST in order to identify dimensional similarities between items. The bundle formation part of this method requires splitting the sample to create exploratory and confirmatory halves, run the HCA, test the clusters from HCA in DIMTEST, and then examine the bundles through human judgment to make final decisions about which items should be tested. Douglas et al. (1996) demonstrated this method on National Assessment of Educational Progress (NAEP) data which resulted in a nine-item bundle cluster. Since items are required to be subjectively examined for content the study resulted in only a six-item bundle because those items were found to be proximally similar. Therefore, three of the nine items in the bundle were removed resulting in a six item bundle to be tested for DBF.

Lastly, bundles can also be formed using cognitive psychology and human development principles to identify a dimensional structure (Kim and Jang, 2009). This method bundles items by identifying the cognitive processes required to answer items. Kim and Jang (2009) reviewed theories in language teaching and learning for native and

non-native speakers. They used prior research to show that different language groups utilize different processing methods when reading.

After reviewing the four methods of creating bundles, it is evident that only with accurate and efficient bundling can one accurately and efficiently detect DBF. The best method of bundle formation may be a combination of the aforementioned techniques. By not adhering to any of the suggested DBF techniques, a practitioner may not be developing the best bundle and thus ignoring dimensionality in the test data which can lead to inaccurate ability estimates. In general, researchers attempt to achieve the greatest validity as possible in a measure so they should strive to detect amplification in a bundle so they can alter the items in the bundle, or force cancellation in a bundle which has been shown to result in no DTF. (Nandakumar, 2003). It should be the goal of practitioners to bundle items such that all items within the bundle function against the same group, or on the other end of the spectrum, to attempt to bundle items to completely cancel out the DIF effect. If practitioners are unable to fully cancel out the effect of DIF, partial cancellation can occur which is not fully understood to this point.

Ability Estimation

There are three commonly used methods to estimate examinee ability in IRT: maximum likelihood estimation (MLE), maximum a posteriori (MAP), and expected a posteriori (EAP). MLE is a frequentist approach while MAP and EAP were developed under a Bayesian framework (Rupp, 2003). MLE estimates can be derived when the item parameters are assumed to be known. As described in Kim and Nicewander (1993) estimation using MLE makes use of the following likelihood function in order to estimate ability:

$$L(u|\theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j}, \quad (7)$$

where u is the vector of observed item responses, n is the number of items, P is the probability of response $u_j = 1$ if correct and 0 if incorrect, and $P_j = P_j(\theta)$ and $Q_j = 1\text{-}P_j$. To solve for θ, the log likelihood function can be used:

$$\ln\ (L)' = \sum \frac{P_j'(u_j - P_j)}{P_j Q_j}, \quad (8)$$

where P_j' and $(L)'$ are the partial derivatives of P_j and (L) conditional on θ. To solve for θ Equation 8 is set equal to 0 and an iterative method can be used to obtain an estimate which meets a chosen criterion. One disadvantage of using MLE in practice is that MLE is unable to produce an estimate when a respondent gets either all items correct or all items incorrect (de Ayala, 2009).

On the other hand, the Bayesian estimators EAP and MAP are generally considered better options than MLE because they can both be used to obtain trait ability estimates for all response patterns. The EAP estimation technique, originally introduced by Bock and Aitkin (1981), makes use of a prior distribution to estimate ability. Again, to estimate ability using EAP item parameters need be known. If the prior distribution of ability is represented by $g(\theta)$ then the EAP estimator for θ is the mean of the posterior distribution as follows:

$$EAP(\Theta) = \frac{\int \theta L(u|\theta) g(\theta) \partial\theta}{\int L(u|\theta) g(\theta) \partial\theta}. \quad (9)$$

where $L(u|\theta)$ is the likelihood conditional on θ from Equation 7.

Lastly, the MAP method, proposed by Samejima (1969) uses the same posterior distribution as EAP; however, instead of using the mean of the posterior distribution to obtain ability estimates, the mode of the posterior distribution is obtained by setting the

derivative of the posterior distribution to 0 and using an iterative method such as the Newton-Raphson method to find the maximum.

Due to its discrete estimation points, as opposed to being estimated on a continuum, EAP is the default setting in the most commonly used scaling programs Bilog-MG (Zimowski, Muraki, Mislevy & Bock, 2003) and MULTILOG (Thissen, Chen, & Bock, 2003). EAP has been shown to produce generally accurate ability estimates (Bock and Mislevy; 1982) and is computationally faster than MAP given that EAP utilizes the mean while MAP finds the mode requiring a more computationally intensive iterative process (de Ayala, 2009). Prior research has shown that EAP has resulted in estimates with lower standard errors (Kim & Nicewander, 1993; Wang & Wang, 2001) than MAP as well as smaller bias (Wang and Vispoel, 1998). Since EAP estimates are the most commonly used estimates of ability and prior studies have demonstrated EAP results in lower standard errors and smaller bias, EAP will be the estimation procedure used in the current study.

DIF's Influence on Ability Estimation

How DIF influences ability estimation is largely an unexplored topic. Researchers have yet to come to consensus as to what degree, if any, items displaying DIF actually lead to test bias or ability estimation bias. To this point, it is not fully understood what magnitude of DIF is required to influence resulting person estimation. Since DIF is assumed to lead to lower validity, items that are found to function differentially are typically re-written or removed from a test; however, knowledge of DIF's influence on ability estimation can help practitioners better understand how and when to treat DIF items. Overall, research investigating the effects of DIF on test bias

are largely mixed. Moreover, initial research investigating DBF's effect on ability estimation has found that ability estimation depends on the total amount of DIF as well as the proportion of items in a bundle to the test length (Walker et al. 2012). Firstly though, several studies will be described which investigated the influence of single item DIF on test bias.

One such investigation by Drasgow (1987) compared groups of individuals that differed based on both sex and race against white males on English and Mathematics tests. Individual items were found to display measurement non-invariance; however, overall results indicated there was little evidence that individual items influenced total test scores. This is because while some items in the English and Mathematics tests were found to favor White males, other items were found to favor the other groups, indicating cancellation was taking place at the test level.

In another study, Roznowski (1987) investigated the relationship between gender and subtest composite scores, which favored either males or females over a variety of topics. Roznowski found that correlations between general intelligence and the composites were consistent regardless of which group the composite favored. These results were used as an indication that items displaying group differences do not necessarily result in poor measures.

As a follow up, Roznowski and Reith (1999) found that items displaying DIF did not have a large influence at the overall test score level. They determined this by creating composite scores after flagging items for DIF using the Maentel-Haenszel method. Their composites indicated which group was favored and where there was no bias, both bias, focal bias, and reference bias. They investigated the correlations between these scores to

show that there was a high relationship between biased and unbiased composite score distributions. The correlations were high, indicating that the order of the test scores was approximately the same regardless of which composite was used. Of importance was the finding that the correlation decreased as the bias in the composite increased, indicating there was an influence in the scores to at least some degree (Roznowski and Reith, 1999). Lastly, Roznowski and Reith (1999) used t-tests to investigate differences between the slopes of regression equations for different composites predicting test scores and found that regression coefficients were different for focal and reference composites only, not the balanced composites.

Similarly, Zumbo (2003) found that item-level DIF did not show test-level invariance using a CFA framework by simulating DIF of different magnitudes in 1 to 16 items of 38 total items. Zumbo (2003) found that regardless of the magnitude of DIF, or the percentage of items containing DIF, the scale scores were measurement invariant; however, it was argued that even though the tests may be measuring the same thing a systematic bias may still interfere with the validity, and therefore, the actual interpretation of the results.

Conversely, Pae and Park (2006) investigated the influence of DIF on DTF using CFA by creating composites of items (using the Maentel Haenszel procedure) of balanced DIF, consisting of equal number of items with DIF favoring each gender group. Pae and Park (2006) found that item level DIF may influence test level performance, when unidirectional DIF is present in a balanced manner, with 5 items functioning against the focal and 5 items functioning against the reference groups. Surprisingly, Pae and Park (2006) found no cancellation at the test level, even when the magnitude of DIF

in their balanced items was equal. Because of these findings, they suggested that the relationship between DIF and DTF is much more complex than they had originally anticipated. They suggested that DIF items be removed from high-stakes test because DIF may influence overall test level performance.

These findings were consistent with prior research by Takala and Kaftandjieva (2000) which developed 4 composites of DIF using the separate calibration t-test method based on the 1PL model in order to determine if they led to gender test bias; the composites were: the whole test (40 items), a test with all items easier for females (18 items), a test with all items easier for males (22 items), and a subtest consisting of the items with no significant difference in ability (29 items). Results indicated when a test favors a group (male or female) their respective mean test scores were higher. These findings were not surprising considering that all items on those subtests were favoring the same group. Most importantly Takala and Kaftandjieva (2000) suggest that DIF items can lead to DTF but did not consider the magnitude of the difference between the male and female groups. They only considered the number of items favoring each group, similar to Pae and Park (2006).

DBF's Influence on Ability Estimation

The influence of DBF on ability estimation has only recently been studied. A simulation study by Walker et al. (2012) manipulated the number of items containing DIF in a bundle (1, 3 or 5), the test length (20 or 40 items), and the magnitude of uniform DIF against the focal group in each bundle (0.0 to 3.0 in increments of 0.2), while maintaining equal ability distributions for the reference and focal groups. Since ability was simulated to be equal, and the only difference between reference and focal groups was the difficulty

parameter of each item in the bundle, Walker et al. (2012) used t-tests between the ability estimates obtained from Bilog to determine if ability estimation bias existed. Results indicated that as the sum of DIF/DBF increased, the t-test rejection rate increased, indicating estimates have an inverse relationship with the magnitude of DIF/DBF, but only to a problematic degree when the number of test items was low. The inverse relationship was more pronounced when the proportion of items in the bundle increased. Walker et al. (2012) also investigated ability estimation bias as a function of DIF. Consistent with the prior findings, they determined that ability estimation bias increased as the test became shorter. They also found that as the sum of DIF/DBF increased, positive ability estimation bias increased for the reference group.

Results from a study by Wells, Subkoviak, and Serlin (2002), investigating the effect of item parameter drift on ability estimates, indicated a similar finding. Item parameter drift is a difference in item parameters (difficulty and/or discrimination) over subsequent testing occasions (Goldstein, 1983). The underlying cause for the changes in parameters are different in DIF and item parameter drift; however, the result is the same in that the difficulty and/or discrimination parameters are different for two groups of individuals, which makes research among these two areas quite comparable. Findings by Wells et al. (2002) determined that ability was most influenced by an interaction among the percentage of drift and test length. These findings were consistent to Walker and colleagues (2012) findings.

Person Fit

Since decisions are often made with IRT ability estimates, a fundamental goal of any educational measure is to estimate the ability of interest as accurately as possible.

Inaccurate measurement can result in negative consequences for individual test takers. For instance, an aberrant response pattern can lead to either a student with a high probability of success in an academic program to miss an opportunity of admission into that program because they have a spuriously low test score while a spuriously high test score could result in someone with a low probability of success in an academic program to be accepted into the same program. Person fit statistics can be used to identify examinees with aberrant item response patterns which result in spuriously low or high test scores (Karabatsos, 2003). Reise (1990) defined a person fit statistic as a statistic developed to assess whether an examinee's responses aggregated across items are congruent with a specified IRT model. Person fit indices then investigate the consistency of a response pattern to the IRT model (Reise, 2000).

Meijer (1996) summarized the following 7 possible causes of misaligned person fit on dichotomously scored tests but did not suggest these are all of the possible causes of misfit: sleeping behavior, guessing, cheating, alignment errors, plodding, creativity, and deficiency of sub-abilities. Sleeping behavior is an instance where an examinee does not check answers to the easier items resulting in an incongruent proportion of incorrect easy items compared to more difficult items. Guessing behavior occurs when an examinee gets an abnormal amount of items correct of medium or high difficulty because they have guessed blindly at the answers. Cheating, or simply copying answers from a third party source, results in a high proportion of difficult and easy items correct and a lower percentage of middle difficulty items correct because most likely only the most difficult items will be copied from another. Next, according to Meijer (1996), plodding can be thought of as "a person who works very slowly and methodically and refuse to

proceed to the next item until they have done their utmost to answer the item correctly" (p 5). A person could also be extremely creative in that they believe the easy items to be too simple to be true and reinterpret the items and thus get easy items incorrect thus resulting in better fit for medium and high difficulty items. Lastly, Meijer introduced the idea of deficiency of abilities which is most similar to instances of DIF. Suppose a measure is assessing sub-abilities θ_a and θ_b. If a subset of easy items measure θ_a while a subset of hard items measure θ_b then a person who is more knowledgeable in θ_b would result in a response pattern with more items correct for θ_b.

It can be argued that DIF is also a cause for person misfit because the same items which are more difficult for focal group examinees will also be less difficult for reference group examinees resulting in an aberrant response pattern specifically for the reference group examinees. It can be argued, when the abilities of reference and focal group examinees are estimated together in a unidimensional IRT model, an aberrant response pattern will result for reference group examinees with close to average ability levels because those examinees will be likely to get a larger proportion of difficult items correct than they should in relation to medium difficulty items. Also, focal group examinees will get a disproportionate amount of easy items incorrect compared to what would be expected given their θ.

Though the person-fit mechanism of deficiency of sub-abilities described by Meijer (1996) is also due to multiple dimensions being measured with a unidimensional IRT model, deficiency of sub-abilities is inherently different from person misfit due to DIF. Put simply, neither DIF nor bundled items in DBF are assumed to be similar in difficulty. The deficiency of sub-abilities mechanism describes an instance where the

sub-ability items causing person misfit are either all easy or all difficult whereas in DIF items can fall along a continuum. There is no assumption that items in a bundle are all similar in difficultly and thus the person misfit would not be due to the same reasons.

Though there are numerous person fit measures, the l_z person-fit statistic has been long considered one of the most powerful and easily implemented person-fit statistics for the detection of nonfitting response patterns (Drasgow, Levine, and McLaughlin, 1991; Li & Olejnik, 1997). The l_z person-fit statistic has been the most widely applied, and thus widely researched, person fit statistic (Seo & Weiss, 2013). The l_z person-fit statistic is a parametric index which assesses misfit in response patterns of individuals. One of the major advantages of the l_z statistic is that it has a theoretical sampling distribution which allows for a hypothesis test to be conducted for misfit. Given that it is a parametric measure, it can easily be applied to both the 2-PL and 3-PL IRT models.

Though the l_z person-fit statistic is frequently used it has been shown to have its downfalls. Nering (1995) investigated the normality of the l_z distribution with an estimated θ and found that the distribution of the index was consistently greater than 0.0 with a SD less than 1. Moreover, it has been demonstrated that the distribution of the index was consistently negatively skewed meaning the null distribution may not always be appropriate (Nering, 1995; Reise, 1995). Furthermore, Reise (1995) demonstrated that the power to detect l_z can be affected by the method used to estimate θ. Seo and Weiss (2013) argue that the l_z index has difficult assumptions to achieve and therefore further research is needed to understand how to use this statistic in an applied fashion resulting from the finding that there was overfit of the model. Because of this, Seo and Weiss

further argue that a Monte Carlo simulation be implemented to determine an appropriate level to detect nonfitting examinees.

Even though the l_z statistic has been found to have downfalls, it is still the most commonly applied index of its type most likely due to its hypothesis test. Given its frequent use in practice and calculable hypothesis test, the l_z person-fit statistic will serve as the detection tool for person misfit for this study. Due to the recommendations by Seo and Weiss, in addition to using the standard normal distribution to classify misfit, an initial Monte Carlo simulation will be completed to determine more appropriate cutoffs given a .05 type I error rate. A more detailed description of the statistic and methods will be discussed later.

CHAPTER 3

METHODOLOGY

A simulation study was coducted to assess the relationship between the presence of DIF/DBF and ability estimation. 500 replications were conducted for each condition in the study of which there were 2000 participants for each replication in total. The simulation was conducted using SAS 9.4 (SAS Institute Inc. 2013). The simulation process first involved simulating ability distributions to be equal or unequal for reference and focal groups to simulate the absence or presence of impact. Next, parameters were generated for the items in the bundle. DIF/DBF was then simulated by either manipulating the *b* parameter, to simulate uniform DIF, or the *a* parameter, to simulate non-uniform DIF. The focal group was simulated to have a larger *b* parameter than the reference group, while the item was better able to differentiate individuals along the θ continuum for the reference group in the study (the reference group was simulated to have a greater *a* parameter value than the focal group). Item responses were then generated for all examinees based on the item parameters. Next, MULTILOG was used to estimate the item parameters and then estimate, or score, ability with the EAP method. Next, person fit estimates were calculated and saved. Lastly, SIBTEST and Crossing SIBTEST were used to estimate the DIF/DBF effect. Performing a DIF/DBF analysis allows for a thorough investigation of the relationship between ability estimation and DIF/DBF detection. Performing both SIBTEST and Crossing SIBTEST DIF/DBF detection methods will also allow for a comparison of detection rates when the incorrect procedure is used (e.g. using Crossing SIBTEST in cases of uniform DIF/DBF).

Item Generation

Item parameters were simulated first, prior to simulating any DIF. DIF data generated from a 3PL model are more generalizable to the standardized testing community, so the 3PL will be used for simulation (presented earlier in equation 2 of this paper). The difficulty parameter (*b*) was simulated from a standard normal distribution with mean of 0 and standard deviation of 1; N (0, 1). Values of *b* outside of the lower and upper limits of -3 to 3 were resampled because difficulty parameters outside of this range would rarely be used in practice. The discrimination parameter (*a*) was simulated from a log normal distribution with mean of 0 and standard deviation of 0.2; LN (0, 0.2). The resulting distribution has a slightly positive skew with a mean of 1.02 and standard deviation of 0.21. This distribution was chosen because it closely resembles the discrimination parameters test developers would encounter in practice. The pseudo-guessing parameter was simulated as the computational probability of guessing an item to be correct, or (1/k) = (1/5) = 0.2.

Design Factors

A set of design factors related to DIF/DBF will be manipulated to determine their influence on ability estimation. Based on prior research, a multitude of items working in concert is believed to be much more influential in ability estimation bias than instances of single item DIF (Pae and Park, 2006; Takala and Kaftandjieva, 2000; Walker et al., 2012). Given results from these prior studies, only a test with a large proportion of items containing DIF will be the focus instead of tests containing a single item of DIF. Specifically, the design factors of interest are the influence of uniform DIF, non-uniform

DIF, impact, balance or unbalance of sample size for the reference and focal groups, total test length, and bundle size. Details are given in totality below.

Uniform DIF (9 levels).

Uniform DIF was simulated as a sum of DIF similar to a study by Walker et. (2012) which simulated the sum of DIF in a bundle (either 1, 3 or 5 items) to range from 0.0 to 3.0 in increments of 0.2. Findings of the study indicated that ability estimation differences between reference and focal groups occurred only in situations with a high proportion of DIF items. Walker et al. found that the Beta Uni Statistic followed a nearly linear trend in single item DIF analysis with both a 20 item and 40 item test. The same study found, on average, Beta Uni Statistics were found to be approximately 0.4 when the magnitude of single item DIF was 3.0, and greater than 0.5 in 3 and 5 item bundles. The simulation also determined that a single item DIF analysis will result in sufficient power with a magnitude of DIF of 0.5.

Even though a Beta Uni Statistic of 0.4 (with DIF equal to 3.0) is very large from a practical perspective and DIF can be detected with a magnitude of DIF of only 0.5, this study will utilize a range of the total sum of DIF in a test from 0.0 to 4.0 to push the boundaries of understanding when there is a large portion of DIF spread among many items. Also, investigating a sum of DIF of 4.00 will help to determine patterns in ability estimation bias, particularly if the bias follows a linear trend.

Uniform DIF was be simulated by adjusting the *b* parameter for the item to function more difficultly for the focal group. The adjustment of the *b* parameter occurred in intervals of 0.50 from 0.00 to 4.00 resulting in a total of 9 levels of the design factor. Because this study investigated the influence of uniform DIF on ability estimation in

bundles, uniform DIF was spread equally across all items in the simulated suspect bundle resulting in an overall sum of DIF. For example, if the sum of DIF is 1.0 for a 10 item test with a 2 item bundle, each item in the bundle would result in a difference in the *b* parameter of 0.5 (0.5 + 0.5 = 1.0). For demonstration purposes an abbreviated condition list for uniform and non-uniform DIF/DBF is given for the 10 item test length with 10 percent and 20 percent bundle sizes in Table 3 of the appendix.

Non-uniform DIF (3 levels).

Non-uniform DIF, or crossing DIF, has been studied in terms of DIF detection, but not in terms of ability estimation bias. Non-uniform DIF will be simulated by adjusting the *a* parameter by increasing the discrimination parameter for the reference group, because decreasing the *a* parameter for the focal group could result in negative discrimination parameters which would customarily result in the item being eliminated from the test in practice. To study the influence non-uniform DIF has on ability estimation, non-uniform DIF will be simulated in three levels: 0.00, 0.40, and 0.80, indicating an absence of non-uniform DIF (or strictly uniform DIF), a moderate amount of non-uniform DIF, and a high amount of non-uniform DIF. These levels were previously used by Narayanan and Swaminathan (1996). Unlike the uniform DIF conditions, the magnitude of non-uniform DIF will be simulated for *each item* in the bundle on the overall test. To better explain generation of non-uniform DIF consider a 10 item test with a 2 item bundle containing non-uniform DIF simulated to be 0.40. This would result in both items' discrimination parameters adjusted upwards 0.40 for the reference group while the focal group discrimination parameters would remain constant.

This method results in these items containing more information for the reference group in relation to the focal group.

Impact (2 levels).

Two levels of impact were simulated to mimic either no impact or a true difference in the underlying ability distributions between the reference and focal groups. In the case of no-impact the ability distribution for both groups was simulated from a standard normal distribution; N (0, 1). In the case of impact, the ability distribution for the reference group was simulated from a standard normal distribution [N (0, 1)] while the ability distribution for the focal group was simulated from a normal distribution with a mean of -0.5 and a standard deviation of 1; N (-0.5, 1). The levels of impact were the same as a recent study by Finch (2012) comparing DBF detection using the MIMIC model and SIBTEST. It can be expected that a 0.5 mean difference between reference and focal group examinees will provide a realistic estimate of the influence of impact on ability estimation in practice and that it can be expected that greater impact will lead to greater ability estimation bias.

Balanced or unbalanced sample size (2 levels).

It has been well referenced that as sample size increases so does detection rate for both uniform and non-uniform DIF detection procedures (Narayanan and Swaminathan, 1996; Finch & French, 2007; Gierl, Jodoin, & Ackerman, 2000). To account for this fact, sample size will be held constant at 2000 subjects throughout the simulation. Even though larger sample sizes result in greater power, SIBTEST has been found to result in acceptable results even with sample sizes of 250 participants per group (Roussos & Stout, 1996). Since, in common practice, the sample size of the focal group is likely to be

smaller than that of the reference group, prior research has utilized sample sizes of 500 or less for the focal group (Narayanan and Swaminathan, 1996). To account for population differences often found in practice, sample sizes for the reference and focal groups were simulated to be either balanced (1000 reference, 1000 focal) or unbalanced (1500 reference, 500 focal). The unbalanced condition containing 1500 reference and 500 focal group simulated examinees was chosen because it represents the population difference between Non-Hispanic Whites and Minority groups in the United States.

Total test length (3 levels).

Three different test lengths were considered: 10, 20, or 40 items. Prior research has determined that ability estimation is biased to a greater degree when estimated from a short test with a high proportion of DIF (Walker et al., 2012). Findings by Walker et al. indicated that 40 item tests (or greater) which contain DIF/DBF are not likely to influence ability estimation, unless the proportion of DIF items in that test are extremely high. Even though most standardized achievement tests are between 35 and 80 items (Narayanan and Swaminathan, 1996), the primary purpose of the current study is to determine in which instances ability estimation is biased. Since Walker et al. did not find a considerable influence for 1, 3, or 5 item bundles in ability estimation with a 40 item test, but did find an influence of DIF/DBF on ability estimation in a 20 item test, there seems little reason to study tests longer than 40 items.

Size of bundle (3 levels).

Given that Walker et al. (2012) simulated a fixed number of items (1, 3, or 5) containing DIF and found that the number of items in relation to the test was influential, the proportion of items to the overall test length is considered instead of a fixed number

of items. The size of the bundle is an important factor in determining whether the influence of DBF on ability estimation depends on the proportion of items in a bundle or simply by the overall sum of DIF in the bundle. This study will follow the same proportion of items in a bundle as a prior study by Narayanan and Swaminathan (1996) which had proportions of DIF items of 10, 20, or 40%. These proportions represent a wide range of DIF contamination levels in a test. These levels will help to clarify research by Walker et al. in order to clarify which is more influential: the sum of DIF or the proportion of the test contaminated with DIF items.

Overall, there were 972 conditions (9 x 3 x 2 x 2 x 3 x 3) considered in the analysis. Given 500 replications for each condition a total of 972 x 500 = 486,000 data sets were generated for the simulation.

DIF Detection Software

SIBTEST.

SIBTEST, or simultaneous item bias test, a nonparametric, model based procedure, was used to estimate the total effect of DIF/DBF. SIBTEST is a procedure which detects uniform DIF or DBF (Shealy & Stout, 1993). The SIBTEST procedure looks for multidimensionality, by determining if the conditional distribution of η, the secondary dimension, at given levels of θ, the target ability, are the same for reference and focal group examinees. If the distributions are not equal then DIF/DBF is present. To phrase it another way, SIBTEST compares each items marginal response function for the reference and focal group to determine if they are equal or not; if they are not found to be equal then DIF is present in an item. SIBTEST calculates an index of DIF at each

ability level and then creates a global index by integrating over the ability levels. The resulting statistic is the Beta Uni Statistic ($\hat{\beta}$).

DBF analyses in SIBTEST investigates a subset of items by weighting all individual items (*U)* equally and summing them to come up with a total subset score h_s (U):

$$h_s(U) = \sum_{i=1}^{n} U_i \tag{10}$$

where *n* is the number of items in the subset and *i* denotes each individual item. SIBTEST matches on ability using items expected by the researcher to be DIF free. DBF is expected to be present in the data when the total score on the subtest of the reference and focal groups are not equal, after matching subsets of individuals who have the same matching subtest score:

$$E_R[h_s(U)|t] \neq E_F[h_s(U)|t] \tag{11}$$

where *E* is the expected value, *R* and *F* refer to the reference and focal group respectively, and *t* is the true score on the matching subtest. The overall test for DBF assumes that examinees from the reference and focal group with the same overall test score on the matching subset will score similarly on the suspect bundle of items. If this is not the case then DBF is present. Generally researchers look to see if DBF exists against the focal group; however, SIBTEST also allows for two tailed hypothesis testing, so all equations will be written with this consideration in mind. The test statistic to test for the null hypothesis of no DBF is *B*:

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)} \tag{12}$$

$\hat{\beta}_U$ is defined as:

$$\hat{\beta}_U = \sum_{k=0}^{n_h} p_k^2 \left(\hat{Y}_{Rk} - \hat{Y}_{Fk}\right) \tag{13}$$

where p_k^2 is the proportion of respondents who have a score *k* on the matching subtest and $\left(\hat{Y}_{Rk} - \hat{Y}_{Fk}\right)$ are the adjusted means of the reference and focal groups (at that score *k)*. The denominator, or the standard error of $\hat{\beta}_U$, is defined as:

$$\hat{\sigma}\left(\hat{\beta}_U\right) = \sum_{k=0}^{n_h} p_k^2 \left(\frac{s^2_{(Y|k,R)}}{N_{Rk}} + \frac{s^2_{(Y|k,F)}}{N_{Fk}}\right) \quad (14)$$

where $s^2_{(Y|k,R)}$ and $s^2_{(Y|k,F)}$ are the variance of the suspect subtest for reference and focal group examinees respectively, while N_{Rk} and N_{Fk} are the sample size of the reference and focal group respectively.

The resulting $\hat{\beta}_U$ can be used to estimate the amount of unidirectional DIF (Nandakumar, 1993). A $\hat{\beta}_U$ value of 0.1 would indicate that reference and focal group examinees of comparable ability would have total test scores that differ by 0.1, on average, meaning that 0.1 is the predicted difference in the probability between reference and focal group examinees of getting the item correct after being matched on ability level (Nandakumar, 1993). Refer to Shealy and Stout (1993) for a more detailed explanation of the entire SIBTEST procedure. Roussos and Stout (1996) described guidelines for interpreting the $\hat{\beta}$ statistic in single item DIF where A-type DIF, or negligible DIF, has a $\hat{\beta}$ < .059 and a rejected null hypothesis, B-type dif, or moderate dif, has a $\hat{\beta}$ <.088 and a rejected null hypothesis, and C-type DIF, or large DIF, has a $\hat{\beta}$ > .088 and a rejected null hypothesis. It should be noted that these guidelines are not applicable for DBF analysis. For DBF Walker et al. (2012) proposed calculating the proportion of DIF, or the sum of DIF (in the bundle) divided by the number of items on the test. Walker et al. recommend

that as the sum of DIF increases, having more items in a respective bundle as well as a large proportion of DBF will lead to significant differences in ability.

Crossing SIBTEST.

SIBTEST is also capable of detecting non-uniform DIF, or crossing DIF, using the program Crossing SIBTEST described by Li and Stout (1996). Li and Stout (1996) defined the statistic, B_{cro}, in which the null hypothesis of no CDIF is:

$$B_{cro} = \frac{\hat{\beta}_{cro}}{\hat{\sigma}(\hat{\beta}_{cro})} \quad (15)$$

$\hat{\beta}_{cro}$ is calculated as

$$\hat{\beta}_{cro} = \sum_{k=0}^{k_{c-1}} \hat{p}_k \, (\bar{Y}^*_{Fk} - \bar{Y}^*_{Rk}) + \sum_{k=k_c+1}^{n} \hat{p}_k \, (\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk}) \quad (16)$$

Where k_c = the point where the two groups ICCs cross, $\hat{p}_k$ = the proportion of examinees with score k, $\bar{Y}^*_{Fk}$ = the mean of the suspect item(s) for the focal group adjusted for group differences on the latent trait, and $\bar{Y}^*_{Rk}$ = the mean of the suspect item(s) for the reference group adjusted for group differences on the latent trait.

Li and Stout (1996) have posited that if crossing DIF is to occur there will be no DIF at the crossing point between the focal and reference groups so the expected value for examinees with score k_c should approximate 0 at this point as demonstrated in equation (17).

$$E[\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk} | X = k_c] = 0 \quad (17)$$

Therefore, k_c is solved for $0 = \beta_0 + \beta_1(k)$ and the closest integer is used for convenience. A positive B_{cro} indicates that low proficiency examinees in the focal group have a greater probability of a correct response compared to low proficiency examinees in the reference group while high proficiency examinees in the reference group have a

greater probability of a correct response than high proficiency examinees in the focal group, and a negative B_{cro} indicates that low proficiency examinees in the reference group have a greater probability of a correct response than low proficiency examinees in the focal group while high proficiency examinees in the focal group have a greater probability for a correct response than high proficiency examinees in the reference group (Li & Stout, 1996). Since B_{cro} is dependent on k_c, no sampling distribution can be derived, so a randomization technique must be used to determine if the test statistic is statistically significant. The process involves randomly assigning a + or – to the difference between adjusted means and then calculating a crossing point estimate, DIF index, and test statistic 1000 times to create a distribution of the test statistic. The proportion of data sets that have B_{cro} greater than or equal to the obtained test statistic determines the p-value.

The Influence of DIF/DBF on Ability Estimation

Ability estimation bias will be assessed through a number of methods. Firstly, because estimates can fall above or below 0, bias will first be calculated independently for reference and focal groups as $B(\hat{\theta}) = \frac{\sum(\hat{\theta}-\theta)}{N}$, where $\hat{\theta}$ is the estimate and θ is the true parameter and N = the number of observations in the group. Mean bias will be calculated by taking an average of the bias values over the number of replications (500). A negative value will indicate the mean of estimates is below the true value while a positive value indicates the mean of estimates is greater than the true value. Secondly, RMSE will be calculated for both reference and focal groups to gain a better understanding of the accuracy and relative magnitude of the residuals. RMSE will be calculated as: $RMSE(\hat{\theta}) = \sqrt{\frac{\sum(\hat{\theta}-\theta)^2}{N}}$ where N = the sample size of the group. Mean RMSE will be

calculated by averaging RMSE by the number of replications in the study. Greater values for RMSE indicate greater variability in the estimate of ability. Thirdly, since the ability estimates θ are known, Pearson correlation coefficients (r) will be computed between the true and estimated θ for both reference and focal groups. Lastly, for each replication an independent samples T-tests will be conducted with reference and focal group examinees as the between subjects grouping factors. A t-test rejection rate will be calculated across replications to determine the rate at which significant differences are found among reference and focal group examines. For conditions with no impact the expected difference will be 0 between groups for the t-test while the expected population difference will be equivalent to the level of impact, or a difference of 0.5 when impact present.

Given that researchers are usually unable to predict the influence of impact a follow up simulation was considered which adjusted the null hypothesis for impact conditions to be 0. Results from this analysis will help to identify the probability of being unable to identify true mean differences when they are present with impact = 0.50.

Analysis of variance (ANOVA) was used to partition the total variance in the dependent variables given the design factors. Since the purpose of the current investigation is descriptive instead of inferential only the η^2 effect size (computed as $\eta^2 = SS_{effect}/SS_{total}$) was considered and not the significance test (Cohen, 1973). Given that this investigation was descriptive and exploratory, a cutoff value of 5 percent explained variance was used to determine which design factors to report.

Assessing Person Fit Using the l_z Statistic

Person fit will be assessed using the *lz statistic.* The *lz statistic*, a parametric statistic introduced by Drasgow, Levine, & Williams (1985) was an extension to the l_o likelihood function expressed by Levine and Rubin (1979). The l_z statistic was developed because l_o was not a standardized measure, meaning that the classification of fit or misfit of an item-score pattern depended on the location of one's θ (Meijer & Sijtsma, 2001). A second issue with the l_o statistic was that it lacked the required distribution of l_o under the null hypothesis in order to classify a given person-score pattern as misfitting. Given that the null is not known for l_o, the resulting classification of fit or misfit is very difficult. Because of these issues, l_z was created. Drasgow et al. (1985) developed the l_z asymptotically standard normal distributed person-fit statistic which can be expressed as:

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}} \tag{24}$$

where $E(l_o)$ is the expectation of l_o, and $Var(l_o)$ is the variance of l_o. Elements l_o, $E(l_o)$, and $Var(l_o)$ can be computed as:

$$l_0 = \sum_{i=1}^{n} \{u_i \ln P_i(\hat{\theta}) + (1 - u_i) \ln[1 - P_i(\hat{\theta})]\} \tag{25}$$

$$E(l_0) = \sum_{i=1}^{n} \{[P_i(\hat{\theta}) \ln P_i(\hat{\theta})] + [1 - P_i(\hat{\theta})] \ln[1 - P_i(\hat{\theta})]\} \tag{26}$$

$$Var(l_0) = \left\{\sum_{i=1}^{n} P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\{\ln P_i(\hat{\theta})/[1 - P_i(\hat{\theta})]\}\right\}^2 \tag{27}$$

where n is the number of items on the measure, u_i is the dichotomous response (0 or 1) of an individual to item *i*, and $P_i(\theta)$ is the probability of a correct response to item *i* given an individual's θ estimate.

The l_z person fit statistic is an asymptotically standard normal distributed person-fit statistic; therefore, the standard normal distribution can be used to detect item-score misfit; however, since Seo and Weiss (2013) have argued that the l_z index has difficult to

achieve assumptions they recommend that a Monte Carlo simulation be implemented to determine an appropriate critical value. Therefore, this simulation will rely on two sets of critical values: one from the standard normal distribution with a critical value of -1.64 for a one-tailed test, and the other developed from an initial simulation to account for the specific design factors in the study. An initial simulation with 1000 replications was conducted which recovered the type I error rate ($\alpha = 0.05$) with the non DIF design factor in the study, impact (0.0/0.0, 0.0/-0.5). Table 2 included in the appendix summarizes the critical values recovered from the initial simulation which will be used in addition to the standard normal critical value.

Person fit will be compared to the magnitude of DIF/DBF in a descriptive manner. First, Pearson correlations will be computed between person-fit statistics and the observed ability estimation bias as well as the magnitude of uniform DIF/DBF and non-uniform DIF/DBF. If relationships among person-fit and the design factors are observed, a regression will be used to predict the misfit given the design factors in the study. Lastly, proportions of reference and focal group examinees displaying person-misfit will be recorded for each condition for comparison.

CHAPTER 4

RESULTS

The effect of the design factors related to DIF/DBF on ability bias was investigated through a series of ANOVAs. Ability estimation bias and RMSE will be investigated separately. Estimation bias will be investigated to determine the direction and size of the mean difference of the ability estimate from the simulated ability parameter for each the reference and focal group, while RMSE will assess a more accurate relative distance of the estimate from parameter. Since the simulation resulted in a large sample size, and thus very large power, η^2 effect sizes were calculated to determine which design factors related to DIF/DBF, or combination of design factors, were influential in explaining ability estimation bias. Given that bias was calculated individually for the reference and focal groups, a group variable was added to the ANOVA. Analyses were completed independently for each test length because results are more interpretable when the test length remains constant and only the number of items in the bundle varies.

Ability Estimation

Ability bias.

Results of the interactions among the design factors uniform DIF/DBF, non-uniform DIF/DBF, group, impact, sample size balance, and bundle size, were reported for each of the three test lengths. Results from an ANOVA indicated the most influential design factor was impact for the 10, 20, and 40 item bundles which was found to explain 60.82, 69.94, and 74.3 percent of the variance in ability estimation bias respectively. The ANOVA also found a main effect for the reference group proportion (balance of

reference/focal groups) which was found to explain 10.63, 10.85, and 10.30 percent of the variance in ability estimation bias for 10, 20, and 40 item bundles respectively. Lastly, the ANOVA revealed that the relationship between impact and ability estimation bias was dependent on the reference group proportion in the sample. An impact by reference group proportion interaction was found to explain 6.88, 7.94, and 8.32 percent of the variance in the three test lengths respectively. A visual inspection of the results paneled by test length revealed that test length has little influence on ability estimation bias so the results portrayed will be averaged across test length. As can be seen in Figure 4, averaged across group, ability estimation bias remained fairly consistent when there was no impact in the data regardless of whether the proportion of reference group examinees was 50 percent (M = 0.06, SD = 0.05) or 75 percent (M = 0.05, SD = 0.04); however, when impact was present in the data, bias was much larger with equal numbers of reference and focal group examinees (M = 0.29, SD = 0.04) and decreased when the reference group comprised 75 percent of the sample (M = 0.16, SD = 0.04).

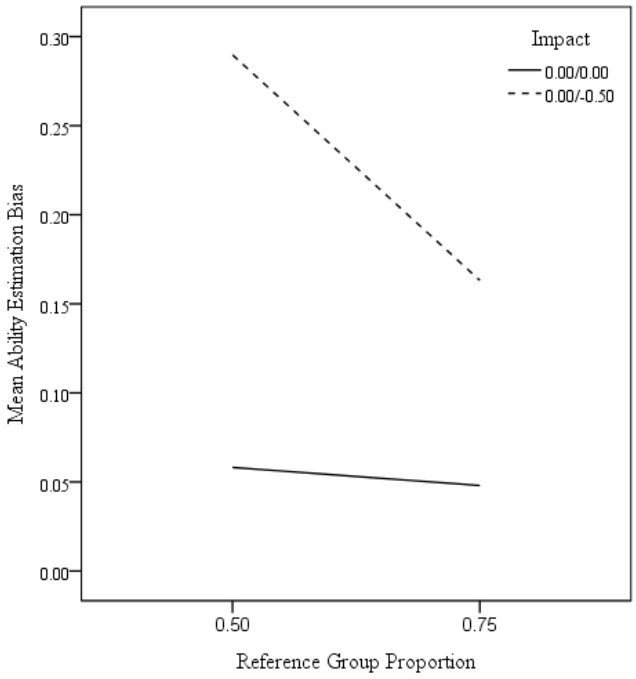

Figure 4. Ability estimation bias for impact by group proportion.

Given that impact is not necessarily related to DIF/DBF, but by definition is manipulating the underlying distribution of ability itself, it was not surprising that the presence of impact resulted in ability estimation bias; therefore, a second analysis of variance was completed to investigate the design factors for the conditions where impact was 0. Therefore, the resulting ANOVA for each test length included the design factors uniform DIF/DBF, non-uniform DIF/DBF, group, sample size balance, and bundle size.

Results of the follow up analysis, excluding impact, revealed that group (reference/focal) accounted for 15.42, 18.39, and 14.07 percent of the variance in ability estimation bias for 10, 20, and 40 item tests respectively, while the sum of DIF accounted for 6.49, 6.92, and 3.96 percent of the variance respectively. The ANOVA revealed that an interaction among group and the sum of DIF accounted for 3.96, 5.15, and 3.55 percent of the variance in ability estimation bias. Plotting this interaction exposed that the bias of the reference group and focal groups were fairly similar in conditions with a sum of DIF of 0; however, as the sum of DIF increased the mean bias of the reference group consistently increased while the mean bias for the focal group did not increase to a practical degree (Figure 5). A consistent pattern was observed among group membership and the sum of DIF across the number of items on the test; however, test length was found to be related to ability estimation bias also. Ability estimation bias was found to decrease across all other factors as test length increased.

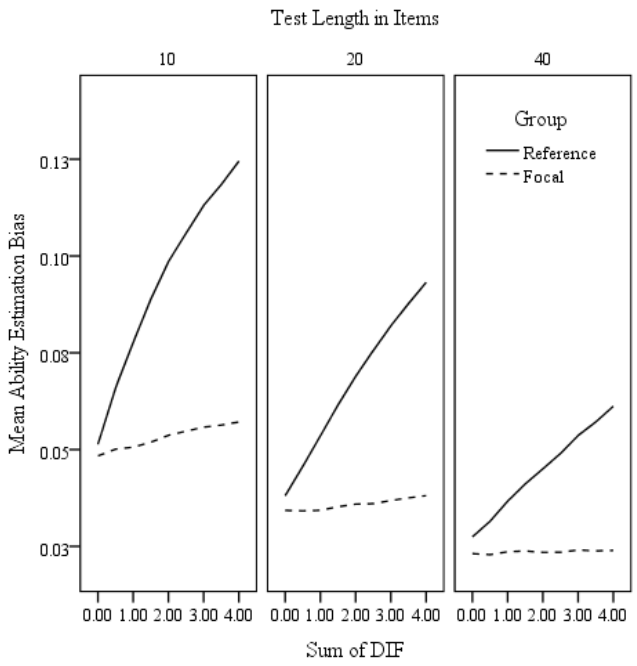

Figure 5. Ability estimation bias for the sum of DIF by group when impact was 0.

RMSE.

Again an ANOVA was used to estimate η^2 for each combination of the design factors, but this time for RMSE. The ANOVA tested the interactions for each test length among the design factors uniform DIF/DBF, non-uniform DIF/DBF, group, impact, sample size balance, and bundle size.

A factorial ANOVA revealed that impact explained the most variance in the RMSE of ability estimation by explaining 8.59, 22.90, and 44.92 percent of the variance in RMSE for the 10, 20, and 40 item tests respectively. Similar to bias, the reference group proportion accounted for a large portion of variance by explaining 2.82, 6.61, and 12.31 percent of the variance of ability estimation RMSE. The factorial ANOVA discovered that again the relationship of impact and bundle proportion with RMSE were better explained through their interaction. The relationship between impact and RMSE which depends on the reference group proportion was found to explain 1.81, 5.06, and

10.61 percent of the variance for the 10, 20, and 40 item tests respectively. The difference in η^2 across RMSEs most likely indicated that test length has an influence on the RMSE of ability estimates. Figure 6 shows that, overall, RMSE was the greatest for the 10 item test and RMSE decreased as the test length increased. This can easily be explained by better estimates of ability given more statistical information. Figure 6 also shows within test length conditions no-impact conditions resulted in fairly consistent RMSE values regardless of the proportion of reference group examinees; however, conditions with impact resulted in greater RMSE with a balanced sample and reduced RMSE the as the reference group comprised a larger portion of the sample (in this case 75 percent).

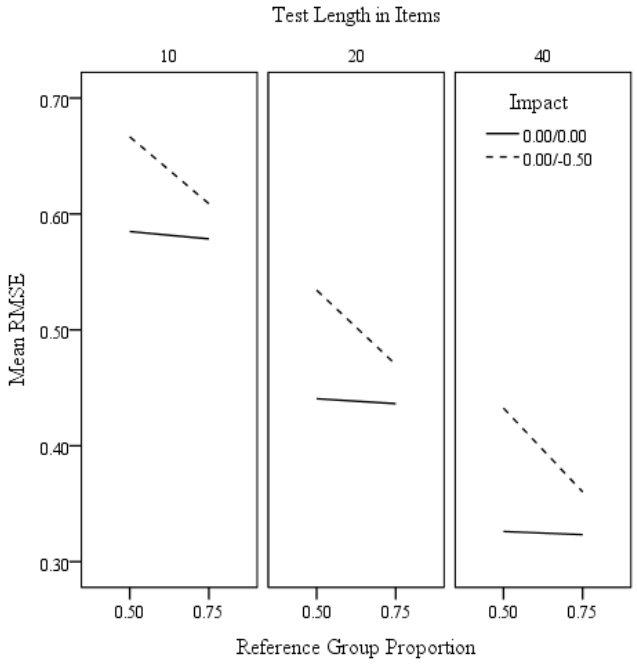

Figure 6. RMSE for the reference group proportion by impact.

Again, since impact was found to be influential in explaining ability estimation RMSE, a second set of analyses were completed to investigate the relationship among the design factors and the RMSE while only cases with impact = 0 were considered. Split by

test length the design factors uniform DIF/DBF, non-uniform DIF/DBF, group, sample size balance, and bundle size, were investigated. Split by test length, the amount of non-uniform DIF in each item was found to explain a significant amount of variance in the RMSE of ability estimation. For the 10, 20, and 40 item tests non-uniform DIF per item was found to explain 0.28, 1.38, and 4.36 percent of the variance in the RMSE of ability estimation respectively. As can be seen in Figure 7, as the magnitude of non-uniform DIF per item increased, the RMSE was found to decrease slightly. Secondly, as test length increased RMSE declined overall. These results can be explained due to the way in which non-uniform DIF was simulated. Non-uniform DIF was simulated by adding 0.00, 0.40, or 0.80 to each item for the reference group which would actually increase the amount of information in each item instead of decreasing it. Therefore, more information was present as non-uniform DIF increased, leading to lower RMSE for the simulated examinees.

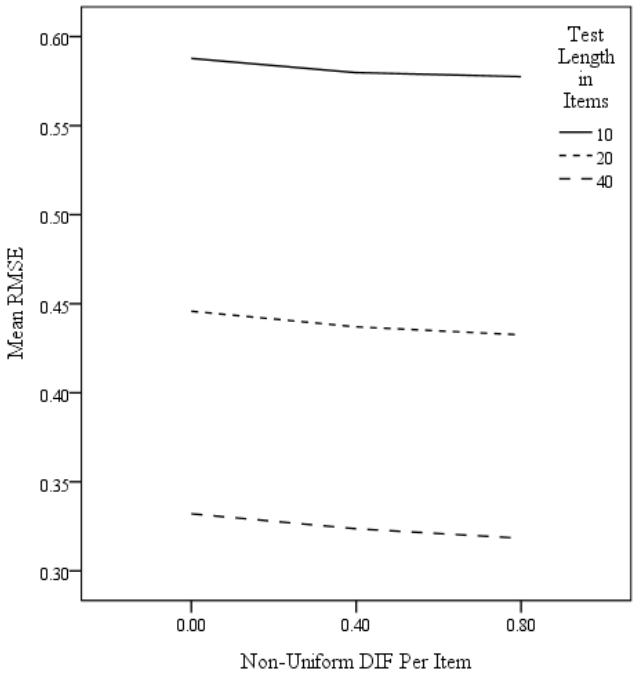

Figure 7. RMSE for non-uniform DIF per item by test length when impact was 0.

Correlation analysis.

Correlations were computed between ability parameters and ability estimates from MULTILOG. An initial investigation with a factorial ANOVA split by test length which investigated the design factors uniform DIF/DBF, non-uniform DIF/DBF, group, impact, sample size balance, and bundle size, indicated that none of the design factors resulted in explaining a practically important amount of variation in the correlations between true and estimated ability. Given that none of the factors were found to be influential, a follow up analysis was completed which also included test length. Given that ability estimates are more accurate if tests have more information, it was not surprising that test length was found to explain a large portion of the variance (η^2 = 45.76) in the correlation between estimated and true θ. Averaged over all other factors the average correlation increased to a large degree from the 10 item (M = 0.81, SD = 0.09), to the 20 item (M = 0.90, SD = 0.05), and lastly to the 40 item (M = 0.95, SD = 0.02) test length (Figure 8). The 95 percent Monte Carlo confidence intervals on the correlation between estimated and true θ were also considered for these test lengths (Figure 8). It can be seen that the confidence interval decreased as the test length increased. Again, this was not surprising given more information is obtained from a longer test. These results indicate that even when uniform and non-uniform DIF are present, as long as a measure has a sufficient number of items (in this case a 40 item test), estimates of ability remain largely in the same order.

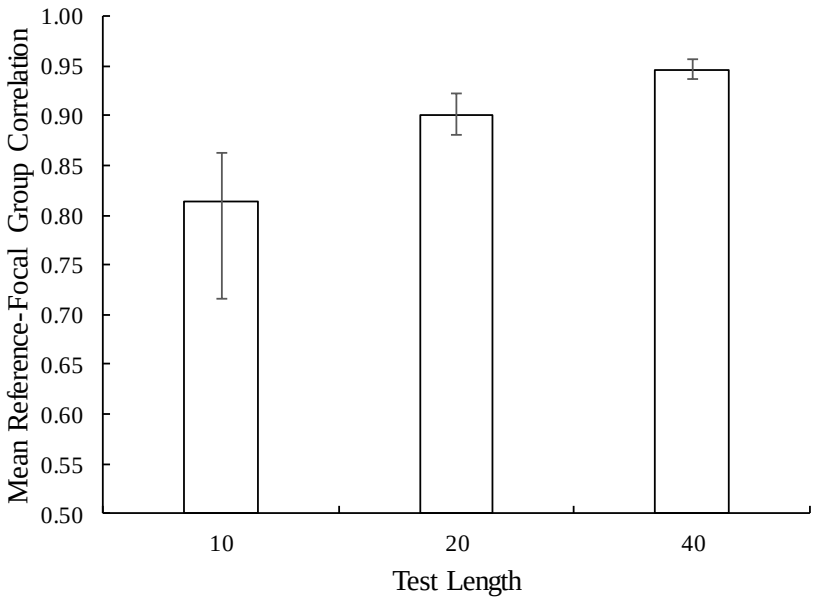

Figure 8. Reference-focal group correlation by test length. 95 percent Monte Carlo confidence intervals provided.

In order to make this more clear, Figure 9 presents the correlations between true and estimated θs across a sum of DIF of 0.0, 2.0, and 4.0 (for simplicity) with all three levels of non-uniform DIF per item. As can be seen within test length, the correlations were found to be largely consistent among the uniform and non-uniform DIF/DBF conditions.

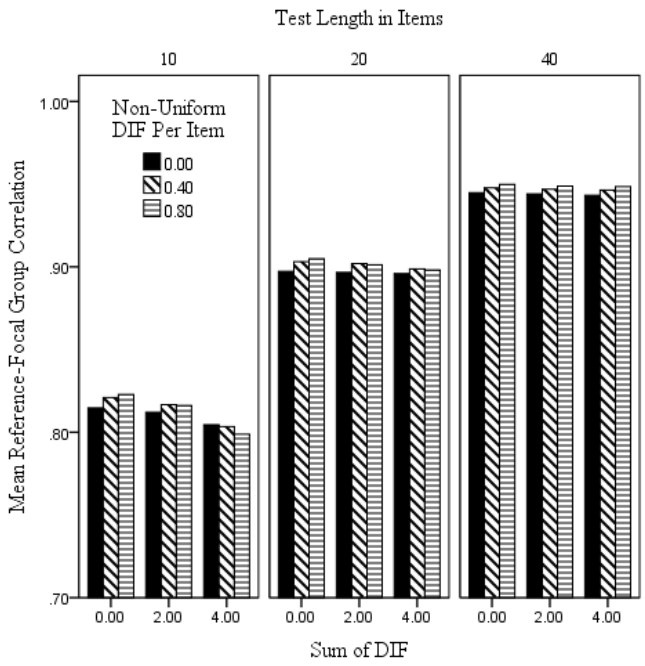

Figure 9. Rank order correlations for the sum of DIF by non-uniform DIF per item.

T-tests.

Independent samples t-tests were conducted for each replication in the study between the reference and focal group ability estimates. Given that impact is a true population difference between the reference and focal groups, the null hypothesis depended on whether impact was present or not. In no-impact conditions the null hypothesis for the t-test was that there were no differences between the population means of the reference and focal groups ($\mu_1 - \mu_2 = 0$), while conditions simulated with impact tested the null hypothesis that the population difference between reference and focal group examinees was 0.5 ($\mu_1 - \mu_2 = 0.5$), equal to the amount of impact simulated. Since the dependent variable was treated as 0 (fail to reject), or 1 (reject), a stepwise logistic regression was used to determine which of the design factors among uniform DIF/DBF, non-uniform DIF/DBF, impact, balance of the reference/focal group, and bundle size, were influential in explaining the rejection rate.

Given that Walker et al. (2012) found that uniform DIF/DBF was related to the t-test rejection rate, and prior analyses involving ability estimation bias and ability estimation RMSE were found to be related to impact, an initial visual investigation was completed investigating the influence of uniform DIF/DBF and impact on the t-test rejection rate. Results indicated that there was an interaction among the sum of DIF and impact (Figure 10). For the non-impact condition, there is a positive relationship between the rejection rates and sum of DIF. At first glance the results of the impact conditions are peculiar because the impact conditions have a high t-test rejection rate between reference and focal group examinees when the sum of DIF equaled 0, and an overall decrease in rejection rates as the sum of DIF increased (Figure 10).

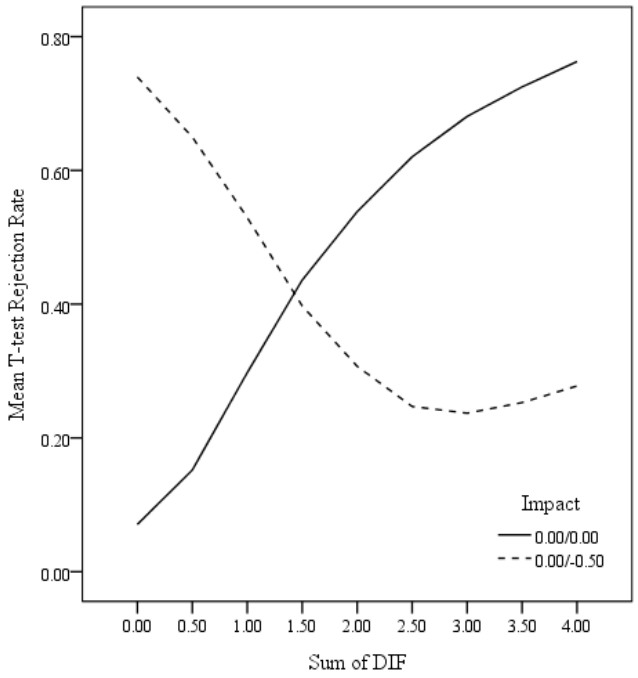

Figure 10. T-test rejection rates for the sum of DIF by impact. The null hypothesis for no-impact was ($\mu 1 - \mu 2 = 0.0$) and the null hypothesis for impact conditions was ($\mu 1 - \mu 2 = -0.5$).

In order to better understand the results from the t-test rejection rate analysis the mean difference between reference and focal groups was plotted in Figure 11. For the

non-impact conditions, across all other design factors, the mean difference between reference and focal groups approached 0 (mean difference = 0.01) when the sum of DIF equaled 0.00, and the mean difference consistently increased as the sum of DIF increased to 4.00. The impact conditions followed the same increasing pattern; however, the observed difference between reference and focal group estimates was smaller than expected. It would be expected that a difference of 0.5 would be observed when the sum of DIF equaled 0 because there were true mean differences between the groups. The observed difference when the sum of DIF equaled 0 was 0.38 and steadily increased to a mean difference 0.50 as the sum of DIF increased to 4.00. A reference line was added to Figure 11 which represents the expected mean difference between reference and focal groups adjusted for impact (0.5 was added to the non-impact conditions). As can be seen, the difference between reference and focal group ability estimates was smaller than expected in impact conditions meaning ability was estimated more similar for the reference and focal group than should be.

Overall, it appears that when no DIF was simulated in the test, the estimates between reference and focal group examinees were estimated more similarly, and as the sum of DIF increased the estimates became more spread apart. An investigation into the mean difference between reference and focal groups explained why the t-test rejection rates were not clear. Results from the impact conditions can be better understood by the definition of the null hypothesis. The null hypothesis test for the impact condition was testing a difference of 0.50 ($\mu_1 - \mu_2 = 0.5$), so it makes sense that the hypothesis test would be less frequently rejected as the mean difference between reference and focal groups increased because the difference between reference and focal groups, or the

simulated level of impact, was not observed when there was no DIF present in the test. If the mean difference between reference and focal group ability estimates in the impact condition followed the reference line in Figure 11, results for impact and non-impact conditions t-test rejections rates would have been similar. It is possible that results from the impact conditions may be simply an artifact of the level of impact simulated (0.00/-0.50). Results may have been different given a different amount of impact; however, the same trend where the mean difference between reference and focal groups is not fully observed will most likely be observed.

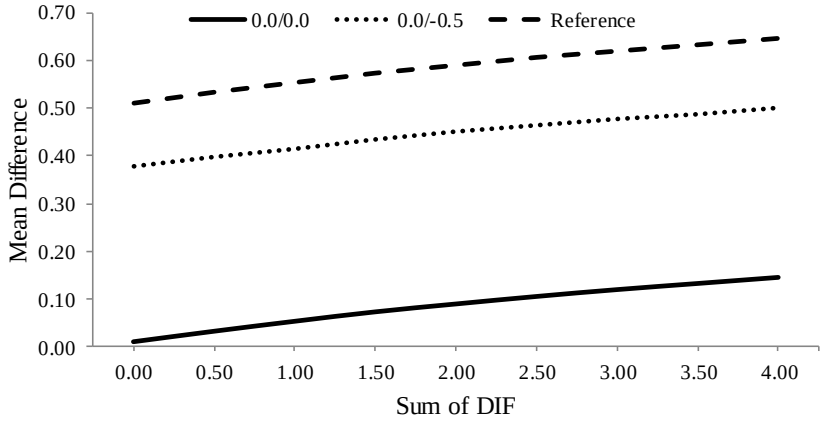

Figure 11. Mean difference between reference and focal group across the sum of DIF by impact. A reference line has been provided to compare the impact conditions mean difference with the expected difference based on the non-impact condition.

Given that the influence of impact on the t-test rejection rate seems to be simply an artifact of the level of impact selected for the study, a follow up logistic regression was completed for only the non-impact conditions. The logistic regression treated uniform DIF/DBF as a continuous variable given it had 9 levels, and non-uniform DIF/DBF, reference group balance, and bundle size as categorical variables. Results from the logistic regression for the 10, 20, and 40 item tests indicated slightly different findings. Using the Nagelkerke R Square the 10 item test found that uniform DIF/DBF

explained approximately 31.1% of the variance in the rejection rate, while bundle size accounted for an additional 9.3% of the variance, non-uniform DIF/DBF an additional 1.00%, and reference/focal group balance no additional variance (Table 4 in the appendix). The 20 item test also found that uniform DIF/DBF was the primary factor related to the t-test rejection rate (Nagelkerke R^2 = .392), while bundle size (Nagelkerke R^2 change = .031), non-uniform DIF/DBF (Nagelkerke R^2 change = .024), and bundle size (Nagelkerke R^2 change = .000) were found to have lesser influences (Table 5). The 40 item test found that uniform DIF/DBF explained 22.1% of the variance in the t-test rejection rate, while non-uniform DIF/DBF was found to explain an additional 4.2% of the variance, bundle size an additional 1.4% of the variance, and balance explaining a small amount of additional variance (0.1%; Table 6).

Given the logistic regression was conducted in an exploratory manner in order to determine which factors were related to the rejection rate, Figure 12 summarizes the relationship among the factors which were found to explain at least 4 percent of the unique variance in the t-test rejection rate given that the influence of non-uniform DIF/DBF was hypothesized to be related to the estimates of ability and it was found to explain an additional 4.2% of the variance above and beyond uniform DIF/DBF in the 40 item test. As can be seen, the t-test rejection rate increased as the sum of DIF in the bundle increased. The increase in the t-test rejection rate was found to be lower as the test length increased indicating that a longer test is less likely to show mean differences than a short one given the same magnitude of DIF/DBF overall. Non-uniform was found to relate to the t-test rejection rate in the 40 item test where there appears to be an interaction among uniform and non-uniform DIF/DBF. Specifically, as the sum of DIF

increased in the 40 item test the t-test rejection rate was found to increase; this increase was found to be greater as the amount of non-uniform DIF per item increased.

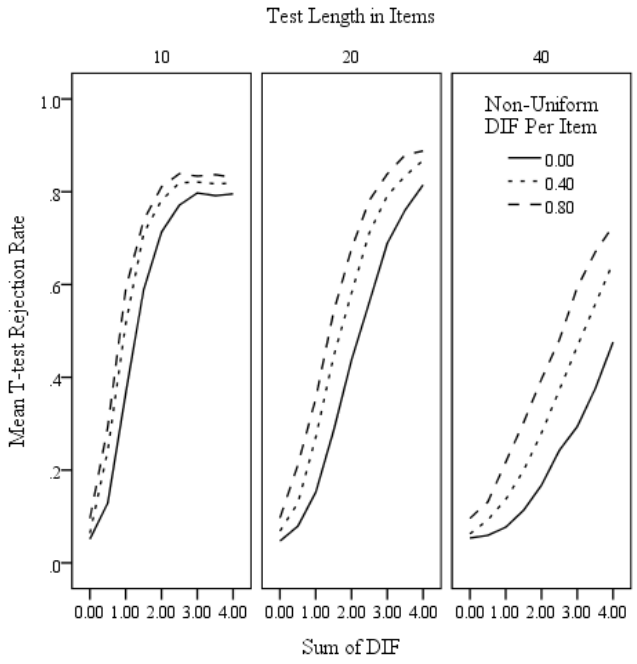

Figure 12. T-test rejection rate for the sum of uniform DIF/DBF and non-uniform DIF/DBF per item.

It could be argued that in practice it is very difficult to predict when impact will occur. When impact is present in data it would be beneficial to know the frequency of identifying impact in the resulting ability estimates through a t-test. Because there were true differences between reference and focal groups in instances of impact, and impact was simulated to a smaller degree (0.50 difference between reference and focal groups), t-tests were conducted testing the null hypothesis that the difference between groups = 0 ($\mu_1 - \mu_2 = 0$). In this test non-significant findings can be treated as error. The analyses were ran only when impact was present in the data. Design factors which were varied were the sum of DIF, non-uniform DIF per item, balance of the groups, and proportion of the test in a 40 item bundle. The follow up t-tests in the simulation unanimously

correctly rejected the null hypothesis for all 81,000 replications incorporating all combinations of design factors.

Standard error of ability estimation.

The standard error of the ability estimates were recorded in order to better understand the relationship between the design factors and confidence in ability estimates. An ANOVA tested the interactions for each test length among the design factors uniform DIF/DBF, non-uniform DIF/DBF, group, impact, sample size balance, and bundle size. An ANOVA was ran to get effect size estimates to determine how much variance in the standard error of the ability estimates the design factors explained. Results for the 10, 20, and 40 item tests indicated that the main effect of non-uniform DIF in each item explained 5.34, 11.31, and 19.24 percent of the variance in the standard error of the ability estimates respectively. The ANOVA also revealed that the bundle size explained an increasing amount of variance from the 10 (7.54%), to the 20, (9.33%), to the 40 item tests (10.29%). These main effects were found to be better explained by their interaction where the relationship between non-uniform DIF per item and the standard error of the ability estimates depended on the bundle size; for the 10, 20, and 40 item tests the percentage of explained variance in standard error was 2.16%, 3.58%, and 5.40% for the interaction of non-uniform DIF per item and bundle size.

A visual representation of the influence of the magnitude of non-uniform DIF per item and bundle size proportion can be found in Figure 13. It was found that there was a decrease in standard error as the amount of non-uniform DIF increased, or put another way as there was more information for the reference group in relation to the focal group the standard error decreased; however, this decrease was found to depend on the bundle

size where instances where the proportion of the bundle to test length increased, the a greater decrease in standard error was observed. Test length also had an influence on the standard error of the estimate. The standard error of the ability estimates was found to decrease as the test length increased; regardless of the test length the relationship among bundle size, non-uniform DIF/DBF, and ability estimate standard error was consistent.

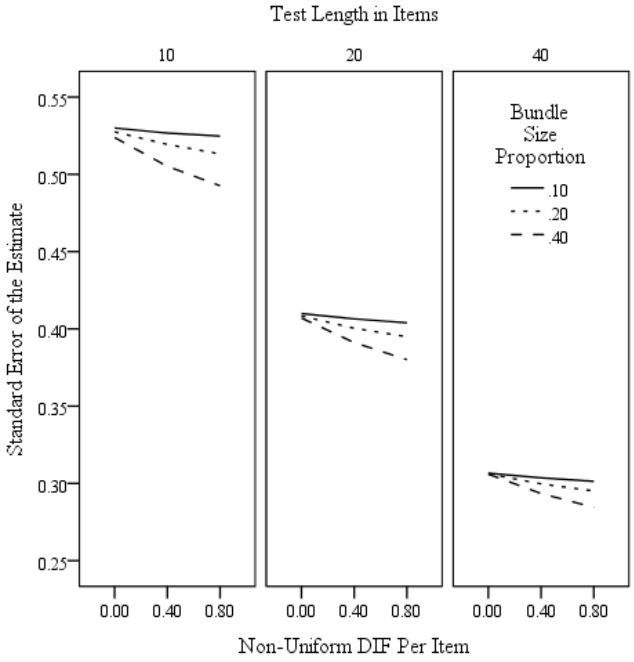

Figure 13. Standard error of ability estimation for test length by non-uniform DIF per item.

Person Fit

Differential item functioning's influence on person fit was investigated in two ways. First, person fit was assessed by investigating the proportion of misfit. Next, the influence of the DIF design factors on the l_z index was investigated. The influence of the total test length was eliminated from analysis for this specific investigation given the l_z index is known to perform less efficiently when there is a short test. Generally tests lengths of 20 to 60 items are investigated with person fit research because the chances of

detection of person misfit increase as the test length gets longer (Rupp, 2013). As the test gets longer more statistical information is gained about the expected response pattern, and thus more accurate estimation of person fit can be achieved. Therefore, only the 40 item test length conditions were considered for person fit. This resulted in the investigation of the following factors influence on person fit: group membership (reference or focal), sum of DIF, non-uniform DIF per item, bundle size, impact, and the proportion of the sample in the reference group.

Person fit was assessed using the standard normal critical value (-1.64). The total number of simulated participants who were flagged as having a misfitting response pattern were flagged independently for the reference and focal groups. Due to the fact that one of the design factors in the study was the balance of the reference to focal group where the reference group accounted for either 50 % of 75% of the total sample, calculating the mean number of cases flagged for misfit was inappropriate. In order to account for this issue a percentage of misfit was calculated from the sample size and proportion of the sample the reference group comprised where the reference group was calculated as [Misfit Percentage = Sample Size * Balance] and focal group was calculated as [Misfit Percentage = Sample Size * (1 – Balance)]. Stepwise regressions were used to determine which, or if any, of the design factors were influential in explaining misfit in the sample. Impact and reference group sample size were dummy coded where no impact and equal group sizes respectively were treated as the reference categories for these analyses. An Analysis of Variance (ANOVA) was also conducted to determine if a combination of the design factors were influential in explaining misfit. Again, given that the sample size was incredibly large for these analyses (N = 324,000),

η^2 was calculated to determine the percentage of variance accounted for instead of relying on the p-values themselves.

Using the standard normal cutoff of -1.64 to flag response pattern misfit, all predictors in the regression were found to be significant at the .01 level; however, it was found that the combination of the design factors only accounted for a total of less than one percent (R^2 = 0.0013) of the variance in the percentage misfitting response patterns with a 40 item test. A follow up factorial ANOVA was completed in order to determine if an interaction among the design factors was better able to explain the variance in misfitting person response vectors. The follow up ANOVA revealed that none of the factors, or interaction among factors, accounted for even one percent of the variance in misfit. Taking the results from the regression and ANOVA it can be concluded that the design factors have little to no practical importance in detecting misfit.

Due to the recommendations by Seo and Weiss (2013), a follow up analysis was completed with Monte Carlo critical values (Table 2). A stepwise regression was ran with the adjusted critical value results on the 40 item test. Again, results indicated that the design factors as a whole explained less than one percent (0.28%) of the variance in the percentage of misfit per group. Similar to before, a follow up ANOVA was conducted and all of the design factors, and interactions among design factors, accounted for less than 1 percent of the variance in the percentage of misfit.

Detection Using the Incorrect Model

SIBTEST.

Lastly, given the study investigated DIF in a fully crossed uniform/non-uniform manner, a descriptive investigation was completed to test the efficiency of testing

bundled non-uniform DIF in SIBTEST and the efficiency of detecting bundled uniform DIF in crossing SIBTEST. Since SIBTEST is known to more accurately detect DIF as the test length is increased (Gierl, Gotzmann, & Boughton, 2004), and given that DIF detection in SIBTEST and Crossing SIBTEST requires an adequate matching subtest, analyses were only conducted for the 40 item test conditions.

Firstly, an investigation was completed to determine the ability of SIBTEST to detect DBF in the presence of multiple items containing non-uniform DIF among the design factors. Results indicated a multitude of things. Firstly, it was found that the reference group proportion had little influence on the detection of DBF in SIBTEST across the included design factors. Given the reference group proportion was not found to be influential only the sum of DIF, non-uniform DIF per item, impact, and the size of the bundle in the 40 item test were investigated further (Figure 14). Findings indicate that the type-I error rate of SIBTEST when non-uniform DIF was 0.00 was close to a nominal level across all conditions. To clarify this, a reference line was drawn at 0.05 in Figure 15. It was also found that as the sum of uniform DIF per item increased, the ability to detect DBF was found to be greatly increased as the number of items in the bundle decreased. Therefore, with equal magnitudes of DIF, SIBTEST was found to detect DBF with higher power as the number of items in the bundle decreased. Thirdly, consistent with recent findings by Finch (2012) as impact increased from 0.00 to 0.50 between the reference and focal group, the power of SIBTEST was found to decease slightly across all other conditions.

The ability of SIBTEST to detect DBF among non-uniform items was also investigated. Results indicated that the amount of non-uniform DIF per item was found

to have an influence on the detection of DIF in SIBTEST. In the presence of only non-uniform DIF, so when the sum of uniform DIF was 0, the power to detect DBF was consistently low even when non-uniform DIF/DBF was high. When the sum of DIF was 0.00 and the magnitude of non-uniform DIF per item 0.40 the power to detect DBF ranged from 0.13 to 0.18 across all conditions; when the sum of uniform DIF was 0.00 and the magnitude of non-uniform DIF per item was 0.80 the power to detect DBF ranged from 0.25 to 0.31. These findings were expected given Crossing SIBTEST was developed for this instance. Interestingly an inverse relationship between the sum of uniform DIF and non-uniform DIF per item was observed. Specifically, as the sum of uniform DIF increased, the power of SIBTEST was lower as non-uniform DIF per item increased. This was particularly true between a sum of uniform DIF between 1.00 and 3.00 for the 8 and 16 item bundles.

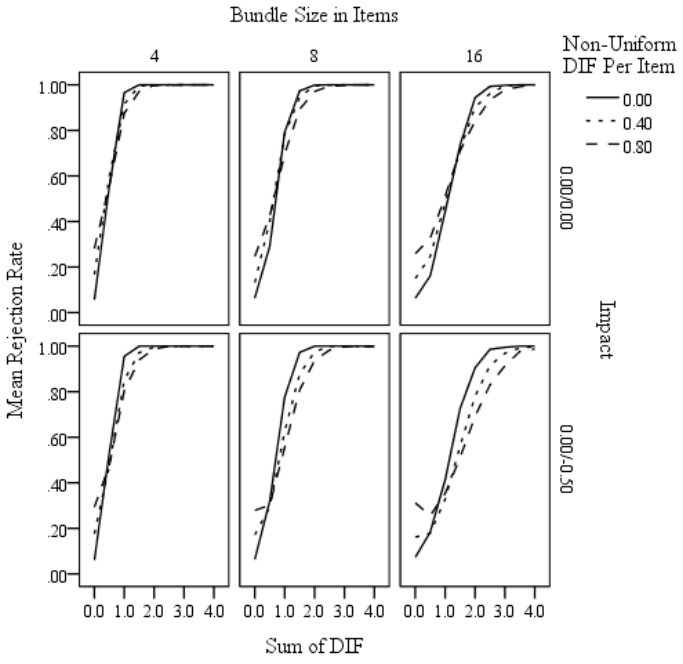

Figure 14. Type-I error and power for SIBTEST across sum of DIF, non-uniform DIF per item, and impact in a 40 item test.

Crossing SIBTEST.

Results for the detection of DBF with Crossing SIBTEST were also considered (Figure 15). Again a reference line was drawn at 0.05 to correspond to a nominal type-I error rate. Results indicated that Crossing SIBTEST maintained a nominal type-I error rate in non-impact conditions; however, 8 and 16 item impact conditions were found to result in a slightly inflated type-I error rate. Firstly, the detection of non-uniform DBF will be investigated. When controlling the sum of uniform DIF to be 0, power for Crossing SIBTEST to detect non-uniform DBF per item when non-uniform DBF was simulated to be 0.40 was 0.39, 0.46, and 0.57 for the 4, 8 and 16 item bundles respectively across impact free conditions. Conditions containing impact resulted in lower power than the same non-impact conditions with powers of 0.28, 0.36, and 0.41 for the 4, 8, and 16 item bundles respectively. Again, when non-uniform DIF per item was simulated to be large, non-impact conditions resulted in powers for the 4 (0.66), 8 (0.75), and 16 (0.84) item bundles that were larger than the same impact conditions for the 4 (0.58), 8 (0.68), and 16 (0.71) item bundles.

Results also indicated that detection of uniform DBF using Crossing SIBTEST resulted in adequate power, when controlling non-uniform DBF to be 0 (represented as the solid line in Figure 15). Uniform DBF power using Crossing SIBTEST appears to be very similar to that of SIBTEST.

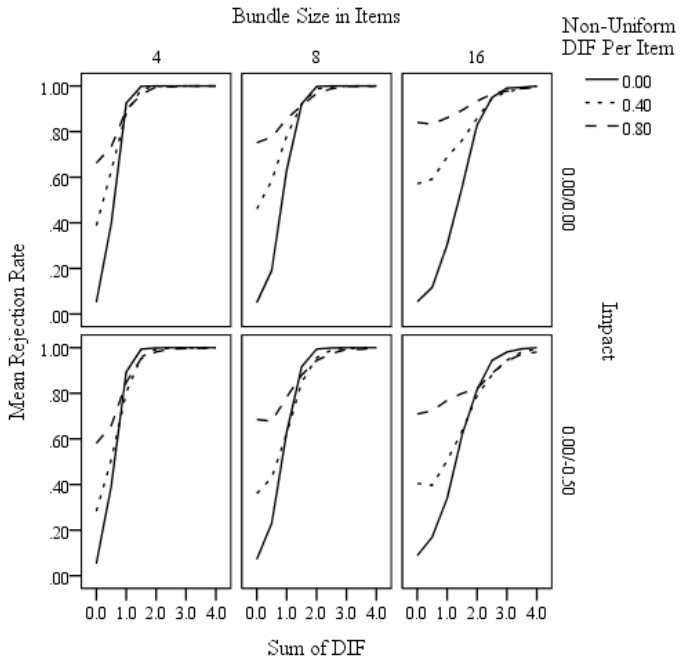

Figure 15. Type-I error and power for Crossing SIBTEST across sum of DIF, non-uniform DIF per item, and impact in a 40 item test.

In order to make a more direct comparison between the two methods, results from SIBTEST and Crossing SIBTEST were overlaid on the same figure (Figure 16). As can be seen in Figure 16, when the bundle proportion to the overall test is not large power remains high using Crossing SIBTEST to detect uniform DBF. This was determined by investigating the detection rate while non-uniform DIF in each item was 0.00. Overall, the detection of both uniform DIF/DBF as well as non-uniform DIF/DBF with Crossing SIBTEST was comparable or better than SIBTEST. Crossing SIBTEST may be a method worth using if testing for DBF regardless of the form considering there is similar power given a smaller bundle proportion, as well as the ability to test for non-uniform DBF.

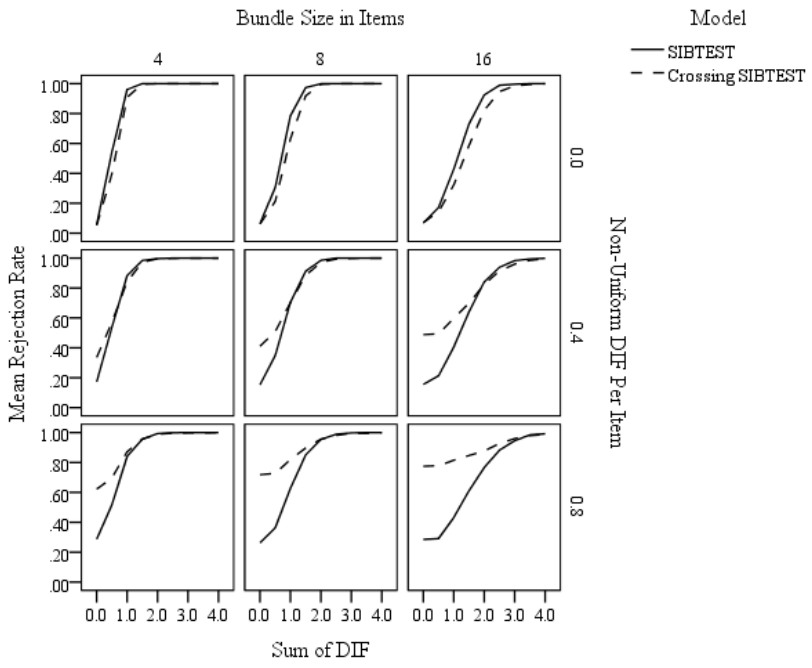

Figure 16. Type-I error and power comparisons between SIBTEST and Crossing SIBTEST across the sum of DIF, non-uniform DIF per item, and bundle size in a 40 item test.

CHAPTER 5

DISCUSSION

This study investigated the influence of uniform and non-uniform DIF/DBF on ability estimation using a unidimensional IRT model and also person fit through a simulation study. DBF was simulated in 10, 20, and 40 item tests with bundle sizes of 10, 20, and 40 percent for each of those test lengths. Uniform DIF/DBF was simulated as a sum of DIF in the difficulty parameter across all bundled items which ranged from a total sum of DIF of 0.00 to 4.00. Non-uniform DIF/DBF was simulated as a change in the discrimination parameter of each item in the bundle of 0.00, 0.40, or 0.80. In addition to investigating the direct influence of uniform and non-uniform DIF/DBF as well as test length and bundle size on ability estimation, other commonly associated design factors related to DIF/DBF were studied including impact and the balance of reference/focal group examinees. ANOVAs were calculated to estimate η^2 effect sizes in order to determine how much variance the design factors, or combination of design factors, were explaining in the bias of ability estimates, RMSE of ability estimates, standard error of the ability estimates, and person misfit. T-tests were also conducted, with particular interest in conditions with no simulated impact given no ability differences were simulated in those conditions, in order to determine how different the means were of the reference/focal group examinees as the magnitude of uniform and non-uniform DIF/DBF was varied; follow up ANOVAs were ran to determine what design factors influenced the t-test rejection results. Lastly, since both SIBTEST and Crossing SIBTEST were used to detect the simulated DBF effect, power and type-I error from SIBTEST and Crossing

SIBTEST were investigated from a descriptive manner across uniform and non-uniform DBF conditions.

Ability Estimation

Bias.

Bias of the ability estimates was calculated separately for the reference and focal groups in the study. Ability estimation bias was calculated in order to determine the direction of the bias taking place. Impact, as well as the reference/focal group balance, was found to play a role in explaining ability estimation bias. This result is not surprising given impact is defined as a difference in the mean ability distributions between the reference and focal groups. One possible reason for the influence of impact is due to the combined ability distribution when it is present. When impact was simulated the resulting distribution for the entire sample is flatter (a lower kurtosis value) with a combined mean somewhere between the mean of the reference and focal group; this mean is largely dependent on the sample size balance between the reference and focal group. It is possible that the resulting increased bias is due to a disagreement between the expected distribution in Multilog and the actual distribution in the sample given that the default setting in Multilog is to set the distribution for ability to standard normal. When impact was present ability estimation bias was greater for the balanced design (1000 reference and 1000 focal) compared to the unbalanced design (1500 reference and 500 focal). These results seem intuitive given when there was a balanced design the proportion of the sample with a lower mean θ was larger than when the sample was unbalanced. The reduction in the number of those in the focal group from balanced to

unbalanced designs resulted in a change from more heterogeneous samples among the reference and focal groups to a more homogeneous one.

Since impact is a direct influence on the distribution of θ it was not surprising this resulted in an influence on the ability estimates, so a follow up analysis was conducted that ignored impact as a factor, thus only investigating conditions where impact equaled 0. Results from this analysis found that an interaction between group, reference or focal, and the sum of uniform DIF was influential in explaining variance in the bias of the ability estimates. Similar to the Walker et al. (2012) study, the current investigation found an increase in reference group ability estimation bias as the sum of DIF increased; however, Walker et al. had previously found that as the sum of DIF increased the focal group bias decreased. The current investigation found as the sum of DIF increased there was little change in focal group ability estimation bias.

Implications of these findings can be better understood by looking at the direction of the bias. In all instances the mean ability estimation bias was positive indicating estimates for the reference group were increasingly more overestimated as the sum of DIF increased, while ability estimation bias remained relatively constant within test length for the focal group regardless of the sum of DIF. These results help to clarify a long unspoken debate in the DIF literature which has been whether DIF/DBF results in the reference group becoming more advantaged or the focal group more disadvantaged. Results from this study seem to indicate that DIF has a larger influence on reference group ability estimates than the focal group ability estimates. The reference group examinees actually appear to become more advantaged as the sum of DIF increases while the focal group examinees ability estimates remains relatively constant regardless the

sum of DIF. Put another way, results from the current investigation clarify that DIF/DBF seems to increase reference group estimates while the focal group estimates remain constant.

RMSE.

Similar to ability estimation bias, ability estimate RMSE was calculated separately for the reference and focal group examinees; however, RMSE was calculated to estimate the magnitude of the error in ability estimation in the presence of the DIF/DBF related design factors. Similar to the bias conditions an impact by reference/focal group balance interaction was observed following the same pattern as bias. The greatest RMSEs were found with the shortest test and smallest with the longest test. Once again this is easily explainable by the added information for more items in a longer test. Again, impact was found to increase RMSE values compared to the non-impact conditions. As the proportion of reference group to the focal group increased for the impact conditions (meaning a larger portion of the sample came from the same θ distribution), the RMSE decreased, while for the non-impact conditions the decrease was not practically meaningful.

Again, impact was removed because it was expected that impact have a negative influence on ability estimation. Results of a follow up ANOVA with impact removed indicated that RMSE was influenced by the test length and non-uniform DIF per item. As the magnitude of non-uniform DIF per item increased, RMSE was found to decrease. The decrease in RMSE can be explained by the simulation methodology. As the non-uniform DIF per item increased the amount of information in each item increased for the reference group, resulting in more accurate measurement for the reference group. Test

length was also considered, where again, more information from longer tests resulted in a decrease in RMSE across all other design factors.

Surprisingly, uniform DIF/DBF was found to be influential for bias while non-uniform DIF/DBF was found to be influential for RMSE. This finding indicates that both uniform and non-uniform DIF/DBF play at least some role in the individual level estimates of ability. The size of the bundle was found to have little influence on the bias and RMSE of ability estimates. The bundle size design factor simply spread out the DIF/DBF effect over multiple items. The fact that bundle size had little impact on resulting bias and RMSE was not surprising. This was consistent with the very nature of the philosophy behind amplification, which predicts that the combined influence of multiple items working together (regardless of size) can result in DTF. Therefore, the overall sum of DIF seems to be more related to the accuracy of estimates compared to the number of items in the bundle based on results from bias and RMSE.

Correlation.

Similar to Roznowski and Reith (1999), this study also considered the correlation among ability parameters and ability estimates. An investigation by Roznowski and Reith (1999) found high correlations among different composites in a test which favored both reference and focal groups. They concluded that even though the different composites may be favoring different groups, high correlations between the composites indicated that the examinees scores fell approximately in the same order. They concluded DIF may not play a large role in DTF; however, they also found that the correlation decreased as the bias in the composite increased which they took to be indicative that DIF can influence resulting scores to at least some degree.

Again an ANOVA was run to determine which design factors were influential in explaining the variance in correlations between true and estimated θ. Interestingly, only test length was found to influence the correlation between true and estimated θ. As test length increased so did the correlation between true and estimated ability. This means that uniform and non-uniform DIF was found to have little influence on the rank ordering of individuals regardless of their group membership. This finding helps researchers to understand that given a sufficiently long test, the influence of DIF/DBF is minimized if the goal of a particular test is to rank order individuals.

An increase in test length was also found to result in the least variation among correlations between true and estimated θ. Similarly, findings indicated regardless of the sum of DIF, or the amount of uniform DIF per item, the correlations were very similar within test length. Again, this helps to understand that the longer the test, the better accuracy of ability estimation, regardless of the presence of DIF. Though the results do not indicate test length negates the influence of uniform or non-uniform DIF, as can be seen in Figure 9, the influence of the interaction of both uniform and non-uniform DIF is very small and not likely to be influential to a practical degree. Results from the correlational analysis have a practical implication for test developers. Though results from the bias and RMSE investigation helped to explain the influence of uniform and non-uniform DIF/DBF on ability estimation, it seems that test length plays a very influential role in ability estimation by demonstrating that even though DIF/DBF may be present in a test, the rank order of individuals can be retained to a high degree by having a sufficiently long test. Practically this means that one way to control for DIF/DBF is to increase the test length.

T-tests.

Given that reference and focal groups were simulated from the same ability distribution, when no impact was simulated, t-tests were conducted between reference and focal group examinees. The true focus of the t-test investigation was on the non-impact conditions because the interpretation is much clearer than when there is impact. Any rejection of the null hypothesis indicated that the design factors were influencing the estimation of θ overall between groups.

Most importantly, in non-impact conditions, the t-test rejection rate was found to increase as the sum of DIF increased. In a practical sense this means that the ability estimates were found to be more and more spread out between reference and focal groups as the sum of DIF in the bundle increased. In fact, results indicated that the difference between reference and focal group means continually increased as the sum of DIF increased. In practice this means that the achievement gap would be more likely to be perpetuated as the sum of DIF increases. Perhaps more interesting however is that the observed mean difference between the means of the reference and focal groups was not found to be overly large when the distributions of reference and focal groups had the same means. Even when the sum of DIF was 4.0, the observed mean difference between reference and focal groups was less than 0.15. Given the scale is standard normal, this means that the difference between reference and focal groups is still less than 0.15 standard deviations, which would be considered a small effect size if using traditional effect size guidelines. Given the observed mean differences to be so small it is difficult to pinpoint a sum of DIF where the differences become practically meaningful. Though any differences between reference and focal group means can lead to perpetuating the

achievement gap, it seems that small amounts of DIF may not largely influence that achievement gap; however, one must remember that when the results of the measure are high stakes even a small difference between group means can be meaningful, so researchers need to be cognizant of this finding regardless of how small the influence of DIF has on the difference between group means.

These results give more insight into the findings of Walker et al. (2012). Walker et al. simulated 1, 3, or 5 item bundles and found that with the 40 item test the rejection rate was nominal regardless of the sum of DIF/DBF in the bundle over those items. The current study considered the proportion of the bundle to the overall test instead of fixed bundle sizes. Based on the results from this study, a similar trend was observed where the t-test rejection rates were decreased as the test length increased. These findings were consistent with Walker et al. (2012) which indicated that the same amount of DIF/DBF over a bundle will influence the estimates to a larger degree with a shorter test than a longer test. This highlights the importance of having enough items to gain as much information as possible to result in stable estimation for each individual.

The relationship between the sum of DIF and t-test rejection rate was also found to depend on the level of non-uniform DIF per item in the 40 item test. As the sum of DIF increased in the bundle the t-test rejection rate increased; however, for the 40 item test the increase was found to be greater as the amount of non-uniform DIF per item increased.

Combing the results from the t-test with the results from the bias investigation earlier, it is evident that the observed difference between reference and focal group examinees is due to the increase in reference group ability estimates and not a change in the focal group examinee estimates. Therefore, as the sum of DIF increases the reference

group continually becomes more and more favored (e.g. higher ability estimates), while the focal group remains unchanged (e.g. similar ability estimates).

In an applied sense researchers need to be cognizant that the combined influence of multiple items working in concert can in fact have an influence on ability estimation. Moreover, large bundles of items containing DBF can influence the difference in ability estimates between the reference and focal groups. This creates a challenging problem for researchers because items which may seem to contain small amounts of DIF on a test, and therefore individually are difficult to detect, may collectively influence the outcome of a measure. This finding is consistent with the literature on DBF detection which predicted that a combined effect of multiple items can bias results from a test (Nandakumar, 2003). Therefore, it is advised that researchers employ a combination of the DBF bundling methods mentioned earlier, particularly content evaluation, to test bundles of items for DBF.

Standard error.

The standard error of ability estimation was investigated with particular interest in the relationship between the standard error of estimation and magnitude of non-uniform DIF/DBF. Non-uniform DIF was believed to have a large influence on the standard error of ability estimation. Because non-uniform DIF/DBF influences the amount of information in each item it was believed that the more information an item has the better the estimation. Therefore, a test with multiple items containing higher discrimination parameters for the reference group would presumably result in more accurate measurement than one without differential discrimination between groups. Results indicated that this hypothesis is true. As the amount of non-uniform DIF per item

increased, thus indicating a greater amount of information for the reference group members, the standard error of the ability estimate decreased. A pattern was observed where a greater bundle size, as well as greater non-uniform DBF, and thus greater discrimination parameters for the reference group, resulted in lower standard errors of the ability estimate. Put another way, this means as the number of items containing non-uniform DIF/DBF in a bundle increase, in combination with an increase in the magnitude of non-uniform DIF/DBF per item, greater confidence can be had in the estimate itself.

Person Fit

Person fit statistics give insight into how well a person's response pattern match the model used to estimate that person. Given it is well known that person fit statistics are more accurate with longer tests only the 40 item test was considered for analyses. The l_z fit index was employed to test the influence of the design factors on person fit. Using the standard normal critical values as well as Monte Carlo critical values calculated for the specific design factors involved in this study resulted with no dominant design factors explaining the variance in misfit.

Results indicated an overall consistent positive skew for the person fit statistics consistent with previous investigations by Nering (1995) and Reise (1995). To investigate the influence of the design factors on person fit a stepwise regression was used to predict the l_z estimates. None of the design factors were found to be influential in explaining the proportion of misfit in the study. It is therefore believed that even though DIF creates differences in item difficulty and discrimination, the influence on the response pattern is not enough to induce person misfit. This is not to say DIF may never

influence person fit; however, it is possible that individual items may need large differences in these parameters, over multiple items, to influence person fit statistics.

Generally person misfit is due to a misfitting response pattern not consistent with what would be expected from the examinees, such as higher ability examinees getting multiple easy questions correct or low ability examinees getting a large amount of difficulty questions correct. DIF/DBF does not influence the response pattern in a consistent way like a typical misfitting response patterns would. The link between the items in DBF does not necessarily require the items be of similar difficulty. If the content area of the items in a bundle all measured examinee ability in a similar location in the ability continuum it is possible DIF/DBF would lead to person misfit; however, this was not the method DBF was simulated for the study. A future investigation should be conducted to investigate this form of DBF on person misfit.

The choice of person fit statistic may also have had an influence on the number of individuals with detectable misfit. A study by Karabatsos (2003) investigated 23 different person fit statistics and found their performance differed by the type of person misfit simulated. Given there are a multitude of person fit statistics, of which many of them are generalizable to the three parameter logistic model, it is possible that the person fit index chosen had an influence on the classification rate of misfit.

DIF/DBF Detection

Lastly, an exploratory investigation into the detection rates of uniform and non-uniform DIF/DBF was conducted on the programs SIBTEST and Crossing SIBTEST. Of particular interest was the detection of non-uniform DBF with traditional SIBTEST and the detection of uniform DBF with Crossing SIBTEST. In order to examine this a fully

crossed design was employed. Results indicated that SIBTEST does a fairly poor job at detecting non-uniform DBF when no uniform DIF is present in the bundle. SIBTEST was also found to result in decreased power with medium levels of uniform DIF in the study when non-uniform DIF was present in the bundle. Not surprisingly, given that SIBTEST was designed to test only for a difference in the difficult of an item, it seemed that SIBTEST performed poorly when non-uniform DIF/DBF was present in the test.

Results from Crossing SIBTEST appear much more promising. Crossing SIBTEST was able to maintain a nominal type-I error rate when there was no simulated uniform and non-uniform DIF in the bundle. The power of Crossing SIBTEST was found to be comparable to that of SIBTEST when only uniform DBF was simulated in the bundle. Comparable power was observed only when the bundle was smaller. As the bundle got larger in size a decrease in power in relation to SIBTEST was observed; however, the observed power difference was not overly large. Though the power for detection of uniform DIF/DBF was slightly lower than SIBTEST, when non-uniform DBF was simulated, Crossing SIBTEST consistently performed better than SIBTEST. Overall, it appears that Crossing SIBTEST is a better tool than SIBTEST at detecting DBF even though in some instances there appears to be a slight power decrease. Further investigations should be conducted to determine exactly when Crossing SIBTEST should and should not be used for uniform DIF/DBF detection.

Further investigation should also be given into how to interpret the Crossing SIBTEST statistic because the directionality of the statistic no longer holds if using Crossing SIBTEST to detect uniform DBF. When using SIBTEST to detect DIF/DBF the direction of the statistic indicates which group is being advantaged; however, a

Crossing SIBTEST statistic indicates which group has a greater probability of a correct response along the ability scale. This makes interpretation of the statistic difficult to interpret when uniform DIF/DBF is present. In practice, if a researcher used Crossing SIBTEST to detect uniform DIF/DBF they would only be able to know that DIF/DBF is present, but not know which kind or in which direction.

Limitations and Future Directions

The current investigation had several limitations. Most notably were the methods chosen to simulate uniform and non-uniform DIF/DBF Uniform DIF/DBF was simulated as a sum of DIF in a manner that reflected amplification. In practice pure amplification would not normally be present in a test. Rather a measure would most likely result in either full or partial cancellation where items favoring the focal group would cancel out at least some of the DIF/DBF favoring the reference group. Therefore, particular interest should be given to partial cancellation of the DBF effect because the partial cancellation of the DBF effect has not been thoroughly studied in the DIF literature. It is also possible that a measure contains full cancellation. Full cancellation may also have negative influences on ability estimation. Even though cancellation should theoretically cancel out the influence of DIF at the test level, this largely depends on the person location. Though these complex situations are difficult to study in a simulation, they need due attention.

Another limitation was the method used to simulate non-uniform DIF/DBF. From a practical sense this simulation maintained acceptable discrimination for both groups by increasing the discrimination parameter for the reference group; however, in practice this may not occur. The simulation methodology of the non-uniform DIF may not be the most appropriate. In practice it is quite possible that one group will have a lower or even

negative discrimination parameter while the other group has an acceptable discrimination parameter. This situation was not considered in the current investigation and needs further inquiry as it will most likely influence the results to a more extreme degree than what was observed in this study. It is recommended that further research be completed to investigate the method of simulating non-uniform DIF/DBF to test its influence on ability estimation.

The current investigation randomly simulated datasets for each replication; however, results may have been clearer had a fixed test been used. Given the design of the study there was no way to know how the item location influenced the ability estimate bias and RMSE. A fixed test would allow an investigation into the relationship between person and item locations on ability estimation.

Another limitation to the study was the ability to determine appropriate cutoffs for the amount of DIF which lead to practically meaningful differences between the reference and focal group. It was predicted that the study would be able to determine practically meaningful cutoffs for the amount of DIF a measure could contain before having a detrimental influence on one group over another. Given the results, this hope was far too optimistic.

Impact was found to be a very influential design factor related to the estimation of ability. Given impact was found to be so highly related to bias and RMSE, it would be highly beneficial to conduct a thorough investigation in order to better understand this relationship. Particularly, the finding that the mean difference between the reference and focal group estimates of ability was not fully observed when estimating with MULTILOG needs to be further investigated. Given that the mean difference between

reference and focal group examinees was simulated to be 0.5 when impact was simulated, and that this mean difference was not fully observed in the estimation of ability, more conditions investigating the magnitude of impact should be studied with particular attention given to instances with large amounts of impact. Given the results of the current investigation, it can be inferred that the mean difference will not be fully estimated between the reference and focal group; however, it would be helpful for practitioners to know the magnitude at which impact is influencing ability estimation.

The current investigation used the 3PL to estimate ability for the simulated examinees. Though this method results in more information by estimating a pseudo-guessing parameter, the pseudo-guessing parameter may have resulted in more statistical noise in the interpretation of the findings. The current investigation was heavily compared and contrasted against the study by Walker et al. (2012) which used the 2PL to estimate ability. Though many of the results were similar to Walker and colleague's results, it is possible that some of the differences in findings were due to the model used to estimate ability and not the design factors themselves.

Practical Implications

There are three main implications that that can be derived from this research study. First, the detection of non-uniform DIF/DBF may be more important than previously thought. Often researchers test for uniform DIF/DBF and then stop because uniform DIF/DBF is directly related to the difficulty parameter, and thus thought to have a large influence on ability estimation. This study has found that non-uniform DIF/DBF can also play a role in influencing ability estimates. Though results showed that the influence of non-uniform DIF/DBF is small in comparison to uniform DIF/DBF in

influencing the estimates of ability, non-uniform DIF/DBF was found to influence both the estimate and the standard error of the estimate. Therefore, non-uniform DIF/DBF has a unique influence on ability estimation and is directly related to the confidence of the accuracy of the estimate. Detection of non-uniform DIF/DBF should be completed in addition to uniform DIF/DBF to ensure that both the reference and focal group estimates are both as accurate as possible as well as ensure that the confidence in the estimates is consistent across groups.

Secondly impact was found to have a large influence on ability estimation, particularly when the sample size was unbalanced. Given the unanimous rejection rate across all replications from the follow up analysis investigating the t-test rejection rate when impact of 0.5 with a sample size of 2000, a t-test may be a good method to initially test for impact, particularly with the subset of items expected to be DIF free. Though impact is a difficult thing to test for it is something that those in measurement need to be cognizant of prior to administering a measure. If test developers assess impact by investigating mean differences on the matching subtest through a t-test they will better understand their sample and can understand how impact will influence their results. Given the results of this study, test developers should be cognizant that the influence of impact will be greater when the proportion of the sample in each group is more balanced as this investigation found that more balanced samples result in more ability bias.

A final recommendation involves test length. Not surprisingly, given the results from this investigation it evident that longer tests result in better estimation and a greater ability to circumvent the influence of DIF/DBF on ability estimation. Consistent with findings from Walker et al. (2012) results from the t-test analysis revealed a decrease in t-

test rejection rates for non-impact conditions as the test length increased. This means that the probability of resulting mean differences between groups due to DIF/DBF can be minimized as the test length increases, so only by writing a sufficiently long test with limited DIF will test developers be able to be confident in their person estimates.

REFERENCES

Ackerman, T. A. (1992). A dyadic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*¸431-444.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34,* 123-139.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factorial ANOVA designs. *Educational and Psychological Measurement, 33,* 107-112.

De Ayala, R.J. (2009). The Theory and Practice of Item Response Theory. NY: The Guildford Press.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35-66. Hillsdale, NJ: Erlbaum.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33,* 465-484.

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

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *Journal of Mathematical and Statistical Psychology*, *38,* 67-86.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. Applied Psychological Measurement, 15, 171-191.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Finch, W. H. (2012). The MIMIC Model as a tool for differential bundle functioning detection. *Applied Psychological Measurement, 36*, 40-59.

Finch, W. H. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement, 71*, 663-683.

Finch, W. 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*, 278-295.

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

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice 20,* 26-36.

Gierl, M. J., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17,* 241-264.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (April, 2000). Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression when the performance of DIF items is large. Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Gierl, M. J., Tan, X., & Wang, C. (2005) *Identifying content and cognitive dimensions on the SAT* (No. 2005-11). New York: College Board.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20,* 369-377.

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications*. Norwell, MA: Kiuwer Academic Publishers.

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

Karabastos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277-298.

Kim, Y., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: a multidimensionality model-based DBF/DIF approach. *Language Learning, 59,* 825-865.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58,* 587-599.

Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647-677.

Li, M. F., & Olejnik, S. (1997). The power of rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 18,*21*,* 215-231.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247-264.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluation person fit. *Applied Psychological Measurement, 25,* 107-135.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9,* 3-8.

Muthen, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10,* 133-142.

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

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

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 121-129.

O'Neill, K. A., & McPeek, W. M. (1993). *Item and test characteristics that are associated with differential item functioning*. In P. W. Holland, & H. Wainer (Eds.), Differential item functioning (pp. 255-276). Hillsdale, NJ: Erlbaum.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11,* 353-369.

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

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

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43,* 295-312.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14,* 127-137.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 15*, 217-226.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*, 543-568.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33,* 215-230.

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

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

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3,* 365-384.

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

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Reponse Theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55,* 3-38.

Russell, S. S. (2005). Estimates of Type I error and power for indices of differential bundle and test functioning. *Dissertation Abstracts International, 66*(5B), 2867. (UMI NO. 3175804).

Samejima, F. (1993). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika, 58,* 119-138.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph,* No 17.

SAS Institute Inc. (2013). *SAS® 9.4 [Computer Program]*. Cary, NC: SAS Institute Inc.

Seo, D. G., & Weiss, D. J. (2013). L_z person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement, 73, 994-1016.*

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

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

Thissen, D. J., Chen, W.H., & Bock r. D. (2003). *MULTILOG* (Version 7.0) [Computer program]. Mooresville, IN: Scientific Software.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197-219.

Walker, C. M., Zhang, B., Banks, K., & Cappaert, K. (2012). Establishing effect size guidelines for interpreting the results of differential bundle functioning analyses using SIBTEST. *Educational and Psychological Measurement, 72,* 415-434.

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

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computer adaptive testing. *Journal of Educational Measurement, 35,* 109-135.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computer adaptive testing. *Applied Psychological Measurement,* 25, 317-331.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35,* 339-361.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG* (Version 3.0) [Computer program]. Mooresville, IN: Scientific Software.

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

Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20,* 36-147.

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

APPENDIX A: TABLES

Table 2

Person Fit Critical Values

	Number of Items		
	10	20	40
Impact	Critical Value	Critical Value	Critical Value
0/0	-1.25	-1.28	-1.31
0/0.5	-1.26	-1.30	-1.34

Note. Only the person fit critical values for the 40 item test were used.

Table 3

Abbreviated Condition List for a 10 Item Test: 10 and 20 Percent Bundle Sizes

		10 Percent Bundle		20 Percent Bundle			
		Item 1		Item 1		Item 2	
Uniform DIF/DBF	Non-uniform DIF/DBF	Uniform DIF	Non-uniform DIF	Uniform DIF	Non-uniform DIF	Uniform DIF	Non-uniform DIF
0.0	0.0	0	0.0	0	0.0	0	0.0
0.0	0.4	0	0.4	0	0.4	0	0.4
0.0	0.8	0	0.8	0	0.8	0	0.8
0.5	0.0	0.5	0.0	0.25	0.0	0.25	0.0
0.5	0.4	0.5	0.4	0.25	0.4	0.25	0.4
0.5	0.8	0.5	0.8	0.25	0.8	0.25	0.8
1.0	0.0	1.0	0.0	0.5	0.0	0.5	0.0
1.0	0.4	1.0	0.4	0.5	0.4	0.5	0.4
1.0	0.8	1.0	0.8	0.5	0.8	0.5	0.8
1.5	0.0	1.5	0.0	0.75	0.0	0.75	0.0
1.5	0.4	1.5	0.4	0.75	0.4	0.75	0.4
1.5	0.8	1.5	0.8	0.75	0.8	0.75	0.8
2.0	0.0	2.0	0.0	1	0.0	1	0.0
2.0	0.4	2.0	0.4	1	0.4	1	0.4
2.0	0.8	2.0	0.8	1	0.8	1	0.8
2.5	0.0	2.5	0.0	1.25	0.0	1.25	0.0
2.5	0.4	2.5	0.4	1.25	0.4	1.25	0.4
2.5	0.8	2.5	0.8	1.25	0.8	1.25	0.8
3.0	0.0	3.0	0.0	1.5	0.0	1.5	0.0
3.0	0.4	3.0	0.4	1.5	0.4	1.5	0.4
3.0	0.8	3.0	0.8	1.5	0.8	1.5	0.8
3.5	0.0	3.5	0.0	1.75	0.0	1.75	0.0
3.5	0.4	3.5	0.4	1.75	0.4	1.75	0.4
3.5	0.8	3.5	0.8	1.75	0.8	1.75	0.8
4.0	0.0	4.0	0.0	2	0.0	2	0.0
4.0	0.4	4.0	0.4	2	0.4	2	0.4
4.0	0.8	4.0	0.8	2	0.8	2	0.8

Note: The 40 percent bundle size was not included for demonstration purposes, but was included in the study.

Table 4

Stepwise Logistic Regression Predicting T-Test Rejection Rates for the 10 Item Test

Variable	*B*	*S.E.*	*Wald*	*df*	*p*	e^{β}	Nagelkerke R^2
Step 1							.311
Intercept	0.93	.01	16019.82	1	< .001	2.54	
Uniform DIF/DBF	-1.27	.02	7174.53	1	< .001	0.28	
Step 2							.404
Intercept	-0.71	.02			< .001		
Uniform DIF/DBF	-0.40				< .001		
Bundle Size			6784.86	2	< .001		
Bundle Size (0.20)	-1.77	.02	6099.22	1	< .001	0.17	
Bundle Size (0.40)	-0.40	.02	339.80	1	< .001	0.67	
Step 3							.414
Intercept	-0.46	.02	405.86	1	< .001	0.63	
Uniform DIF/DBF	1.06	.01	16509.11	1	< .001	2.88	
Bundle Size			6839.33	2	< .001		
Bundle Size (0.20)	-1.79	.02	6149.81	1	< .001	0.17	
Bundle Size (0.40)	-0.41	.02	343.80	1	< .001	0.67	
Non-Uniform			796.85	2	< .001		
Non-Uniform (0.40)	-0.60	.02	761.99	1	< .001	0.55	
Non-Uniform (0.80)	-0.20	.02	81.93	1	< .001	0.82	
Step 4							.414
Intercept	-0.41	.02	291.04	1	< .001	0.66	
Uniform DIF/DBF	1.06	.01	16510.29	1	< .001	2.88	
Bundle Size			6840.94	2	< .001		
Bundle Size (0.20)	-1.79	.02	6151.30	1	< .001	0.17	
Bundle Size (0.40)	-0.41	.02	343.92	1	< .001	0.67	
Non-Uniform			797.12	2	< .001		
Non-Uniform (0.40)	-0.60	.02	762.24	1	< .001	0.55	
Non-Uniform (0.80)	-0.20	.02	81.95	1	< .001	0.82	
Balance (1500/500)	-0.09	.02	23.65	1	< .001	0.92	

Note: Bundle size = 0.10 was treated as the reference category, non-uniform = 0.00 was treated as the reference category, balance = 1000/1000 was treated as the reference category.

Table 5

Stepwise Logistic Regression Predicting T-Test Rejection Rates for the 20 Item Test

Variable	*B*	*S.E.*	*Wald*	*df*	*p*	e^{β}	Nagelkerke R^2
Step 1							.392
Intercept	-2.13	.02	14735.63	1	< .001	0.12	
Uniform DIF/DBF	1.10	.01	19735.16	1	< .001	3.00	
Step 2							.423
Intercept	-1.74	.02	7202.81	1	< .001	0.18	
Uniform DIF/DBF	1.15	.01	19818.72	1	< .001	3.13	
Bundle Size			2522.69	2	< .001		
Bundle Size (0.20)	-1.07	.02	2440.37	1	< .001	0.34	
Bundle Size (0.40)	-0.36	.02	293.06	1	< .001	0.70	
Step 3							.447
Intercept	-1.33	.02	3262.42	1	< .001	0.26	
Uniform DIF/DBF	1.18	.01	19867.02	1	< .001	3.25	
Bundle Size			2591.19	2	< .001		
Bundle Size (0.20)	-1.11	.02	2506.93	1	< .001	0.33	
Bundle Size (0.40)	-0.37	.02	301.98	1	< .001	0.69	
Non-Uniform			2074.88	2	< .001		
Non-Uniform (0.40)	-0.99	.02	2045.15	1	< .001	0.37	
Non-Uniform (0.80)	-0.39	.02	331.77	1	< .001	0.68	
Step 4							.447
Intercept	-1.40	.03	3136.55	1	< .001	0.25	
Uniform DIF/DBF	1.18	.01	19868.19	1	< .001	3.25	
Bundle Size			2593.23	2	< .001		
Bundle Size (0.20)	-1.11	.02	2508.91	1	< .001	0.33	
Bundle Size (0.40)	-0.38	.02	302.25	1	< .001	0.69	
Non-Uniform			2076.56	2	< .001		
Non-Uniform (0.40)	-1.00	.02	2046.80	1	< .001	0.37	
Non-Uniform (0.80)	-0.39	.02	332.07	1	< .001	0.68	
Balance (1500/500)	0.14	.02	62.56	1	< .001	1.15	

Note: Bundle size = 0.10 was treated as the reference category, non-uniform = 0.00 was treated as the reference category, balance = 1000/1000 was treated as the reference category.

Table 6

Stepwise Logistic Regression Predicting T-Test Rejection Rates for the 40 Item Test

Variable	*B*	*S.E.*	*Wald*	*df*	*p*	e^{β}	Nagelkerke R^2
Step 1							.221
Intercept	-2.53	.02	16705.27	1	< .001	0.08	
Uniform DIF/DBF	0.77	.01	11247.02	1	< .001	2.15	
Step 2							.263
Intercept	-2.10	.02	9344.19	1	< .001	0.12	
Uniform DIF/DBF	0.80	.01	11471.57	1	< .001	2.22	
Non-Uniform			2802.10	2	< .001		
Non-Uniform (0.40)	-1.14	.02	2801.99	1	< .001	0.32	
Non-Uniform (0.80)	-0.48	.02	564.77	1	< .001	0.62	
Step 3							.277
Intercept	-1.80	.02	5560.50	1	< .001	0.17	
Uniform DIF/DBF	0.81	.01	11544.56	1	< .001	2.25	
Non-Uniform			2835.37	1	< .001		
Non-Uniform (0.40)	-1.16	.02	2835.25	1	< .001	0.32	
Non-Uniform (0.80)	-0.48	.02	572.36	1	< .001	0.62	
Bundle Size			934.67	2	< .001		
Bundle Size (0.20)	-0.65	.02	932.25	1	< .001	0.52	
Bundle Size (0.40)	-0.33	.02	250.38	1	< .001	0.72	
Step 4							.278
Intercept	-1.88	.03	5329.75	1	< .001	0.15	
Uniform DIF/DBF	0.81	.01	11551.42	1	< .001	2.25	
Non-Uniform			2838.52	2	< .001		
Non-Uniform (0.40)	-1.16	.02	2838.40	1	< .001	0.31	
Non-Uniform (0.80)	-0.48	.02	73.09	1	< .001	0.62	
Bundle Size			935.81	2	< .001		
Bundle Size (0.20)	-0.65	.02	933.38	1	< .001	0.52	
Bundle Size (0.40)	-0.33	.02	250.70	1	< .001	0.72	
Balance (1500/500)	0.16	.02	89.11	1	< .001	1.18	

Note: Bundle size = 0.10 was treated as the reference category, non-uniform = 0.00 was treated as the reference category, balance = 1000/1000 was treated as the reference category.

APPENDIX B: SIMULATION SYNTAX

```
libname Outfiles 'c:\ability\simout';

/*Set Options*/
options NOXWAIT XSYNC mprint mlogic;

/*Set Where Log and Output Save*/

PROC PRINTTO LOG= "C:\Ability\SimOut\LL.txt"
                        PRINT="C:\Ability\SimOut\OUTPUT.TXT" ;

%MACRO SIMU(MC);

/*DO-LOOP FOR UNIFORM DIF/DBF (5)(5)*/;
%DO UNIFORM1=1 %TO 9;

        %IF &UNIFORM1=1 %THEN %LET UNI=0.00;
        %IF &UNIFORM1=2 %THEN %LET UNI=0.50;
        %IF &UNIFORM1=3 %THEN %LET UNI=1.00;
        %IF &UNIFORM1=4 %THEN %LET UNI=1.50;
        %IF &UNIFORM1=5 %THEN %LET UNI=2.00;
        %IF &UNIFORM1=6 %THEN %LET UNI=2.50;
        %IF &UNIFORM1=7 %THEN %LET UNI=3.00;
        %IF &UNIFORM1=8 %THEN %LET UNI=3.50;
        %IF &UNIFORM1=9 %THEN %LET UNI=4.00;

/*DO-LOOP FOR NON-UNIFORM DIF/DBF (3)*/;
%DO NONUNI=1 %TO 3;
   %IF &NONUNI=1 %THEN %LET NUNI=0.00;
   %IF &NONUNI=2 %THEN %LET NUNI=0.40;
   %IF &NONUNI=3 %THEN %LET NUNI=0.80;

/*DO-LOOP FOR IMPACT (2)*/;
%DO IMPACT1=1 %TO 2;
   %IF &IMPACT1=1 %THEN %LET IMPACT=0.00;
   %IF &IMPACT1=2 %THEN %LET IMPACT=0.50;

/*DO-LOOP FOR REFERENCE SAMPLE SIZE*/;
%DO SAMPLE=1 %TO 2;
        %IF &SAMPLE=1 %THEN %LET RGSS=0.50;
        %IF &SAMPLE=2 %THEN %LET RGSS=0.75;

/*DO-LOOP FOR TOTAL SAMPLE SIZE - FIXED*/;
%DO TSAMPLE=1 %TO 1;
   %IF &TSAMPLE=1 %THEN %LET SS=2000;
```

```
/*      DO-LOOP FOR TEST LENGTH*/;
%DO TEST1=1 %TO 3;
   %IF &TEST1=1 %THEN %LET NITEMS=10;
   %IF &TEST1=2 %THEN %LET NITEMS=20;
   %IF &TEST1=3 %THEN %LET NITEMS=40;


/*      DO-LOOP FOR BUNDLE SIZE*/;
%DO BUNDLE1=1 %TO 3;
   %IF &BUNDLE1=1 %THEN %LET BUNDLE=0.10;
   %IF &BUNDLE1=2 %THEN %LET BUNDLE=0.20;
   %IF &BUNDLE1=3 %THEN %LET BUNDLE=0.40;


/*DO-LOOP FOR replications*/;
%DO C=1 %TO &MC;

**********************************************************;

/*Additional Simulation Variables*/

%LET DPI = %sysevalf((&UNI)/(&NITEMS*&BUNDLE));
%LET SIZE1 = %sysevalf(&NITEMS+2);
%LET SIZE2 = %sysevalf(&NITEMS+1);
%LET SIZE3 = %sysevalf(&SS*&NITEMS);
%LET SIZE4 = %sysevalf(&SS*&RGSS);
%LET SIZE5 = %sysevalf(&SS*&RGSS+1);
%LET TTESTIMPACT = %sysevalf(&IMPACT*-1);

/*Generate Item Parameters*/

%include "c:\Ability\ParameterGeneration.sas";

/*Save Response Data*/

Data save2;
        set Group;
                file "c:\Ability\SIBTEST\sibFULL.txt";
                if &NITEMS=10 then put COl1 1 COL2 2 COL3 3 COL4 4 COL5 5
COL6 6 COL7 7 COL8 8 COL9 9 COL10 10 COL11 11;
                else if &NITEMS=20 then put COl1 1 COL2 2 COL3 3 COL4 4 COL5 5
COL6 6 COL7 7 COL8 8 COL9 9 COL10 10 COL11 11
                                                        COL12 12 COL13 13 COL14
14 COL15 15 COL16 16 COL17 17 COL18 18 COL19 19 COL20 20 COL21 21;
```

```
		else if &NITEMS=40 then put COl1 1 COL2 2 COL3 3 COL4 4 COL5 5
COL6 6 COL7 7 COL8 8 COL9 9 COL10 10 COL11 11
                                                        COL12 12 COL13 13 COL14
14 COL15 15 COL16 16 COL17 17 COL18 18 COL19 19 COL20 20 COL21 21 COL22
22
                                                        COL23 23 COL24 24 COL25 25
COL26 26 COL27 27 COL28 28 COL29 29 COL30 30 COL31 31 COL32 32
                                                        COL33 33 COL34 34 COL35 35
COL36 36 COL37 37 COL38 38 COL39 39 COL40 40 col41 41;
run;

Data save3;
	set Group;
		file "c:\Ability\multilog\data.dat";
		if &NITEMS=10 then put COl1 1 COL2 2 COL3 3 COL4 4 COL5 5
COL6 6 COL7 7 COL8 8 COL9 9 COL10 10;
		else if &NITEMS=20 then put COl1 1 COL2 2 COL3 3 COL4 4 COL5 5
COL6 6 COL7 7 COL8 8 COL9 9 COL10 10 COL11 11
                                                        COL12 12 COL13 13 COL14
14 COL15 15 COL16 16 COL17 17 COL18 18 COL19 19 COL20 20;
		else if &NITEMS=40 then put COl1 1 COL2 2 COL3 3 COL4 4 COL5 5
COL6 6 COL7 7 COL8 8 COL9 9 COL10 10 COL11 11
                                                        COL12 12 COL13 13 COL14
14 COL15 15 COL16 16 COL17 17 COL18 18 COL19 19 COL20 20 COL21 21 COL22
22
                                                        COL23 23 COL24 24 COL25 25
COL26 26 COL27 27 COL28 28 COL29 29 COL30 30 COL31 31 COL32 32
                                                        COL33 33 COL34 34 COL35 35
COL36 36 COL37 37 COL38 38 COL39 39 COL40 40;
run;

*Create DATA FOR HT PERSON FIT;

%if &NITEMS=10 %then %do;
	data ht_data;
		infile "c:\Ability\SIBTEST\sibFULL.txt";
		input i1 1 i2 2 i3 3 i4 4 i5 5 i6 6 i7 7 i8 8 i9 9 i10 10;
	run;
%end;

%else %if &NITEMS=20 %then %do;
	data ht_data;
		infile "c:\Ability\SIBTEST\sibFULL.txt";
		input i1 1 i2 2 i3 3 i4 4 i5 5 i6 6 i7 7 i8 8 i9 9 i10 10
			i11 11 i12 12 i13 13 i14 14 i15 15 i16 16 i17 17 i18 18 i19 19 i20
20;
```

```
	run;
%end;

%else %do;
	data ht_data;
		infile "c:\Ability\SIBTEST\sibFULL.txt";
		input i1 1 i2 2 i3 3 i4 4 i5 5 i6 6 i7 7 i8 8 i9 9 i10 10
			i11 11 i12 12 i13 13 i14 14 i15 15 i16 16 i17 17 i18 18 i19 19 i20
20
			i21 21 i22 22 i23 23 i24 24 i25 25 i26 26 i27 27 i28 28 i29 29 i30
30
			i31 31 i32 32 i33 33 i34 34 i35 35 i36 36 i37 37 i38 38 i39 39 i40
40;
	run;
%end;

data outfiles.ht_data;
	set ht_data;
run;

*Create Data Sets For Later Use;

Data ThetaTrue;
	set ThetaNew;
Run;

Data RefDifficulty;
	set B_;
run;

Data FocDifficulty;
	set B;
run;

Data RefDiscrimination;
	set A;
run;

Data FocDiscrimination;
	set A_;
run;

/*USE NEXT LINE IF FINDING MC LZ CRITICAL VALUES*/

/*%include "c:\ability\lz_True.sas";*/
```

```
/*Clear MULTILOG Files*/

x 'del  c:\ability\multilog\ML_Item_3PL.MLG
            c:\ability\multilog\ML_Item_3PL.OUT
                  c:\ability\multilog\ML_Item_3PL.PAR
                  c:\ability\multilog\ML_Item_3PL.TEN
                  c:\ability\multilog\ML_Item_Score_3PL.MLG
                  c:\ability\multilog\ML_Item_Score_3PL.OUT
                  c:\ability\multilog\ML_Item_Score_3PL.SCO
                  c:\ability\multilog\ML_Item_Score_3PL.TEN
         ';

/*MULTILOG*/

/*Copy MULTILOG Files*/

x copy "c:\ability\multilog\Infiles\&SS.\ML_item_3PL_&NITEMS..MLG"
"c:\ability\multilog\ML_Item_3PL.MLG";
run;

x copy "c:\ability\multilog\Infiles\&SS.\ML_item_score_3PL_&NITEMS..MLG"
"c:\ability\multilog\ML_Item_score_3PL.MLG";
run;

/*Run MULTILOG*/

x "c:\ability\multilog\mlg_ML_Item_3PL.bat";

/*Save Theta Estiames*/

data Estimates3PL_C;
         infile "c:\ability\multilog\ML_item_score_3PL.SCO";
         input Theta_Est Theta_SE;
run;

data Estimates3PL_C2;
         set Estimates3PL_C;
                  keep Theta_Est;
run;

data Theta3PL_C;
         set Estimates3PL_C2;
         Theta3PL_C=Theta_Est;
         Keep Theta3PL_C;
run;
```

```
/*Save A B C*/

data junk;
	infile "c:\ability\multilog\ml_item_3PL.out" truncover;
	input Item $ 2-6 A 38-43 B 46-57 C 58-64;
	num=_n_;
run;

data junk2;
	set junk;
		drop num;
run;

data junk3;
	set junk2;
		do i = 1 to 2000;
			 if Item = "ITEM" then num = _n_ + 2;
		end;
	if num = . then delete;
run;

data junk4;
	set junk3;
		if _n_ = &SIZE1 then delete;
		if _n_ = 1 then delete;
		keep num;
run;

proc sql noprint;
	CREATE TABLE ABC AS
	select *
	FROM Junk4 INNER JOIN JUNK
	on JUNK4.num = JUNK.num
	;
quit;

data ABC;
	set ABC;
		drop num item;
run;

data Discrimination3PL_C;
	set ABC;
	Discrimination3PL_C=A;
		if Discrimination3PL_C < -5 then delete;
```

```
                if Discrimination3PL_C > 25 then delete;
        keep Discrimination3PL_C;
run;

data Difficulty3PL_C;
        set ABC;
                        Difficulty3PL_C = B;
                        if Difficulty3PL_C > 25 then delete;
                        if Difficulty3PL_C < -25 then delete;
                        keep Difficulty3PL_C;
run;

Data Guessing3PL_C;
        set ABC;
                Guessing3PL_C=C;
                if Guessing3PL_C>1 then Guessing3PL_C=.;
                keep Guessing3PL_C;
run;

%include "c:\ability\lz_3PL.sas";
run;

/*Combine Data*/

data Combined_3PL;
        merge ThetaTrue Estimates3PL_C;
run;

/*Calculate Bias & Seperate Data*/

data Combined_3PL;
        set Combined_3PL;
                ThetaBias=Theta_Est-ThetaNew;
                ThetaRMSE=(Theta_Est-ThetaNew)**2;
run;

data Ref_Combined_3PL;
        set Combined_3PL;
                if _n_ <= &SS*&RGSS;
run;

proc means data=Ref_Combined_3PL noprint;
        var ThetaBias ThetaRMSE Theta_SE;
        output out=Ref_3PL_C_Out(drop=_type_ _freq_)
                mean=Ref_3PL_C_Bias Ref_3PL_C_RMSE_mean
Ref_3PL_C_Theta_SE;
```

```
run;

Data Ref_3PL_C;
	Set Ref_3PL_C_Out;
		Ref_3PL_C_RMSE=sqrt(Ref_3PL_C_RMSE_mean);
	keep
		Ref_3PL_C_Bias
		Ref_3PL_C_RMSE
		Ref_3PL_C_Theta_SE;
run;

proc corr data=Ref_Combined_3PL outp=Ref_3PL_C_Cor noprint;
	var ThetaNew Theta_Est;
run;

data Ref_3PL_C_Corr;
	set Ref_3PL_C_Cor;
		if _n_ ne 4 then delete;
		Ref_3PL_C_Corr=theta_est;
		drop _type_ _name_ theta_est thetanew;
run;

data Foc_Combined_3PL;
	set Combined_3PL;
		if _n_ > &SS*RGSS;
run;

proc means data=Foc_Combined_3PL noprint;
	var ThetaBias ThetaRMSE Theta_SE;
	output out=Foc_3PL_C_Out(drop=_type_ _freq_)
		mean=Foc_3PL_C_Bias Foc_3PL_C_RMSE_mean
Foc_3PL_C_Theta_SE;
run;

Data Foc_3PL_C;
	Set Foc_3PL_C_Out;
		Foc_3PL_C_RMSE=sqrt(Foc_3PL_C_RMSE_mean);
	keep
		Foc_3PL_C_Bias
		Foc_3PL_C_RMSE
		Foc_3PL_C_Theta_SE;
run;

proc corr data=Foc_Combined_3PL outp=Foc_3PL_C_Cor noprint;
	var ThetaNew Theta_Est;
run;
```

```
data Foc_3PL_C_Corr;
	set Foc_3PL_C_Cor;
		if _n_ ne 4 then delete;
		Foc_3PL_C_Corr=theta_est;
		drop _type_ _name_ theta_est thetanew;
run;

/*Calculate Adjusted Person Fit Statistics*/

%if &NITEMS=40 & &SS=2000 & &Impact=0.00 %then %do;

*3PL;
data LZ_3PL_Ref_Count_Adj;
	Set LZ_3PL_Ref_Count_;
		If LZ_3PL < -1.31 then counter3PL_Ref_Adj = _n_;
Run;

data LZ_3PL_Foc_Count_Adj;
	Set LZ_3PL_Foc_Count_;
		If LZ_3PL < -1.31 then counter3PL_Foc_Adj = _n_;
Run;

%end;

%else %do;

*3PL;
data LZ_3PL_Ref_Count_Adj;
	Set LZ_3PL_Ref_Count_;
		If LZ_3PL < -1.34 then counter3PL_Ref_Adj = _n_;
Run;

data LZ_3PL_Foc_Count_Adj;
	Set LZ_3PL_Foc_Count_;
		If LZ_3PL < -1.34 then counter3PL_Foc_Adj = _n_;
Run;

%end;

**********************************************************************
****;

/*Create Person Fit Output*/

Proc means data=LZ_3PL_Ref_Count_Adj noprint ;
```

```
	var counter3PL_Ref_Adj;
		output out = Ref3PL_Count_Adj;
run;

Proc means data=LZ_3PL_Foc_Count_Adj noprint ;
	var counter3PL_Foc_Adj;
		output out = Foc3PL_Count_Adj;
run;

data Ref3PL_Count_SD_Adj;
	set Ref3PL_Count_Adj;
		if _n_ ne 1 then delete;
		keep counter3PL_Ref_Adj;
run;


data Foc3PL_Count_SD_Adj;
	set Foc3PL_Count_Adj;
		if _n_ ne 1 then delete;
		keep counter3PL_Foc_Adj;
run;

/*SIBTEST*/

/*Generate Data Files for SIBTEST*/

%include "c:\Ability\sibtest.sas";
run;

/*Copy SIBTEST Input File Into Directory*/

x copy "C:\Ability\SIBTEST\SIB_&NITEMS._&BUNDLE..INP"
"C:\Ability\SIBTEST\sib.in";
run;

/*Run SIBTEST*/

x "c:\Ability\SIBTEST\sibtest.bat";
run;

/*Copy SIBTEST Output File*/

x copy "C:\Ability\SIBTEST\SIB.out" "c:\Ability\SIBTEST\uniform.out";
run;

/*Read Output From SIBTEST*/
```

```
%if %sysevalf(&NITEMS=20 and &BUNDLE=0.40) %then %do;
	%include "c:\Ability\read_SIBTEST3.sas";
	run;
%end;

%else %if %sysevalf(&NITEMS=40 and &BUNDLE=0.20) %then %do;
	%include "c:\Ability\read_SIBTEST3.sas";
	run;
%end;

%else %if %sysevalf(&NITEMS=40 and &BUNDLE=0.40) %then %do;
	%include "c:\Ability\read_SIBTEST2.sas";
	run;
%end;

%else %do;
	%include "c:\Ability\read_SIBTEST.sas";
	run;
%end;

/*Crossing SIBTEST*/

/*Copy Crossing SIBTEST Input File Into Directory*/

x copy "C:\Ability\SIBTEST\SIB_&NITEMS._&BUNDLE..INP"
"C:\Ability\SIBTEST\sib.in";
run;

/*Run Crossing SIBTEST*/

x "c:\Ability\SIBTEST\csib.bat";
run;

/*Copy Crossing SIBTEST Output File*/

x copy "C:\Ability\SIBTEST\SIB.out" "c:\Ability\SIBTEST\NONuniform.out";
run;

/*Read Output From Crossing SIBTEST*/

%include "c:\Ability\read_Csibtest.sas" ;

/*T-Test Between Reference/Focal Groups*/

data ttest3pl;
```

```
	Set Estimates3PL_C2;
		if _n_ <= &SIZE4 then Group=0;
		else Group=1;
run;

%if &Impact=0.00 %then %do;

ods _all_ close;
ods results off;

ods	output TTESTS=Ttest3pl_P;
ods output Statistics=Ttest3pl_stats;

proc ttest data=ttest3pl H0=0;
	var Theta_Est;
	class Group;
run;
quit;

ods results on;
ods _all_;

%end;

%else %do;

ods _all_ close;
ods results off;

ods	output TTESTS=Ttest3pl_P;
ods output Statistics=Ttest3pl_stats;

proc ttest data=ttest3pl H0=&IMPACT;
	var Theta_Est;
	class Group;
run;
quit;

ods results on;
ods _all_;

%end;

Data Three_PL_TTEST;
	set Ttest3pl_p;
		if _n_ ne 1 then delete;
```

```
                PvalueT_3PL = probt;
                if probt<.05 then Treject_3PL=1;
                        else Treject_3PL=0;
                keep PvalueT_3PL Treject_3PL;
Run;

%if &Impact=0.00 %then %do;

Data Effect_Size_3PL;
        set Ttest3pl_stats;
                if _n_ = 3;
                MeanDif_3PL = mean;
                Cohens_D_3PL = mean / stddev;
                keep Cohens_D_3PL MeanDif_3PL;
run;

%end;

%else %do;

Data Effect_Size_3PL;
        set Ttest3pl_stats;
                if _n_ = 3;
                MeanDif_3PL = mean;
                Cohens_D_3PL = (mean - .5) / stddev;
                keep Cohens_D_3PL MeanDif_3PL;
run;

%end;

/*Merge Results*/

Data ALL;
        merge
                Three_PL_TTEST Effect_Size_3PL Ref_3PL_C Ref_3PL_C_Corr
Foc_3PL_C Foc_3PL_C_Corr
                Ref3PL_Mean Foc3PL_Mean Ref3PL_Count_SD Foc3PL_Count_SD
Ref3PL_Count_SD_Adj Foc3PL_Count_SD_Adj;
                                Uniform=&UNI;
                                NonUniform=&NUNI;
                                Impact=&IMPACT;
                                RefGrpSS=&RGSS;
                                SampSize=&SS;
                                NumItems=&NITEMS;
                                BundleSize=&BUNDLE;
Run;
```

```
Proc Append data=ALL out=ALL_RESULT force;
Run;

/*End Do Loops*/

%END;
%END;
%END;
%END;
%END;
%END;
%END;
%END;

%MEND SIMU;

%SIMU(500);
QUIT;

/*Save File*/

data outfiles.FINAL_RESULTS;
        set All_result;
run;
```

APPENDIX C: PARAMETER GENERATION SYNTAX

```
proc iml;
	call randseed(0) ;

	mu = 0;
	sigma = 0.2;

	a_ = j(&NITEMS,1) ;
	a = j(&NITEMS,1) ;

	b_=j(&NITEMS,1) ;
	b=j(&NITEMS,1) ;

	c=j(&NITEMS,1,.2) ;

	theta=j(&SS,1) ;
	thetaNEW=j(&SS,1) ;

	Group=j(&SS,&NITEMS+1) ;

	call randgen(theta, "Normal") ;

	do i=1 to &NITEMS;
		a_[i,1]= exp(mu) * rand('LOGNORMAL')**sigma;
	end;

	call randgen(b_, "Normal") ;

	do i=1 to &NITEMS;
		do until(b_[i,1]<3 & b_[i,1]>-3);
			b_[i,1]=rand('NORMAL');
		end;
	end;

	thetaNEW[1:&RGSS*&SS] = theta[1:&RGSS*&SS]+0 ;
	thetaNEW[&RGSS*&SS+1:&SS] = theta[&RGSS*&SS+1:&SS]-&IMPACT ;

	create a_ from a_ [colname ='a_'];
	append from a_;

	create b_ from b_ [colname ='b_'];
	append from b_;

	create c from c [colname ='c'];
```

```
append from c;

create thetaNEW from thetaNEW [colname ='thetanew'];
append from thetaNEW;

do i=1 to &NITEMS;
        if i <= &NITEMS*&BUNDLE then do;
                a[i,1] = a_[i,1]+&NUNI;
                b[i,1] = b_[i,1]+&DPI;
        end;
        else do;
                a[i,1] = a_[i,1];
                b[i,1] = b_[i,1];
        end;
end;

create a from a [colname ='a'];
append from a;

create b from b [colname ='b'];
append from b;


do i=1 to &RGSS*&SS;
        do j=1 to &NITEMS;
                        d=-1.7 ;
                        Rp=c[j]+((1-c[j])/(1+exp(d*a[j]*(thetaNEW[i]-b_[j])))) ;
                        u=rand('Uniform');
                        if Rp<u then Group[i,j]=0;
                        if Rp>=u then Group[i,j]=1;
                        Group[i,&NITEMS+1] = 0;
        end;
end;

do i=&RGSS*&SS+1 to &SS;
        do j=1 to &NITEMS;
                        d=-1.7 ;
                        Fp=c[j]+((1-c[j])/(1+exp(d*a_[j]*(thetaNEW[i]-b[j])))) ;
                        u=rand('Uniform');
                        if Fp<u then Group[i,j]=0;
                        if Fp>=u then Group[i,j]=1;
                        Group[i,&NITEMS+1] = 1;
        end;
end;

create Group from Group;
```

```
        append from Group;

run;
quit;

data Rpar;
        merge a b_ c;
run;

data Fpar;
        merge a_ b c;
run;
```

Kevin J. Cappaert
Curriculum Vitae

Contact

2400 E. Hartford Ave, Enderis 785
University of Wisconsin, Milwaukee
Milwaukee, WI 53211
cappaer3@uwm.edu

Research Interests

Educational statistics and measurement, ability estimation, differential bundle functioning, differential item functioning, multidimensionality, item response theory, missing data

Education

2009-2014	Ph.D. Educational Psychology	University of Wisconsin - Milwaukee
	Concentration: Educational Statistics and Measurement Advisor: Cindy Walker, PhD	Milwaukee, WI 53201
	Dissertation: "Dissecting the Impact of DIF/DBF on Ability Estimation and Person Fit." Committee: Cindy Walker, Razia Azen, Holmes Finch, Raji Swaminathan, Bo Zhang	
2007-2009	M.A. Cognitive and Social Psychology Certificate - Institutional Research Advisor: Thomas Holtgraves, PhD	Ball State University Muncie, IN 47306
2002-2007	B.S. Psychology	University of Wisconsin – Whitewater
	B.S. Sociology	Whitewater, WI 53190

Professional Experience

2013-2014	Research Assistant, Consulting Office for Research and Evaluation University of Wisconsin - Milwaukee, Milwaukee, WI Supervisor: Cindy Walker, PhD
2011-Present	Statistical Consultant Center for Self Sufficiency, Shorewood, WI
2013	Adjunct Instructor, Educational Statistical Methods II (ED PSY 724) University of Wisconsin - Milwaukee, Milwaukee, WI
2011-2013	Teaching Assistant, Educational Statistical Methods II (ED PSY 724) University of Wisconsin - Milwaukee, Milwaukee, WI Supervisors: Wen Luo, PhD; Cindy Walker, PhD; Razia Azen, PhD
2012	Adjunct Instructor, Behavioral Science Statistics and Research Design (CON 630) Mount Mary College, Milwaukee, WI

2012	Research Assistant University of Wisconsin - Milwaukee, Milwaukee, WI Supervisor: Nadya Fouad, PhD
2009-2011	Research Assistant, Consulting Office for Research and Evaluation University of Wisconsin - Milwaukee, Milwaukee, WI Supervisor: Cindy Walker, PhD
2009	Institutional Research Intern, Academic Assessment and Institutional Research Ball State University, Muncie IN Supervisor: Sherry Woosley, PhD
2007-2009	Research Assistant, Psychological Science Department Ball State University, Muncie IN Supervisor: Thomas Holtgraves, PhD

Publications

Luo, W., Cappaert, K.J. Modeling partially cross-classified multilevel data. (Revised Manuscript Under Review).

Singh, R., Fouad, N. A., Fitzpatrick, M .E., Liu, J. P., Cappaert, K. J., & Figuereido, C. (2013). Stemming the tide: Predicting women engineers' intentions to leave. *Journal of Vocational Behavior, 83,* 281-294.

Habeck, T.M., Rice, N.E., & Cappaert, K.J. (2012). Examining perspectives on reading in general, reading and special education: Are the fields ready to respond to response to intervention? *Wisconsin State Reading Journal, 50,* 1-9.

Schapira, M. M., Walker, C. M., Cappaert, K. J., Ganschow, P. S., Jacobs, E. A., McGinley, E. L., Del Pozo, S. , … Fletcher, K. E. (2012). The Numeracy Understanding in Medicine Instrument (NUMI): A measure of health numeracy developed using item response theory. *The Journal of Medical Decision Making, 32,* 851-865.

Walker, C. M., Zhang, B,, Banks, K., & Cappaert, K. (2012). Establishing effect size guidelines for interpreting the results of differential bundle functioning analyses using SIBTEST. *Educational and Psychological Measurement, 72,* 415-434.

Holtgraves, T., McNamara, P., Cappaert, K., & Durso, R. (2010). Linguistic correlates of asymmetric motor symptom severity in Parkinson's Disease. *Brain and Cognition, 72,* 189-196.

Presentations

Cappaert, K. J., Finch, W. H., & Walker, C. M. (2014, April). *Simultaneous uniform and non-uniform DBF detection: A MIMIC Model approach.* Presented at the 2014 annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Cappaert, K. J., & Wen, Y. (2014, April). *Missing not at random: A cause of DIF?* Presented at the 2014 annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Cappaert, K. J., Walker, C. M., & Zhang, B. (2013, April). *Partial cancellation in differential bundle functioning: Influences on the detection rate, ability estimates, and the standard error of the beta statistic*. Presented at the 2013 annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cappaert, K., Peterson, J., & Luo, W. (2012, April). *Modeling partially cross-classified multilevel data.* Presented at the 2012 annual meeting of the American Educational Research Association, Vancouver, British Columbia.

*Walker, C. M., Zhang, B,, Banks, K., & Cappaert, K. (2011, April). *Establishing effect size guidelines for interpreting the results of differential bundle functioning analyses using SIBTEST*. Presented at the 2011 annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Cappaert, K. J., Holtgraves, T. M., McNamara, P., Felton, A. D., & Waters, J. M. (2008, May). *Cognitive and emotional correlates of asymmetric motor symptom severity in Parkinson's Disease*. Poster session presented at the annual convention of the Association for Psychological Science, Chicago, IL.

Felton, A. D., Waters, J. M., Cappaert, K. J., & Holtgraves, T. M. (2008, May). *The right hemisphere: A significant role in the comprehension of indirect replies*. Poster session presented at the annual convention of the Association for Psychological Science, Chicago, IL.

Waters, J. M., Cappaert, K. J., Felton, A. D., & Holtgraves, T. M. (2008, May). *Differential hemispheric activation and creativity*. Poster session presented at the annual convention of the Association for Psychological Science, Chicago, IL.

**NOTE: An earlier version of this paper was presented at the 7th Conference of the International Test Commission, Hong Kong (2010).*

Manuscripts in Progress

Cappaert, K. J., Finch, W. H., & Walker, C. M. *Simultaneous uniform and non-uniform DBF detection: A MIMIC Model approach.*

Cappaert, K. J., Walker, C. M., & Zhang, B. *Partial cancellation in differential bundle functioning: Influences on the detection rate, ability estimates, and the standard error of the beta statistic.*

Cappaert, K. J., & Wen, Y. *Missing not at random: A cause of DIF?*

Technical and Evaluation Reports (Abbreviated List)

* Turner, A.M., Walker, C., & Cappaert, K. Woods, S., Kapper, L., & Porterfield, M. (2010-2011, 2011-2012, 2012-2013). *Year 3 Better Family Life Inc. teen pregnancy prevention grant performance measurement and fidelity analysis report.* Shorewood, WI: Center for Self Sufficiency.

**Polifka, S., Turner, A.M., Walker, C., & Cappaert, K. (2011-2012, 2012-2013). *Center for Self-Sufficiency: Adult healthy marriage and youth healthy relationships education grant evaluation report*. Shorewood, WI: Center for Self Sufficiency.

Graunke, S., Lacy, K., Cappaert, K., & Costomiris, R. (2009*). Fall 2008 making achievement possible survey: Sophomore transition summary report*. Retrieved from http://cms.bsu.edu/About/AdministrativeOffices/Assessment/Surveys.aspx\.

Graunke, S., Woosley, S., Sitzman, J., Cappaert, K., & Costomiris, (2009). *Fall 2008 making achievement possible survey: First-year student summary report.* Retrieved from http://cms.bsu.edu/About/AdministrativeOffices/Assessment/Surveys.aspx\.

**NOTE: Similar reports prepared for the following sites/programs: Mission West Virginia, OIC of America, OIC of Broward, OIC of South Florida, and Trinity Church.*

***NOTE: Similar reports prepared for the following sites/programs: Mission West Virginia, and Operation Keepsake.*

Professional Activities

2013-2014	Treasurer, Co-founder Educational Psychology Student Association University of Wisconsin - Milwaukee
2011-2014	Reviewer National Council on Measurement in Education Graduate Poster Session
2010-2011	Co-President Educational Statistics and Measurement Student Association University of Wisconsin - Milwaukee
2008-2009	Cognitive and Social Psychology Graduate Representative Psychological Sciences Department Ball State University

Professional Qualifications

Software Packages: BILOG, BMIRT, DIMTEST, FORTRAN, Java, LISREL, Mplus, MULTILOG, NOHARM, SAS, SIBTEST, SPSS, TESTFACT

Grants/Awards

2014	Department of Educational Psychology Graduate Research Presentation Travel Grant
2011, 2012, 2013	Graduate Student Travel Award
2009, 2010	Chancellor's Graduate Student Award

Professional Organizations

Alpha Kappa Delta
American Educational Research Association
    Division D: Measurement and Research Methodology
    Division H: Research, Evaluation, & Assessment in Schools
American Psychological Association
    Division 5: Evaluation, Measurement, Statistics
American Statistical Association
National Council on Measurement in Education
Psi Chi
Psychometric Society