A DIFFERENTIAL RESPONSE FUNCTIONING FRAMEWORK FOR UNDERSTANDING ITEM, BUNDLE, AND TEST BIAS

ROBERT PHILIP SIDNEY CHALMERS

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

October 2016

Abstract

This dissertation extends the parametric sampling method and area-based statistics for differential test functioning (DTF) proposed by Chalmers, Counsell, and Flora (2016). Measures for differential item and bundle functioning are first introduced as a special case of the DTF statistics. Next, these extensions are presented in concert with the original DTF measures as a unified framework for quantifying differential response functioning (DRF) of items, bundles, and tests. To evaluate the utility of the new family of measures, the DRF framework is compared to the previously established simultaneous item bias test (SIBTEST) and differential functioning of items and tests (DFIT) frameworks. A series of Monte Carlo simulation conditions were designed to estimate the power to detect differential effects when compensatory and non-compensatory differential effects are present, as well as to evaluate Type I error control. Benefits inherent to the DRF framework are discussed, extensions are suggested, and alternative methods for generating composite-level sampling variability are presented. Finally, it is argued that the area-based measures in the DRF framework provide an intuitive and meaningful quantification of marginal and conditional response bias over and above what has been offered by the previously established statistical frameworks.

Acknowledgements

This dissertation would not have been possible without the support from my two supervisors, Dr. David Flora and Dr. Jolynn Pek, who provided me the tools, guidance, and freedom I needed to explore my research and development interests. I am also grateful for those who provided feedback on earlier versions of this work; namely, Matthew Sigal and Victoria Ng, and for my committee members who dedicated their time and effort to critique this body of work.

Contents

List of Tables

List of Figures

List of Symbols

Symbols	Definition
N	number of participants
J	number of items
K	number of item response categories
G	number of independent groups
c	composite score for one item
C	composite score containing more than one item
$\mathbf{Y}$	matrix of observed item responses
y_{ij}	response from the ith row and jth column of $\mathbf{Y}$
$\boldsymbol{\alpha}$	vector of item slope/discrimination parameters
$\boldsymbol{\theta}$	vector of latent trait values
$g(\boldsymbol{\theta})$	probability density function for $\boldsymbol{\theta}$
m	number of elements in $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$
δ	item intercept
$\boldsymbol{\psi}$	vector of item parameters for a single item
p	number of unique elements in $\boldsymbol{\psi}$
$\boldsymbol{\Psi}$	vector of item parameters for full model
P	number of unique elements in $\boldsymbol{\Psi}$
$\boldsymbol{\Sigma}(\boldsymbol{\Psi}\|\boldsymbol{Y})$	parameter variance-covariance matrix
$\boldsymbol{\Psi}^*$	matrix of plausible parameter estimates
M	number of parametric samples drawn
R	number of independent Monte Carlo replications

1 Introduction

Psychological traits, attributes, abilities, pathologies, and so on are often understood through modeling behavioral response data from educational and psychological tests. The most popular method for collecting test data in psychology and education is to administer a test or questionnaire whereby individuals are 'scored' according to their responses to the item-level stimuli. These responses to individual items are then simplified by forming some meaningful composite value (or set of values) that summarizes the overall response behavior, often to serve as an approximate quantification of the underlying construct which the test purports to measure.

Numerous methods have been proposed to create summary scores of test performance. The simplest and most popular method of scoring a test is to simply tally the number of correct (or positively endorsed) item responses and interpret this sum score as a representation of the overall performance on the test. This approach is commonly used in popular psychological assessment tools (e.g., Beck Depression Invetory-II; Beck, Steer, & Brown, 1996) and in aptitude measurement situations, such as when grading course examination material. However, this simple scoring method rests on a set of important and strong assumptions. Namely, that the items are internally

reliable; the items are exchangeable and of equal importance; the test is unidimensional (i.e., the items measure only one construct or latent trait); every item has the same functional relationship with the unobserved trait (e.g., linear); the test is valid (i.e., items measure what they are intended to measure); and the test items are not unfairly influenced by cultural or personal demographic information (e.g., content does not favor female participants over equally abled male participants).

Psychometricians have long recognized the measurement limitations associated with unweighted composite scoring procedures such as the previously described sum-score method (e.g., Lord & Novick, 1968). These early objections are one reason why the *true-score* test-analysis paradigm was developed to form what is now considered 'classical test theory' (CTT; Crocker & Algina, 1986; Lord & Novick, 1968). Very soon after CTT was realized, psychometricians turned their focus towards understanding item response phenomena directly at the item level by utilizing statistical modeling methods such as linear factor analysis (McDonald, 1999) and item response theory (Lord, 1980). Factor analysis and item response theory generally belong to same family of latent variable statistical models used to model the relationship between latent variables and response stimuli. Compared to CTT, latent trait methods offer a more rigorous and testable statistical paradigm for understanding the relationships between latent variables and items (Borsboom, 2005).

By and large, the latent variable conceptualization and statistical modeling framework of item response data dominates the current zeitgeist in psychometric research. The methods themselves include powerful techniques for determining the underlying structure and goodness-of-fit of the tests and their respective item response models, while also providing helpful techniques to deter-

mine whether items are biased across different populations. Detecting response bias in individual items, item bundles, and tests as a whole using methods from the item response theory paradigm will be the primary theme of this dissertation.

1.1 Background Information

Item response theory (IRT) is a general framework of probabilistic models which provide specific structures to explain variation in observed item-response data. The framework postulates two distinct and separable entities which, when considered jointly, are responsible for the manifest pattern of responses in a given psychological test. The first entity is the set of unobserved values (or relative standing) on the latent traits, or abilities, which each individual test taker possesses denoted by $\boldsymbol{\theta}$. These latent traits provide a rank ordering of respondents along one or more unobserved continua and represent meaningful constructs which the test attempts to quantify (Reckase, 2009). For example, the locations along a continuum may indicate a level of proficiency in educational settings, such as having some mastery of a mathematical subject matter, or may represent the psychological intensity in the context of measuring psychopathologies (e.g., depression).

The second and often more technical entity from the IRT paradigm is the set of characteristics inherent in the item-level stimuli. Such properties may reflect how difficult or extreme the items are, how well they discriminate individuals along the unobserved continua, whether the item response probability is monotonically related to the latent trait, and so on. Given some functional relationship specifying how these two entities interact, a probabilistic response model can often be

constructed to model or explain an individual's overt response behavior.

Expressing the above ideas more concisely, let $\boldsymbol{\theta}$ represent an m-dimensional vector of latent trait values, and let $\boldsymbol{\psi}$ represent a p-dimensional vector of item parameters. The probability of responding with category k for a single item, where $k = 0, 1, \ldots, K-1$ and K is the total number of response categories, is $P(y = k|\boldsymbol{\theta}, \boldsymbol{\psi})$. The probability mapping from $\boldsymbol{\theta}$ to $P(y = k|\boldsymbol{\theta}, \boldsymbol{\psi})$ is often taken to be some parametric model with a monotonically increasing function given $\boldsymbol{\theta}$. Depending on the parameters modeled, items can be organized to include more than one latent trait (i.e., have different discrimination properties), can include background information about the population distributions (e.g., latent regression models; Adams, Wilson, & Wu, 1997), can be constructed to have non-monotonic relationships with the latent traits, and so on.

Although IRT conceptualizes response behavior in terms of probabilistic models alone, several ancillary functions can be specified for capturing other useful properties of the items. One such function is the expected item score,

$$S(c|\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{k=0}^{K-1} k \cdot P(y = k|\boldsymbol{\theta}, \boldsymbol{\psi}), \tag{1.1}$$

which transforms the K probability functions for a given item into a single function denoting which observed response (c) is expected given the respective item parameters and ability values. In the case where the item response function contains only two categories (e.g., true or false, correct or incorrect; scored as $y = 1$ and $y = 0$, respectively), the expected score function is equivalent to $P(y = 1|\boldsymbol{\theta}, \boldsymbol{\psi})$.

To demonstrate the joint behavior of multiple items in a test, the expected score function in

Equation 1.1 can be extended to include a bundle of items by summing over the J possible items in the test, indexing the desired items with a binary indicator function $I(j)$:

$$T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}) = \sum_{j=1}^{J} S(c|\boldsymbol{\theta}, \boldsymbol{\psi} = \boldsymbol{\Psi}_j) \cdot I(j), \tag{1.2}$$

where $C = \sum_{j=1}^{J}(K_j - 1) \cdot I(j)$ and $\boldsymbol{\Psi}$ represents the vector of item parameters for the full model. If item j should be included in the composite response function above then $I(j) = 1$, otherwise $I(j) = 0$. When $I(j) = 1$ for only one item then (1.2) necessarily reduces to (1.1); hence, (1.1) can be understood as an item bundle of length one. Additionally, when all J items are included in (1.2) then this equation will represent the expected test score function (Embretson & Reise, 2000). As we will see in the next sections, as well as in subsequent chapters, these simple functions are of pivotal importance investigating response bias between different populations of respondents.

1.2 Item Response Theory

IRT methods consist of a wide array of probabilistic models for representing the structure of response patterns by using quantitative item-level properties and latent trait values. One such parametric model often used to model binary response data (where the number of categories $K = 2$) is the multidimensional two parameter logistic model, or M2PL (Reckase, 1997). For a binary item with only two potential empirical realizations ($y = 0$ or $y = 1$), the M2PL model has the form

$$P(y = 1|\boldsymbol{\theta}, \boldsymbol{\psi}) = P(y = 1|\boldsymbol{\theta}, \boldsymbol{\alpha}, \delta) = \frac{\exp(\boldsymbol{\alpha}'\boldsymbol{\theta} + \delta)}{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{\theta} + \delta)}, \tag{1.3}$$

where $\boldsymbol{\alpha}$ is an $m \times 1$ vector of discrimination or slope parameters that combine with the commensurate latent trait scores $\boldsymbol{\theta}$ and δ is the overall item intercept representing how 'easy' an item is to answer. Because binary items have only two empirical realizations ($y = 1$ or $y = 0$), the complementary probability for $y = 0$ is $P(y = 0|\boldsymbol{\theta}, \boldsymbol{\psi}) = 1 - P(y = 1|\boldsymbol{\theta}, \boldsymbol{\psi})$. When additional parameter constraints are added to the M2PL model other popular IRT models can be obtained. For instance, when $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ contain only one element (i.e., are scalar values) the M2PL model reduces to the unidimensional 2PL model, and when $\boldsymbol{\alpha} = 1$ the 2PL model further reduces to the Rasch or 1PL model (Rasch, 1960). Although there are many additional item response models for dichotomous and polytomous items, we need only focus on the M2PL to grasp the subsequent methodological developments for the purpose of this body of work.

To better understand the effect of the parameters in (1.3), Figure 1.1 was constructed to visualize prototypical probability response curves. These figures depict the isolated effects of the slope and intercept for a unidimensional IRT model. In practice, however, response models may contain any combination of these parameters. As demonstrated in Figure 1.1, increasing the slope parameters has the effect of increasing the probability of positive endorsement rate given θ, where the rate of change is greatest at the function's inflection point. Items with larger slope parameters tend to discriminate between individuals with different latent trait values better than items with smaller slopes, especially near the inflection point. Items with differing intercepts, on the other hand, have systematically shifted response curves, and therefore resemble the effect of item 'easiness'. Easier items are associated with large positive intercept values and indicate that all individuals have a

higher probability to answer positively (i.e., respond with $y = 1$) compared to more difficult items.

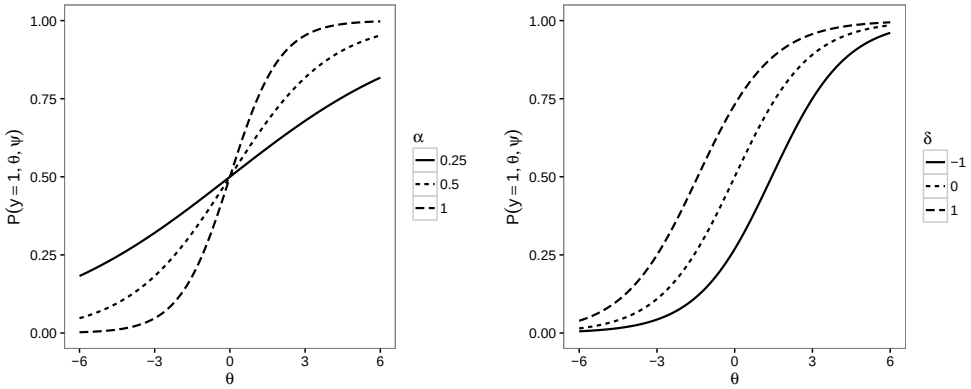

Figure 1.1: Expected probability functions with differing parameters for unidimensional IRT models. The left figure demonstrates the effect of modifying the slope (α), while the right figure demonstrates the effect of modifying the intercepts (δ).

1.2.1 Estimation

In practice, item parameters must be estimated using available response data by optimizing some discrepancy function. One extremely popular approach which currently is the de facto standard is maximum-likelihood estimation (Reckase, 2009). Maximum-likelihood estimates are obtained after maximizing the likelihood function

$$L(\mathbf{Y}|\Psi, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(y = 1|\boldsymbol{\theta}, \boldsymbol{\psi} = \boldsymbol{\Psi}_j)^{y_{ij}} P(y = 0|\boldsymbol{\theta}, \boldsymbol{\psi} = \boldsymbol{\Psi}_j)^{(1-y_{ij})}, \tag{1.4}$$

where y_{ij} represents the observed response to the jth item by person i, $\mathbf{Y}$ is an $N \times J$ a matrix containing all binary response patterns for each respective individual and item combination, and

$\boldsymbol{\Psi}_j$ is the subset of parameters relevant to the jth item. Because the $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ parameters are not sufficiently estimable from the data alone, the $\boldsymbol{\theta}$ terms are often treated as random effects to be integrated out of the likelihood function (Bock & Lieberman, 1970). The marginal likelihood after integrating across $\boldsymbol{\theta}$, given the probability density function $g(\boldsymbol{\theta})$ (typically understood to be multivariate normal density function), is

$$L(\mathbf{Y}|\boldsymbol{\Psi}) = \prod_{i=1}^{N}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{j=1}^{J}P(y=1|\boldsymbol{\theta},\boldsymbol{\psi}=\boldsymbol{\Psi}_j)^{y_{ij}}P(y=0|\boldsymbol{\theta},\boldsymbol{\psi}=\boldsymbol{\Psi}_j)^{(1-y_{ij})}g(\boldsymbol{\theta})d\boldsymbol{\theta}\right). \tag{1.5}$$

Equation 1.5 (or its more numerically manageable logarithmic counterpart) is the objective function to be optimized to obtain P maximum-likelihood parameter estimates; however, its form is restricted by two main computational issues.

The first issue with optimizing Equation 1.5 directly is that the dimensionality of the parameter space grows exponentially as more items are included, where numerically evaluating the integrals becomes overwhelmingly computationally demanding (Bock & Lieberman, 1970). As Bock and Aitkin (1981) demonstrated, however, a powerful solution to this optimization problem is to decompose the likelihood function into a more manageable complete-data likelihood form by estimating the parameters with an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). The EM algorithm effectively solves the issue of estimating parameters in longer tests because maximization can be performed independently over the J items using the more manageable complete data-table rather than over the problematic P-dimensional parameter space. The second issue relates to the dimensionality of $\boldsymbol{\theta}$ in that as the number of latent dimensions m in-

creases, the complexity of the integration grid increases exponentially (Cai, 2010; Chalmers & Flora, 2014). This issue will be discussed in more detail in Chapter 4 where the topic of multidimensional differential response functioning is discussed.

1.2.2 Multiple-Group IRT and Linking

Test developers are often interested in whether their tests behave the same in different populations. Ideally, the expected response equations expressed in (1.1) and (1.2) should be identical in all populations so that response differences can be explained solely in terms of distributional differences with respect to $\boldsymbol{\theta}$. However, before comparing the behavior of expected response functions across different populations it is important to first place the groups on a similar metric; this process is generally termed *linking* or *equating*. The purpose of linking is to take into account population differences in the form of latent trait distributions so that item parameters and observed responses are on a comparable metric. For instance, when an aptitude test is administered to two different age groups in a high-school setting (e.g., seniors versus freshman), we would naturally expect the older group to perform better overall on the test than the younger group. Therefore, composite measures will naturally reflect that the older individuals perform better on the test compared to the younger population. However, overt differences in performance do not necessarily imply that the test items are biased. In this example, group differences in observed sum scores, for instance, might only reflect that the older group has higher ability on average than in the younger group. Linking methods take into account the latent distribution effects by re-scaling the item parameters in each respective

group, thereby accounting for the effect of latent trait distributional differences (Kolen & Brennan, 2004).

Suppose now that there are $G = 2$ groups under investigation. The process of linking parameters between two populations often starts by estimating separate sets of IRT models for each population. After applying a linking method which specifies how differences in latent trait values affect the observed responses, item parameters from one group are then scaled according to the metric of the other group. Numerous linking methods have been proposed (e.g., see Kolen & Brennan, 2004), but these generally fall into two distinct classes: complete-item methods and anchor-based methods. Complete-item linking methods attempt to rescale all of the test items using all available item information simultaneously, while anchor-based methods attempt to rescale the items by using only a subset of items which are assumed to have no bias. Linking tests using complete-item methods is generally not recommended for investigating response bias primarily because the bias in the item response functions will generally contaminate the linking process (Millsap, 2011).

A model fitting approach based on multiple-group estimation techniques is one possible approach to linking the G groups. When multiple groups are included in the model, the following likelihood equation is maximized instead of (1.5):

$$L(\mathbf{Y}|\boldsymbol{\Psi}) = L(\mathbf{Y}|\boldsymbol{\Psi}, g = 1)L(\mathbf{Y}|\boldsymbol{\Psi}, g = 2)\cdots L(\mathbf{Y}|\boldsymbol{\Psi}, g = G). \tag{1.6}$$

In this equation a selection of parameters can be constrained to be equal across groups while other parameters can freely vary. Thissen, Steinberg, and Wainer (1993) demonstrated that the respective group parameters can be equated by setting a small selection of item parameters to be equal across

groups during estimation so that one groups' hyper-parameters can be freely estimated. The hyper-parameters which are freely estimated typically include the mean and variance of the latent trait distributions. By selecting a set of anchor items and freely estimating these hyper-parameters all other item parameters will be naturally expressed in the same scale within the respective groups. In this sense, likelihood-based parameter linking is built into the multiple-group maximum likelihood estimation framework of IRT models. Consequently, multiple-group estimation methodology can be used to investigate response bias directly by comparing nested models via likelihood-based statistics such as the likelihood-ratio test, information criteria statistics (e.g., AIC), and so on.

Another benefit of the likelihood-based linking approach is that parameter sampling variability can be approximated by forming a second-order Taylor series estimate of the log-likelihood function. After computing an estimate for the parameter covariance matrix $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$, standard errors and associated large sample Wald (1943) tests can be obtained to evaluate the reliability (i.e., sampling variability) of the estimated parameters. Under mild regularity conditions and large sample sizes, $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ will provide a sufficient approximation to the shape of the log-likelihood function at the ML estimates, thereby providing a more manageable form of the likelihood function's behavior (Fisher, 1925). As is demonstrated below, as well as in subsequent chapters, the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix has important applications when studying differential item functioning.

1.3 Differential Item Functioning

IRT has an intuitive and meaningful conceptualization regarding how unbiased items should behave. For simplicity, I will focus only on models for two groups of interest in the following descriptions; however, comparing more than two groups is possible. Groups are often selected based on prior demographic information, such as gender or education level, but groups may also be formed artificially by combining other sources of information (such as categorizing individuals above or below some threshold from an auxiliary test or survey). When comparing response functions between groups, one group is generally denoted as the *reference* group while the other groups are considered *focal* groups. The reference group hyper-parameters are typically assumed to form a standard Gaussian distribution, while the Gaussian distribution hyper-parameters in the focal group are freely estimated to allow for proper item parameter linking.

Using the subscripts R and F to indicate which response function is from the reference or focal groups, respectively, items are considered unbiased when

$$\forall \boldsymbol{\theta} : \; S(c|\boldsymbol{\theta}, \boldsymbol{\psi}_R) = S(c|\boldsymbol{\theta}, \boldsymbol{\psi}_F). \tag{1.7}$$

Expressing Equation 1.7 in words, an item is considered unbiased when all individuals in different groups, with equivalent latent trait values, receive equivalent expected scores. When (1.7) is not true then the item will favor one of the groups at one or more levels of $\boldsymbol{\theta}$; hence, the item contains bias.

Numerous statistical techniques have been proposed to assess the veracity of (1.7) through the

use of null hypothesis significance tests (NHST). These techniques have generally appeared in the literature under the umbrella term differential item functioning (DIF). The study of DIF has a long history in the psychometrics literature, and numerous statistical methods have been proposed to test different types of DIF. Primarily, DIF detection methods have been broken into two distinct classes which are based on whether the response curves have potential *uniform* or *non-uniform* effects (Millsap, 2011). Uniform differences occur when item response curves only differ by a constant arising from differences in intercept terms, thereby shifting the response probability function systematically in one direction[1]. Non-uniform DIF, on the other hand, occurs when the response functions are allowed to differ by more than an intercept term, allowing the probability functions to (potentially) cross at one or more locations. Consequently, with respect to non-uniform DIF, the overall difference between the response curves may vary dramatically depending on the level of $\boldsymbol{\theta}$; therefore, some latent trait levels will be affected more than others. These DIF effects can be seen in Figure 1.2, which depicts two types of parametric IRT models demonstrating uniform (DIF Item 1 and 2) and non-uniform (DIF Item 3 and 4) DIF effects.

[1]For readers already familiar with Rasch or 1PL models, note that near the tails of the expected probability functions the difference between the respective groups approaches 0 due to the non-linearity in the functions (the difference in the logit of the response functions will be constant, however). Therefore, even for these simple IRT models the response bias is technically non-uniform and varies along $\boldsymbol{\theta}$.

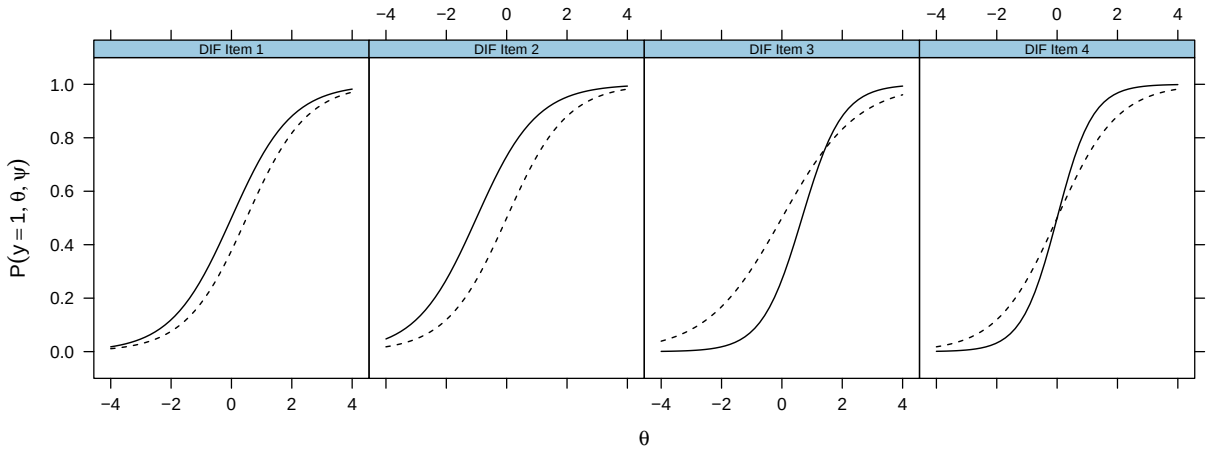

Figure 1.2: Probability functions for uniform (left two images) and non-uniform (right two images) DIF, where the latent trait values are organized along the x-axis. In these graphics the solid lines following a sinusoidal pattern refer to the response function of the reference group, while the dashed line represents the response function of the focal group.

1.4 Differential Bundle and Test Functioning

Much like DIF, differential bundle functioning (DBF) and differential test functioning (DTF) can be understood as a type of response bias between the expected scoring functions (cf. Equation 1.7). More formally, DBF or DTF are said to occur when the following equation does not hold,

$$\forall \boldsymbol{\theta} : \; T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) = T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F). \tag{1.8}$$

In the equation above, C is a composite score and T_B is a scoring function containing B item response functions (where $1 < B \leq J$) indexed from the $I(j)$ indicator function in (1.2). When DBF or DTF is present then $T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) \neq T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F)$ at one or more locations on$\boldsymbol{\theta}$. Note that DTF and DIF are in fact special cases of DBF: DTF is realized when the composite functions in (1.8) contain all J items, and DIF is realized when only a single item is indexed to form the required single item bundle.

Unfortunately, constructing statistical tests for DTF and DBF is not as straightforward as it is for DIF. When inferring population-level DIF from sample data, any differences in the expected item scoring functions can be understood primarily as population differences in the $\boldsymbol{\psi}$ parameter sets across groups for each respective item in isolation because population parameter differences will necessarily result in unequal expected item response curves. Methods for detecting DTF and DBF, on the other hand, are more complicated because they necessarily require the evaluation of nonlinear aggregate functions implied by the estimated $\boldsymbol{\Psi}$ vectors. The composite functions consist of a mixture of nonlinear response functions which are *implied* rather than *estimated*. As

such, differences in population parameters with respect to particular items do not easily translate to population differences in the composite response functions. In other words, the presence of DIF is necessary for DBF and DTF to occur but DIF is not sufficient to establish their existence, particularly at any given $\boldsymbol{\theta}$ location. Depending on the direction and magnitude of DIF within each item, DTF or DBF may manifest in different ways, such as being present across the entire range of $\boldsymbol{\theta}$ (analogous to uniform DIF), present at only specific levels of $\boldsymbol{\theta}$ (analogous to non-uniform DIF), or, in the case of no DTF and DBF, not present at all across any level of $\boldsymbol{\theta}$ (complete cancellation). The last scenario is equivalent to a situation in which there is no DIF in any items, thus indicating that the test is truly unbiased across groups regardless of the $\boldsymbol{\theta}$ level. Tests with no DTF are the gold standard which practitioners developing psychometric instruments should strive for.

1.4.1 Compensatory and Non-Compensatory DTF

Another important topic to consider when measuring DTF and DBF effects is how differences across response functions should be aggregated given the $\boldsymbol{\theta}$ parameters. Integrating (or marginalizing) over the latent trait distributions is often done to help automate the detection of differential item, bundle, and test effects rather than focusing on particular $\boldsymbol{\theta}$ locations. In the case where the response functions between the groups intersect at one or more locations of $\boldsymbol{\theta}$, practitioners must decide whether the total differential effects should be allowed to compensate across the response functions or whether a definition based on the total magnitude of the difference should be adopted.

These ideas can be formally expressed as

$$\int \left[T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) - T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F)\right] g(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{1.9}$$

and

$$\left(\int \left[T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) - T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F)\right]^2 g(\boldsymbol{\theta}) d\boldsymbol{\theta}\right)^{1/2} \tag{1.10}$$

for the respective bundle scoring functions. Equation 1.9 represents an overall compensatory effect across the latent trait distributions, where differences across the latent trait values may cancel out if the functions cross at one or more locations. For example, positive differences at a higher $\boldsymbol{\theta}$ level may combine with negative differences in a lower $\boldsymbol{\theta}$ level to create a small bias effect overall. Equation 1.10, on the other hand, represents the average discrepancy between response functions in absolute terms and therefore is non-compensatory in nature. The $g(\boldsymbol{\theta})$ term is a weight function used to obtain the weighted average across the difference in the response functions given $\boldsymbol{\theta}$. For simplicity, all values of $g(\boldsymbol{\theta})$ can be set to a constant value with the property that $\int g(\boldsymbol{\theta}) = 1$, although other weighting functions can be used if different $\boldsymbol{\theta}$ regions are deemed to be of greater or lesser importance (additional information on this topic is presented in Chapter 4). Finally, an alternative definition for non-compensatory effects may be expressed as

$$\int \left|T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) - T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F)\right| g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{1.11}$$

which represents the absolute difference between the scoring functions instead of the deviation form presented in Equation 1.10. Within the literature, non-compensatory item bias has been the primary target for differential functioning statistics because practitioners are often interested in

whether *any* differences exist between response functions rather than if response functions have cancellation effects (Millsap, 2011).

Figure 1.3 presents three possible response curve combinations which can arise when investigating DBF and DTF (including DIF, as a special case of DBF). Each of these images contain the expected total score functions for the focal and reference group to help determine where the bias occurs given $\boldsymbol{\theta}$. For example, if the dotted line were treated as the reference group in the middle graphic of Figure 1.3 then at $\boldsymbol{\theta} = -2$ we would expect the reference group to obtain a total score around 2.5, while the focal group only obtains a total score of 1.5. Therefore, at $\boldsymbol{\theta} = -2$ the reference group has a favorable bias at this $\boldsymbol{\theta}$ location because individuals with equal latent trait values are more like to obtain a higher total score simply because they belong to the reference group. Alternatively, it is easy to use these figures to compare the $\boldsymbol{\theta}$ locations when both groups expected total scores are equal to 2.5 which, for example, occurs at approximately -2 for the reference group and -1 for the focal group.

The images in Figure 1.3 can also be used to demonstrate compensatory and non-compensatory DTF for the above equations. Within the left image there does not appear to be any significant compensatory or non-compensatory DTF effects because the response curves essentially overlap across every $\boldsymbol{\theta}$ location; hence, whether Equation 1.9, 1.10, or 1.11 is used the values would all be relatively close to 0. The middle image, on the other hand, demonstrates a non-compensatory DTF effect but potentially little compensatory DTF effect depending on the range of integration. For example, if the integration range in Equation 1.9 were between $[-4, 4]$ then the positive and nega-

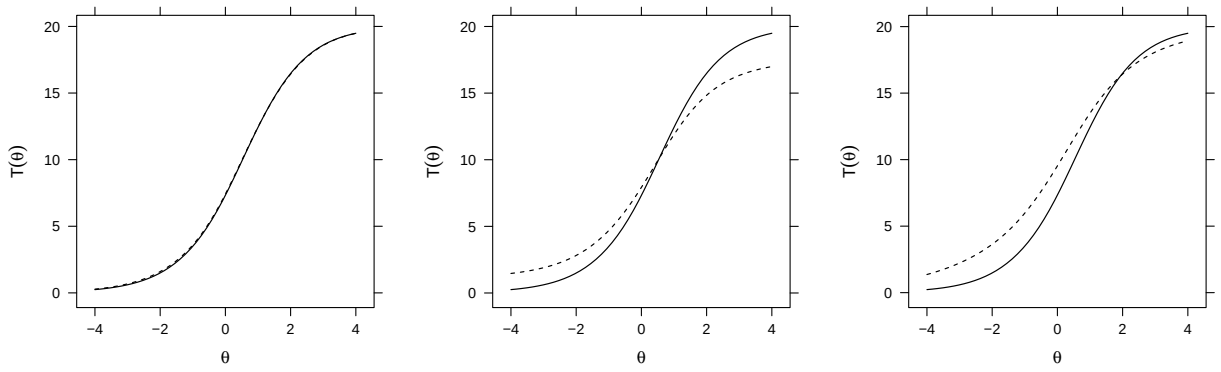

Figure 1.3: Three possible compensatory and non-compensatory expected score plots within the range $\boldsymbol{\theta} = [-4, 4]$. Left image suggests little to no compensatory and non-compensatory effects, middle suggests no compensatory effect but a substantial non-compensatory effect, and the right indicates both compensatory and non-compensatory effects.

tive differences between the response functions would result in an aggregate very close to the value 0. Finally, the rightmost image largely demonstrates compensatory and non-compensatory DTF; however, if the integration range were set within, say, $[-1, 4]$ then there may be little compensatory DTF present.

1.4.2 DIF Amplification and Cancellation

As previously mentioned, a necessary but insufficient requirement for DTF or DBF to occur is the presence of DIF in one or more items[2]. Furthermore, the manner in which the DIF effects propagate

[2]When only one item contains DIF, then non-compensatory DBF and DTF will necessarily occur; however, compensatory DTF and DBF may not.

across multiple items determines the magnitude of the DTF and DBF effects. For example, when the DIF effects combine to cause greater separation between the expected response curves then this combination leads to a phenomenon that has been termed "DIF amplification" (Shealy & Stout, 1993). DIF amplification is akin to the description of DTF and DBF above, but amplification generally implies that the expected response functions become more disparate within the aggregate functions. In turn, the larger the DIF amplification the easier it becomes to detect bias in the composite response functions (in other words, relevant hypothesis tests become more powerful).

DIF cancellation, on the other hand, occurs when multiple items with DIF demonstrate opposite directional effects between the groups, resulting in more subtle and minute differences in the aggregate response functions. In the extreme case when item response functions have perfectly balanced DIF effects, no DTF or DBF may be present (i.e., complete cancellation). This phenomenon is what led Chalmers, Counsell, and Flora (2016) to assert that differential item effects 'might not make a big DIF' in the test as a whole. Namely, if the number of items in the test is large compared to the number of items with DIF, or the DIF effects are small and do not form a large composite difference at any given $\boldsymbol{\theta}$ level (possibly due to DIF cancellation), then the DIF effects may be of little consequence to the test developer because the overall response bias when scoring the test will be negligible. Hence, DBF and DTF should be studied separately in tests which contain known DIF effects. DBF will be more effective at detecting bias in bundles of items that contain DIF, while DTF will quantify the overall impact of the items containing DIF at the test level at the cost of decreased power and increased sampling variability.

The reason why DIF cancellation is important ultimately relates to how $\hat{\boldsymbol{\theta}}$ estimates are built for each group. When forming secondary estimates for individuals (i.e., assigning them suitable scores) the presence of DIF, DBF, and ultimately DTF will cause the secondary estimates to be either be systematically too high or too low, given where the associated population $\boldsymbol{\theta}$ value is. When cancellation occurs, however, neither the focal nor the reference group will have systematic bias in the $\hat{\boldsymbol{\theta}}$ estimates because the values in the population contains no bias. Therefore, each group may be scored separately given the associated fixed item parameters and the resulting $\hat{\boldsymbol{\theta}}$ values may be interpreted as free from differential effects.

More realistically, however, DTF and DBF are more likely to be substantial in clustered locations rather than across the complete $\boldsymbol{\theta}$ range. Hence, the presence of DBF and DTF may not necessarily invalidate a test if the effects occur within locations that are of lesser importance, or in $\boldsymbol{\theta}$ locations that individuals rarely populate (i.e., the extreme tails of the $\boldsymbol{\theta}$ distribution). Therefore, test analysts should consider the implications for differential functioning in their tests with respect to $\boldsymbol{\theta}$ rather than making a global — and often largely overly simplistic — binary decision as to whether DTF, DBF, or even DIF are present. A more useful approach is to supplement statistical tests with graphical representations and effect sizes to describe the conditional and marginal effects of DIF, DBF, and DTF in a metric or form that is meaningful to the test developer. This topic is discussed further in Chapter 4.

1.5 Statement of the Problem

While there has been a wealth of methods for detecting and quantifying DIF using techniques from CTT, IRT, and regression theory, there are considerably fewer methods for detecting DTF and DBF. The most prominent framework for detecting compensatory and non-compensatory DIF and compensatory DTF with IRT methods was proposed by Raju, van der Linden, and Fleer (1995), and later extended to include non-compensatory DBF by Oshima, Raju, Flowers, and Slinde (1998). An alternative framework based on methods from CTT was also presented for compensatory DIF, DBF, and DTF by Shealy and Stout (1993). Shealy and Stout's approach was later generalized to include non-compensatory DIF for 2PL models by Li and Stout (1996). The two frameworks approach the problem very differently, where the latter is derived from a CTT perspective using a regression adjustment technique to account for latent group differences and the former was based on a two-stage IRT approximation with which latent trait estimates are computed and compared across groups.

1.5.1 IRT Perspective of Differential Functioning: DFIT

In their first article regarding IRT methods for differential response functioning, Raju et al. (1995) proposed a statistical testing framework which they termed "differential functioning of items and tests" (DFIT) which appeared, on the surface at least, to be a potentially effective paradigm for detecting compensatory and non-compensatory DIF and DTF. Unfortunately, the authors' ini-

tial findings were that their statistics detected DIF and DTF too often when no such effects were present in the population (i.e., had liberal Type I error rates). With respect to the DIF statistics, an ad-hoc cutoff value was proposed to improve Type I error rates under the conditions that they studied. Follow-up work by Oshima et al. (1998) extended the DFIT framework to incorporate DBF, which unfortunately did not present the theoretical or empirical sampling behavior of the proposed non-compensatory DBF statistics. Instead, the authors only provided a demonstration of how researchers could use DFIT to diagnose non-compensatory DBF. The DFIT framework has since been amended to improve the ad-hoc cut-off values for the DIF statistics through computer intensive subroutines (Oshima, Raju, & Nanda, 2006). Around the same time of the original publication the methodology was also generalized to multidimensional IRT models (Oshima, Raju, & Flowers, 1997). More information regarding this framework will be presented in the Chapter 2.

Since the early work of Raju et al. (1995), a handful of articles purportedly extended the DFIT framework for investigating DTF and DBF. However, no published work — including the original methodological articles previously mentioned — has demonstrated the sampling behavior of the proposed DTF or DBF statistics from the DFIT framework. Following the statistical justification of the compensatory DTF statistics in Raju et al. (1995), later work by Oshima et al. (1997, 1998, 2006) rarely discussed the application of the DTF and DBF statistics. Several of the more recent articles, which purportedly extend the DFIT framework for DTF (e.g., Oshima et al., 2006), often focused exclusively on DIF with minimal discussion of DTF and DBF in the introduction and discussion sections; otherwise, these articles only present example analyses using their variants of

the DFIT framework. Hence, the majority of the authors who have advocated for the use of the DTF statistics from the DFIT framework have provided little to no information about the sampling properties of the respective DBF and DTF statistics. Hitherto, the sampling distribution, Type I error rates, and power to detect true DTF and DBF from the DFIT framework remain largely undocumented.

1.5.2 Non-Parametric Perspective of Differential Functioning: SIBTEST

The second framework intended for testing individual and composite differential response effects was proposed by Shealy and Stout (1993), who termed their approach the simultaneous item bias test (SIBTEST). Their proposed framework did not rely on methods or information from IRT models but instead used a regression adjustment technique, derived from methods in CTT (e.g., use of coefficient α; Guttman, 1945), to scale latent group differences after a suitable set of matched items was selected. Matched items serve the same purpose as anchor items in the IRT modeling approach in that the items are used to equate the groups, and are presumed not to contain DIF. The SIBTEST framework was investigated by Shealy and Stout for detecting DIF. The authors found that SIBTEST performed favorably in controlling Type I error rates under the conditions chosen, and demonstrated reasonable detection behavior when true DIF was present. However, much like the DFIT framework, very little research has focused on how SIBTEST behaves when investigating DBF and DTF, especially with respect to important issues such as: determining the optimal number of anchor items to select to ensure that the latent traits of the groups are properly

equated, whether the length of the test or bundle influences the detection rates, or the effects of including sub-optimal anchor items affect the detection of DBF and DTF.

Because the SIBTEST framework does not rely on fitting IRT models, and instead arguably adopts a non- or semi-parametric approach to DIF detection, the required computations are often reasonably efficient and easy to obtain without specialized software. However, as I describe in subsequent chapters, there are many less attractive issues which make the application of SIBTEST less than ideal for DTF, DBF, and potentially even DIF detection. Additionally, because SIBTEST is generally based on methods from CTT, the inherent benefits obtained from fitting IRT models are, by and large, not capitalized upon. These and other issues will be elaborated upon in the subsequent chapters after further background information on the technique is presented.

The SIBTEST and DFIT frameworks have existed for at least twenty years now, yet little is known about the performance of either framework with respect to DBF and DTF. To help determine how well these frameworks behave with respect to DBF and DTF, this dissertation will employ Monte Carlo simulations under a variety of conditions commonly encountered in real-world applications. Additionally, as will be more apparent after the more in-depth presentation of these frameworks in Chapter 2, both frameworks have their own apparent strengths and weaknesses which make their general application to empirical data difficult. For instance, the DFIT framework does not inherently account for sampling variability in the respective IRT models, and therefore lacks proper asymptotic properties for the proposed hypothesis tests. SIBTEST, on the other hand, has not been extended to include non-compensatory testing for polytomous items or

for non-compensatory DBF and DTF more generally. Ideally, these frameworks should be further developed to address these and other issues; otherwise, a new framework should be developed which is not hampered by these limitations. The differential response functioning (DRF) framework developed in the next chapter provides one such potential methodology.

1.6 Purpose

A more recent approach for detecting DTF using IRT methods was developed by Chalmers et al. (2016). In their article, a parametric sampling framework which used information from the parameter covariance matrix was presented for multiple-group IRT models that were fit using full-information maximum-likelihood estimation. Under the conditions investigated in their Monte Carlo simulation studies, the authors demonstrated consistent Type I error rates for their compensatory DTF statistic that were either at or slightly below the nominal α level. The authors presented several power tables to demonstrate how effective their proposed statistics were at capturing DTF effects, and provide multiple justifications as to why their sampling framework should behave better than the DFIT framework. In the work to be presented, further evaluations and extensions of the framework proposed by Chalmers et al. (2016) will be explored.

One focus of this dissertation will be comparing the methods proposed by Chalmers et al. (2016), Raju et al. (1995), and Shealy and Stout (1993) by simulating data from models with known IRT parameters and comparing results across distinct Monte Carlo simulation conditions. Conditions based on test length, sample sizes (equal and unequal between groups), latent variable

distributions (e.g., Gaussian distributions with different mean and variance-covariance combinations), number of anchor items, and various types of DIF effects will be investigated to determine the Type I error and power rates for all three frameworks. These simulation conditions will help demonstrate how effective SIBTEST and DFIT are when testing DIF, DBF, and DTF, and will also illustrate how the newly proposed framework in this dissertation behaves under the same conditions. This dissertation will also explore additional, more advanced methods which extend and improve upon the work of Chalmers et al. (2016), including the use of alternative test scoring functions, non-parametric estimation of parameter variability, DIF and DBF detection methods, hypothesis testing for non-compensatory effects, extensions for multidimensional test structures, and applications for equivalence testing paradigms. A unified approach to detecting and understanding differential response functioning will be developed and presented, and the collection of the measures proposed in this dissertation will ultimately be framed as a new statistical approach for detecting response bias termed the differential response functioning (DRF) framework.

Based on the parametric sampling methodology presented in Chalmers et al. (2016), it is anticipated that the statistics from the DRF framework will outperform the DFIT and SIBTEST frameworks both in terms of Type I error rates and power rates for DIF, DBF, and DTF. Overall, I argue that this newly proposed framework will provide a more useful set of statistical tools for studying DTF, DBF, and DIF in different testing applications. For example, the DRF framework methods can be used to build graphical realizations of differential functioning to visually depict response bias across θ while simultaneously accounting for sampling variability; this aspect of

the DRF framework is invaluable to test analysts and is a property not easily attainable with the DFIT or SIBTEST frameworks. Finally, the goal of this dissertation is to acquire a more intimate understanding of these newly proposed measures, ultimately helping test developers properly understand, interpret, and capitalize on these improved bias detection methods in their own psychometric research.

2 Differential Functioning Frameworks

This chapter provides the theory underlying the DFIT and SIBTEST frameworks and introduces a new set of methods for testing DIF, DBF, and DTF, called the DRF framework. For simplicity, the following exposition will focus only on the unidimensional 2PL model for dichotomous response data. For extensions regarding polytomous item response models for the DFIT, SIBTEST (excluding the non-compensatory SIBTEST), and DRF frameworks, see Flowers, Oshima, and Raju (1999), Chang, Mazzeo, and Roussos (1996), and Chalmers et al. (2016), respectively.

2.1 SIBTEST Framework

The SIBTEST framework was introduced by Shealy and Stout (1993) as a methodology to test for compensatory DIF, DBF, and DTF effects in unidimensional tests. SIBTEST requires the items in the dataset to be partitioned into a set of *matched items*, which are analogous to anchor items that are used to equate the groups, and *focal items* which are suspected to contain compensatory DIF effects. Note that not all non-focal items are required to be used in the matched set of items and instead can be completely removed from the analysis. However, excluding items from the matched

set has the effect of reducing the overall length of the test, and therefore will result in a loss of information and possibly less effective equating.

The general strategy for the SIBTEST family of statistics is to first select a set of matched items and use information from this set to adjust the differences on a separate composite measure (obtained from the sum score across items). Let Y represent the unweighted composite score computed using only the focal items, and let X represent the unweighted composite score using items only from the matched set. The SIBTEST procedure begins by computing the proportion of individuals $\hat{p}_c$ who obtained the composite score $X = c$ in the focal group. These proportions are then used as weights to be combined with the average composite score difference in Y, which is obtained at each unique value of the composite in X. In equation form,

$$\hat{\beta} = \sum_{c=0}^{C_X} \hat{p}_c(\bar{Y}_{Rc} - \bar{Y}_{Fc}), \tag{2.1}$$

where C_X is the maximum composite score in X. The sampling error of (2.1) is approximated by

$$\hat{\sigma}(\hat{\beta}) = \left(\sum_{c=0}^{C} \hat{p}_c^2 \left(\frac{\hat{\sigma}^2(Y_R|c)}{N_{Rc}} + \frac{\hat{\sigma}^2(Y_F|c)}{N_{Fc}} \right) \right)^{1/2}. \tag{2.2}$$

The $\hat{\sigma}^2(Y_R|c)$ and $\hat{\sigma}^2(Y_F|c)$ terms in (2.2) are the sample variances of the collection of Y_{Rc} and Y_{Fc}, respectively, while the N_{Rc} and N_{Fc} terms represent the frequency of the selected composite X when $X = c$ within the reference and focal groups, respectively. Following the computation of these two values, the test statistic $\hat{B}$ is formed from the ratio

$$\hat{B} = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})}. \tag{2.3}$$

Under the null hypothesis $H_0 : B = 0$, Equation (2.3) follows a standard normal distribution if and only if the focal and reference groups have similar population distributions of the latent trait scores (i.e., equal hyper-parameters).

Shealy and Stout (1993) recognized that (2.3) is not a useful statistical test if the latent trait distributions for each group are unequal. Therefore, the authors' major contribution was to form a true-score regression technique to lessen the effect of the latent trait distributional differences. Shealy and Stout suggested using a Taylor series adjustment for the $\bar{Y}$ composite terms by regressing the values towards the CTT true-score information determined from the matched item-set[3]. After the KR-20 reliability formula (Kuder & Richardson, 1937) is used to obtain the regression weights in the focal and reference group these weights are then used to compute adjusted composite scores $\bar{Y}^*_{Rc}$ and $\bar{Y}^*_{Fc}$. The adjusted composite scores are then substituted into (2.1) to form the improved weighted difference estimate

$$\hat{\beta}_{uni} = \sum_{c=0}^{C} \hat{p}_c(\bar{Y}^*_{Rc} - \bar{Y}^*_{Fc}), \tag{2.4}$$

which is in turn used to evaluate $\hat{B}$ in (2.3) instead of (2.1); note, however, that Equation 2.2 remains unchanged. For more specific details regarding the implementation of SIBTEST readers should refer to Shealy and Stout (1993).

The SIBTEST procedure has several attractive properties. To start, the statistic is easy to implement and efficient to compute. Because the procedure is based on methods from CTT it does

[3]In situations where a Taylor series approximation is known to be poor, such as in 3PL models or when the item slope parameters are large, see the piecewise regression method proposed by Jiang and Stout (1998).

not rely on estimating IRT models; therefore, it does not have issues such as non-convergence in smaller sample sizes or longer computations in larger samples and tests. Next, the statistic can be organized to test compensatory DIF, DBF, and DTF by selecting different sets of focal items (i.e., compensatory DIF is tested when only one item is included in the focal set, DBF is tested when some but not all of the non-matched items are included in the focal set, and DTF is tested when all non-matched items are included in the focal set). Third, although Shealy and Stout (1993) presented SIBTEST for dichotomous items the procedure naturally extends to ordered polytomous items by replacing the KR-20 reliability formula with coefficient α (Chang et al., 1996; Guttman, 1945). Finally, given the recommendations proposed by the authors when forming the p_c values, the inferences become invariant to the selection of the focal and reference group; in other words, switching the reference and focal groups will only flip the sign of $\hat{B}$. Invariance with respect to the selection of the focal group is important because the test statistics will not be affected by different focal and reference group sizes or the decision regarding which group should be considered the reference group.

Further generalizations of SIBTEST have been proposed but have seen little use due to their limited scope and ad-hoc nature. For instance, Stout, Li, Nandakumar, and Bolt (1997) generalized SIBTEST to accommodate tests which have a specific type of two-dimensional latent structure. This extension requires two distinct sets of unidimensional items to be used as the matching criteria for identifying each dimension separately so that the remaining items with cross-loadings can be properly equated. After organizing these two matched sets, a two-dimensional Taylor se-

ries approximation is used to correct for latent differences between the conditioned composite terms. Following these initial adjustments, the bivariate SIBTEST procedure followed the same steps described above for unidimensional models. However, generalizations to other structures, such as three or more dimensions, which are commonly found in tests with second-order or bi-factor structures (see Yung, Thissen, & McLeod, 1999), do not appear feasible under the SIBTEST framework. Additionally, the requirement that the matched items are strictly unidimensional may be viewed as a strong limitation in empirical studies. Further limitations regarding the SIBTEST framework in general will be discussed in the subsequent section.

An alternative approach for quantifying DIF with the SIBTEST method was proposed by Li and Stout (1996), who presented a modified version of SIBTEST which could be used for detecting simple non-compensatory DIF effects for dichotomous data. The authors argued that when DIF is present in response models such as the 2PL model the expected probability functions may cross, thereby decreasing the magnitude of SIBTEST (compare the left and right graphics in Figure 1.2). Hence, the compensatory nature of the original SIBTEST procedure would result in a reduction in power to detect these effects because the differences above and below the crossing location will cancel out. Li and Stout therefore recommended using a crossed-SIBTEST (CSIBTEST) variant which combines the information above and below the response curve crossing location with

$$\hat{\beta}_{cross} = \sum_{c=0}^{k_c-1} \hat{p}_c(\bar{Y}^*_{Rc} - \bar{Y}^*_{Fc}) + \sum_{c=k_c}^{C_X} \hat{p}_c(\bar{Y}^*_{Fc} - \bar{Y}^*_{Rc}). \tag{2.5}$$

In this equation, k_c represents the score on X where the response curves are believed to cross. In situations with no crossing in the item, $\hat{\beta}_{cross}$ reduces to the original SIBTEST statistic. Finally, the

$\hat{B}_{cross}$ ratio is obtained by dividing $\hat{\beta}_{cross}$ by the unadjusted $\hat{\sigma}(\hat{\beta})$ term from the original SIBTEST procedure.

2.1.1 Limitations

Unfortunately, there is a number of limitations with the SIBTEST framework when testing for differential response functioning. Beginning with CSIBTEST, there are two serious limitations in the computations which Li and Stout (1996) made note of and attempted to address. The first issue is that (2.5) requires the specification of a crossing location prior to beginning the computations. By first assuming the crossing location is linear with respect to the sub-scores $X = c$, the authors suggested using a weighted least squares regression approach to approximate an intercept location and rounding the $k_c = -\beta_0/\beta_1$ intercept estimate to the nearest integer value. Li and Stout stated that, in their experience, this approach worked well as long as the selected frequency weights contained more than 1% of the total sample size. The second limitation to CSIBTEST is that the B_{cross} test does not have a known sampling distribution; therefore, the asymptotic properties relevant to SIBTEST are not applicable. To circumvent this issue, the authors proposed a signed permutation approach for stochastically building a suitable sampling distribution to obtain empirical p-value estimates. Combining both these techniques appeared to be an effective strategy for testing non-compensatory DIF, and Li and Stout report reasonable Type I error rates which were comparable to the original SIBTEST procedure, as well as improved power estimates under the DIF conditions studied. However, generalizing this procedure to DBF and DTF, as well as to polytomous items

and multidimensional tests, appears problematic due to the potential for multiple crossing locations. This concern may be one reason that additional non-compensatory SIBTEST developments have not appeared in the literature.

In addition to the aforementioned limitations of CSIBTEST, several other practical problems often arise in empirical applications which the SIBTEST family of statistics is not well equipped to manage. One such issue is the selection of which items to include in the matched set. Ideally, the matched set should have the same properties as the previously described anchor items when equating groups in IRT applications; therefore, these matched items ought to contain no DIF. Including items with DIF effects as anchor items naturally contaminates the linking process, which is why this approach is generally not recommended (Millsap, 2011). Unfortunately, the general consensus when using SIBTEST for DIF testing is to include *all* non-focal items as anchors, regardless of whether the items contain DIF[4]. Fortunately, the contamination issue can be remedied by including only a smaller number of matched items, where non-focal items which potentially contain DIF are forced to have no influence on the current focal items under investigation; however, doing so comes at the cost of discarding auxiliary information, thereby reducing the total length of the test. The performance of SIBTEST when only a few matched items are selected as suitable anchors has not been thoroughly investigated in the literature. This issue is extremely important for DBF and DTF in particular because the size of the focal bundle relative to the number of available anchor

[4]Shealy and Stout (1993) report that the Type I error rates when detecting DIF were largely unaffected in the presence of contaminated anchors. This result seems implausible though, and therefore is investigated in the subsequent Monte Carlo simulations.

items may negatively affect the behavior of the test statistics. Therefore, one important factor investigated in the Monte Carlo simulations presented in the next chapter is the number of anchor or matched items used when testing for DIF, DBF, and DTF.

Another potentially problematic issue with the SIBTEST framework is that it cannot inherently handle missing response data. If item responses are missing in either the focal or matched set then the entire response pattern associated with the missing items must be removed before applying the procedure[5]. This problem is viewed as a major limitation for test designs where missing responses are due to design effects, often caused by administering multiple testing forms; planned missingness is one area where IRT-based methods are superior because of their full-information ML nature (Bock, Gibbons, & Muraki, 1988).

Finally, more specific limitations of the SIBTEST framework are that: 1) the statistics are typically limited to ordinal data for the match set; hence, unordered models (e.g., Thissen, Cai, & Bock, 2010), ideal point models (e.g., Maydeu-Olivares, Hernández, & McDonald, 2006), non-monotonic response functions (e.g., Bock & Aitkin, 1981), and so on cannot be included in the matched item set; 2) the inclusion of a lower-bound parameter to account for guessing is problematic because negative reliability estimates can arise (cf. the ad-hoc adjustment in step 5 of Shealy & Stout, 1993, p. 192); 3) the effect size measure (Equation 2.4) is somewhat difficult to interpret due to the regression adjustment effects, and established DIF effect size cutoffs only exist because of their approximate relationship to the popular Mantel-Haenszel log-odds ratio procedure (Holland

[5]Alternatively, multiple imputation methodology could be implemented to amend the problem; however, this procedure may introduce other challenges and does not appear to be a feasible strategy for the CSIBTEST procedure.

& Thayer, 1988, more on this topic is in Chapter 4); 4) the procedures provide little to no graphical means of plotting the data to depict the magnitude or location of the differential functioning (cf. the methods in TestGraph by Ramsay, 2000; however, these plots do not include the regression correction for latent distribution differences, and are therefore not directly related to SIBTEST); and 5) the statistics are exclusively marginal estimates of the overall bias present in an item or bundle in that they offer little insight to conditional differential effects given $\boldsymbol{\theta}$ (i.e., they provide little insight as to whether the response bias is larger for examinees with lower or higher latent trait values). SIBTEST has been modified to asses the fifth limitation for DIF; however, similar to CSIBTEST, this modification requires a substantial change to the original SIBTEST methodology and does not appear to behave optimally (see Douglas, Stout, & DiBello, 1996).

2.2 DFIT Framework

Instead of focusing on methods derived from CTT, Raju et al. (1995) and Oshima et al. (1997) approached the detection of response bias using a two-step approximation technique. Their differential functioning of items and test procedure, or DFIT, begins by estimating two separate single-group IRT models which are subsequently equated using some parameter linking method. Following this initial setup, predicted values of the latent traits ($\hat{\boldsymbol{\theta}}$) within the focal group are computed by treating the obtained item parameter estimates as stand-in values for the true population parameters. Predicted values for $\hat{\boldsymbol{\theta}}$ (e.g., expected a posteriori, maximum a posteriori, maximum likelihood, etc) are then computed using the information from the focal ($\hat{\boldsymbol{\theta}}_{iF}$) and reference groups

($\hat{\boldsymbol{\theta}}_{iR}$) separately: once using the linked parameter estimates in the focal group with the response data from the focal group, and again using the parameters from the reference group given the response data from the focal group. This process ensures that response patterns are exactly equal across groups when computing the $\hat{\boldsymbol{\theta}}$ values; therefore, only information that would indicate DIF or DTF can be attributed to differences in the item parameter estimates (i.e., not due to different response patterns).

After $\hat{\boldsymbol{\theta}}_{iR}$ and $\hat{\boldsymbol{\theta}}_{iF}$ values are formed for N_F response patterns from the focal group, Raju et al. (1995) suggested using these predictions to form compensatory and non-compensatory DIF, as well as a compensatory DTF, statistics. Their proposed non-compensatory DIF statistic for the jth item was

$$\widehat{NCDIF} = \frac{1}{N_F}\sum_{i=1}^{N_F}[P(y=k|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{iF},\boldsymbol{\psi}=\hat{\boldsymbol{\Psi}}_{jF}) - P(y=k|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{iR},\boldsymbol{\psi}=\hat{\boldsymbol{\Psi}}_{jR})]^2, \tag{2.6}$$

while their compensatory DIF and DTF statistics were

$$\widehat{CDIF} = \left(\frac{1}{N_F}\sum_{i=1}^{N_F} P(y=k|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{iF},\boldsymbol{\psi}=\hat{\boldsymbol{\Psi}}_{jF}) - P(y=k|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{iR},\boldsymbol{\psi}=\hat{\boldsymbol{\Psi}}_{jR})\right)^2, \tag{2.7}$$

and

$$\widehat{DTF} = \left(\frac{1}{N_F}\sum_{i=1}^{N_F}[T(y=C|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{iF},\boldsymbol{\psi}=\hat{\boldsymbol{\Psi}}_{jF}) - T(y=C|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{iR},\boldsymbol{\psi}=\hat{\boldsymbol{\Psi}}\hat{j}_R)]\right)^2, \tag{2.8}$$

respectively. For hypothesis test-based inferences regarding DIF and DTF the authors proposed different variations of χ^2 and t distributions for their proposed statistics; however, they recommended using the χ^2 distribution over the t distribution because the statistical conclusions are

essentially equivalent in larger samples. The χ^2 tests are built as follows. Let $\hat{D}_i = T(y = C|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{iF}, \boldsymbol{\psi} = \hat{\boldsymbol{\Psi}}_{jF}) - T(y = C|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{iR}, \boldsymbol{\psi} = \hat{\boldsymbol{\Psi}}_{jR})$ and $\hat{d}_i = P(y = k|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{iF}, \boldsymbol{\psi} = \hat{\boldsymbol{\Psi}}_{jF}) - P(y = k|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{iR}, \boldsymbol{\psi} = \hat{\boldsymbol{\Psi}}_{jR})$ represent the difference between the conditional test and item response functions, respectively. Using these difference estimates Raju et al. proposed that a test for $H_0 : NCDIF = 0$ is

$$\chi^2_{NCDIF} = \frac{\sum_{i=1}^{N_F} \hat{d}_i^2}{\hat{\sigma}^2(\hat{d})}, \tag{2.9}$$

with degrees of freedom equal to N_F. Analogously, the authors suggested that a test for $H_0 : DTF = 0$ is

$$\chi^2_{DTF} = \frac{\sum_{i=1}^{N_F} \hat{D}_i^2}{\hat{\sigma}^2(\hat{D})}, \tag{2.10}$$

again with degrees of freedom equal to N_F. The authors did not propose a statistical test for $H_0 : CDIF = 0$, and instead use the statistic only as a post-hoc diagnostic tool when significant DTF is detected[6].

2.2.1 Limitations

After simulating data with no population DIF effects, Raju et al. (1995) discovered that their test statistics reported positive signs for DIF and DTF far too often. For the $\widehat{NCDIF}$ statistic, Raju et al. suggested using the cut-off value of $\widehat{NCDIF} = .006$ as an approximate significance flag when the nominal detection rate was set to $\alpha = .01$. This criterion appeared to behave relatively well

[6]Using the same line of reasoning that the authors followed, it appears that $CDIF$ should have the same sampling distribution as $NCDIF$.

under the conditions they studied (though the cuttoff did not generalize well outside the conditions studied; see Oshima et al., 2006). Unfortunately, however, the authors offered no such cutoff for the $\widehat{CDIF}$ or $\widehat{DTF}$ statistics. Therefore, it is unclear how practitioners should use the proposed compensatory statistics in their own work because the behavior of these statistics is unclear.

Generalizations of the DFIT framework have suffered similar limitations in terms of expected sampling behavior. For instance, although the DFIT framework was originally proposed only for non-compensatory DIF and compensatory DTF detection a simple generalization of $\widehat{CDIF}$ was presented by Oshima et al. (1998) for compensatory and non-compensatory DBF, as well as for non-compensatory DTF. The procedure they presented was to replace the test scoring function in (2.8) with a response function which contained fewer than J items. Using a smaller number of items led to the bundled $\widehat{CDBF}$ statistic, and for non-compensatory DBF and DTF, the authors simply summed across more than one item in (2.6) to form

$$\widehat{NCDTF} = \frac{1}{N_F}\sum_{i=1}^{N_F}\left(\sum_{j=1}^{J}[P(y=k|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{iF},\boldsymbol{\psi}=\hat{\boldsymbol{\Psi}}_{jF}) - P(y=k|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{iR},\boldsymbol{\psi}=\hat{\boldsymbol{\Psi}}_{jR})]^2\right), \quad (2.11)$$

where the J number of items included was defined based on whether DTF or DBF was being investigated. Unfortunately, the definitions presented were only conceptual. After presenting these extensions the authors offered no information regarding the statistical sampling characteristics or interpretation of effect sizes, and only presented an empirical case study demonstrating the use of these heuristic methods. From the definitions, however, the $\widehat{DBF}$ statistic would have the same sampling properties as the $\widehat{DTF}$ statistic; therefore, the proposed theoretical sampling distribution for $\widehat{DTF}$ was adapted for $\widehat{DBF}$ in the Monte Carlo simulation study contained in the following

chapter, as well as for the $\widehat{CDIF}$ statistic.

In later published work, Oshima et al. (2006) attempted to address the problematic Type I error behavior of the $\widehat{NCDIF}$ statistic by adopting a simulation-based cutoff procedure. Using a parametric bootstrap-type approach, the authors demonstrated that their procedure worked well to determine better cutoff values for the conditions under study (at the cost of being computationally intensive). However, the authors did not propose a similar solution for the $\widehat{CDIF}$ or $\widehat{DTF}$ statistics.

Another potential limitation and complication to the approach described by Raju et al. occurs when items contain DIF effects but a complete-item linking method is adopted (rather than linking by a set of anchor items). To my knowledge, all articles which have attempted to extend the DFIT framework have used the complete-item linking method, thereby creating contamination in the linking process (Millsap, 2011). This renders the results of these published works difficult to interpret, and largely inconclusive with regards to the expected behavior of these extensions. Contaminated linking can, however, be avoided if models are estimated using a multiple group estimation method whereby only a pre-determined set of anchor items are modeled. This potential solution to the linking issue will be explored in the subsequent Monte Carlo simulation studies in Chapter 3.

In addition to the aforementioned technical limitations of the DFIT framework, Chalmers et al. (2016) have argued that there are numerous fundamental limitations with the DFIT framework for detecting DIF and DTF. The authors argue that the primary concern with the DFIT framework is that the sampling variability of the item parameter estimates is not taken into account, and instead

point estimates for the $\boldsymbol{\theta}$ values are used as a proxy for obtaining the person and item variability. This assumption leads to the well known problem of shrunken estimates that are influenced by numerous factors such as sample size, test length, type of prediction method (e.g., expected a posteriori, maximum a posteriori, maximum-likelihood, etc, see see Chalmers, 2016a, for examples), linking methods, unequal group sizes, selection of focal group, and so on (Mislevy, 1991), while also ignoring the inherent reliability of the item parameter estimates themselves (see Chalmers & Ng, in press).

Additionally, due to the use of response data from only the focal group (i.e., ignoring the response patterns from the reference group entirely) a loss of information from the sample data necessarily occurs. Therefore, despite the inflated Type I error rates that have been previously reported — and even if the Type I error behavior were somehow amended — the detection rates for these statistics will still be less powerful than other frameworks which use the sampling information from all groups.

Finally, similar to the SIBTEST framework, the DFIT framework offers no generalizable interpretation of effect sizes for differential effects, does not handle missing data efficiently (i.e., the point estimates do not consider the decrease in measurement precision for the $\hat{\boldsymbol{\theta}}$ predictions or in the item parameter estimates), and provides no additional means of graphically depicting the detected differential effects. Many of the practical implications of these limitations are empirically examined in Chapter 3, however it is clear that there are numerous fundamental issues present in the DFIT framework; hence, readers should be skeptical about the framework's overall usefulness

in detecting and quantifying response bias.

2.3 Differential Response Functioning Framework

Prior to 2016, the SIBTEST and DFIT frameworks were the only statistical methods designed to detect DTF effects; however, these methods were also capable of estimating DIF and DBF. As outlined above, each of the previous frameworks has unique strengths and limitations. While the SIBTEST framework takes into account sampling variability, as demonstrated by its ability to control Type I error rates across a number of simulated conditions, it is not based on a parametric model fitting approach; hence, it cannot deal with real-world issues such as the presence of missing data, items with unordered categories, non-monotonic response functions, and so on. On the other hand, while the DFIT framework focuses exclusively on IRT methods it does not account for sampling variability due to its ad-hoc two-stage approximation approach. This approach leaves the DFIT framework with its own unique statistical limitations which currently appear difficult to overcome.

Instead of attempting to amend the two aforementioned frameworks, Chalmers et al. (2016) proposed an alternative statistical approach for detecting DTF that is rooted in IRT methods and maximum-likelihood estimation theory (e.g., see Fisher, 1925). Chalmers et al. presented a parametric sampling method to the capture variability in IRT composite functions for measuring DTF by using the estimated sampling information present in the IRT parameters. The method draws from the graphical concept of Thissen and Wainer (1990), who demonstrated how to generate

confidence envelopes in single-group IRT models for individual item response functions.

Thissen and Wainer (1990) proposed the following scheme for generating confidence envelopes for non-linear response functions, which also may be used to generate parametric confidence intervals given the same sampled parameter set at some desired $\boldsymbol{\theta}$ level. The steps required for the parameter sampling method are:

- Estimate the IRT model using maximum-likelihood. Following convergence, compute an estimate of the parameter covariance matrix $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$.

- Using $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$, create a subset of the matrix containing only the estimated sampling variability of a single isolated item; call this sub-matrix $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\psi}}_j|\mathbf{Y})$. For example, if the item selected follows a 2PL model, then $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\psi}}_j|\mathbf{Y})$ is a 2×2 matrix consisting of the variance of the slope and intercept parameters with the covariance between these parameters in the off-diagonal.

- Generate a set of plausible population realizations (i.e., samples) for the desired $\hat{\boldsymbol{\psi}}_j$ parameters by randomly sampling from the distribution $\boldsymbol{\psi}_j^* \sim \mathcal{N}(\hat{\boldsymbol{\psi}}_j, \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\psi}}_j|\mathbf{Y}))$. Here, $\boldsymbol{\psi}_j^*$ is a single plausible set of population parameters $\boldsymbol{\psi}_j$ drawn stochastically around the ML estimates $\hat{\boldsymbol{\psi}}_j$. Repeat this process until M plausible sets have been obtained.

- Using the M plausible sets of values, generate non-linear confidence envelopes or intervals by substituting the $\boldsymbol{\psi}_j^*$ values into the desired response functions using a predetermined grid of $\boldsymbol{\theta}$ values. After these functions have been evaluated, the complete set of sampled parameters ($\boldsymbol{\Psi}^*$) that fall within the joint $1 - \alpha$ confidence envelope can be selected and used for

understanding the joint parameter distribution for the selected item. Alternatively, the set of plausible samples can be rank ordered to locate suitable $\alpha/2$ and $1-\alpha/2$ confidence interval limits given each element in the $\boldsymbol{\theta}$ grid (cf. Efron & Tibshirani, 1998).

The above scheme follows from the fact that, for unbounded and continuous parameter spaces, the sampling distribution of the total set of ML parameter estimates is multivariate normal in large samples (Fisher, 1925). Therefore, subsets of the complete $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix will themselves be multivariate normal (Johnson & Wichern, 2007), and the implied variability of each item can be generated within the associated parameter space and subsequently mapped onto probability space. Finally, the parametric sampling scheme is generally not too computationally demanding and, if desired, each sampled set may be distributed across different computing cores because the draws are completely independent.

The inspiration behind the parametric sampling method is simple and straightforward: to generate variability in the non-linear response functions, one need only sample values from a well behaved parameter space and use these sampled values to construct a set of plausible values by plugging the parameter sets into the respective non-linear response functions. What is meant by well behaved parameter space is that $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ represents a reasonable approximation of the sampling variability, and therefore should only be used for unbounded parameters. Bounded parameters, such as the lower bound term in the 3PL model (see Lord & Novick, 1968), should be re-parameterized so that $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ will better approximate the quadratic curvature of the log-likelihood

function[7].

The parametric sampling approach described above can be useful for augmenting statistical interpretations when a large amount of sampling variability is present in the estimates of single-group models (Yang, Hansen, & Cai, 2012). However, there are other uses of the plausible parameter sets outside of plotting functions and generating sampling uncertainty in single-group IRT models. When random draws are obtained from the complete parameter covariance matrix then these respective values represent the sampling variability of the entire fitted model; this property holds even in multiple-group IRT models.

As demonstrated in Chalmers et al. (2016), using the complete $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix to create sampled parameter sets captures a large degree of the joint parameter variability in the estimated IRT models. Subsequently, the full set of sampled values can be used for hypothesis testing with composite functions. For instance, Chalmers et al. proposed two area-based statistics for detecting DTF and referred to these as 'signed' and 'unsigned' differential test functioning measures ($sDTF$ and $uDTF$, respectively). These marginal DTF measures were defined as

$$sDTF = \int \left(T(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) - T(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F)\right) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.12}$$

and

$$uDTF = \int |T(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) - T(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F)| g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.13}$$

respectively, to capture compensatory and non-compensatory response behavior by integrating

[7]In the case of the 3PL model, the lower-bound parameter can be transformed using $\gamma' = \log\left(\frac{\gamma}{1-\gamma}\right)$, where γ' is estimated in place of γ. Compared to γ which is constrained to fall between 0 and 1, $-\infty \leq \gamma' \leq \infty$.

over a range of $\boldsymbol{\theta}$ values, where $\int g(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$ and all values of $g(\boldsymbol{\theta})$ are constant (i.e., a uniform distribution). Note that these definitions are essentially identical to Equations 1.9 and 1.11, respectively, if the item bundle contains all items in the test. The detection of DTF at individual $\boldsymbol{\theta}$ locations was also developed by computing a measure called $sDTF_{\boldsymbol{\theta}}$ (c.f., Equation 1.8), which is evaluated at isolated $\boldsymbol{\theta}$ values instead of integrating across a range of $\boldsymbol{\theta}$. Computing $sDTF_{\boldsymbol{\theta}}$ across a set of independent $\boldsymbol{\theta}$ values provides conditional estimates of DTF, as well as their associated variability when combined with the parametric sampling method, which is useful for visually approximating where DTF is most prominent in the test. Chalmers et al. (2016) further argued that because $sDTF$ and $uDTF$ represent more realistic and meaningful population definitions of DTF than are available in DFIT, and because the measures contain no reference to the observed data, their measures should be adopted when evaluating DTF in IRT applications.

Utilize the above measures effectively a small set of anchor items must be modeled so that the required latent variable hyper-parameters can be freely estimated in a multiple-group IRT model. Properly linking the groups allows a suitable estimate of $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ to be obtained for the complete set of parameter estimates. After the multiple group model has been estimated by ML methods the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix and vector of parameter estimates can then be used in the parametric sampling scheme to form point-wise confidence intervals and hypothesis tests for the DTF statistics.

The sampling of M sets of ψ^* can be organized to form a $M \times P$ matrix of plausible parameter values, $\boldsymbol{\Psi}^*$. The full set of sampled parameters can then be used to form a null hypothesis test to evaluate $H_0 : sDTF = 0$. More specifically, to perform this test one need only apply a slight

variation of the aggregation strategy proposed by Rubin (1987). After obtaining M independent ψ^* sets the collected values can be used to evaluate the equation for $sDTF$, the results of which can be stored in a vector $\mathbf{m}_{sDTF}$. Given the values in $\mathbf{m}_{sDTF}$, the sampling variability of $\widehat{sDTF}$ is obtained by computing the sample variance estimate $\hat{\sigma}^2(\mathbf{m}_{sDTF})$. This variance estimate serves as the marginal sampling variance for $\widehat{sDTF}$, and the value $\sqrt{\hat{\sigma}^2(\mathbf{m}_{sDTF})}$ may be used as the Monte Carlo approximated standard error of the statistic. Next, the ratio

$$z = \frac{\widehat{sDTF}}{\sqrt{\hat{\sigma}^2(\mathbf{m}_{sDTF})}} \tag{2.14}$$

is formed, where z is distributed $\mathcal{N}(0, 1)$ when the number of parametric samples is large (say, greater than 200). When the number of samples is small, Rubin (1987) recommended using a t distribution with $df = M - 1$. Chalmers et al. (2016) investigated the performance of this statistic under various simulated conditions and found Type I error rates which were consistently close to the nominal α level in larger sample sizes.

2.3.1 Extension of the Parametric Sampling Framework: Hypothesis Test for Non-Compensatory Response Functions

Although the hypothesis test in (2.14) can be used to evaluate the $sDTF$ measure, a hypothesis test for $uDTF$ should not be evaluated using the same approach because a closed-form version of the sampling distribution does not exist. However, in that (2.14) can be viewed as a z statistic, which is formally equivalent to a $\sqrt{\chi^2}$ with $df = 1$, we may generalize the hypothesis testing methodology for unsigned DTF by adopting the alternative definition of non-compensatory DTF from

Equation 1.10. This definition is applicable within the parametric sampling framework because the χ^2 distribution is a distribution for squared random deviates. Hence, the quantity defined in the following equation also follows an approximate χ^2 distribution in large samples. Let the following measure represent the average deviation between two response functions:

$$dDTF = \left(\int \left(T(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) - T(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F) \right)^2 g(\boldsymbol{\theta}) d\boldsymbol{\theta} \right)^{1/2}. \tag{2.15}$$

Assuming an analogous process to what was used to construct the sampled variability in (2.14), a statistical test of $H_0 : \ dDTF = 0$ can be expressed as

$$X^2 = \frac{\widehat{dDTF}^2}{\hat{\sigma}^2(\mathbf{m}_{dDTF})}. \tag{2.16}$$

As was the case for the distribution of $sDTF$, after a large number of samples have been obtained Equation 2.16 will have an approximate χ^2 distribution. Preliminary simulation work studied across a wide variety of conditions (including, but not limited to, sample size, test length, different latent trait distributions, number of anchors, equal/unequal group sizes, number of response categories, and IRT models) suggested that X^2 is approximately χ^2 distributed with $df = 2$ instead of the theoretical $df = 1$. Therefore, in all subsequent simulation work $df = 2$ was used to construct p-values and confidence intervals for this measure.

Given this new hypothesis test both signed and unsigned DTF hypothesis tests and confidence intervals can now be constructed within the parametric sampling framework. This statistical extension is important because neither the SIBTEST nor the DFIT frameworks has presented compelling statistics for detecting non-compensatory effects in bundles or tests. Hence, a formal hypothesis

test for non-compensatory DTF is unique across all previously presented frameworks. Because the Monte Carlo simulation study presented by Chalmers et al. (2016) did not evaluate this statistic it will therefore be investigated in the Monte Carlo simulation study in the next chapter.

2.3.2 Extension of the Parametric Sampling Framework for DIF and DBF Testing

Testing DIF and DBF is another important area where SIBTEST and DFIT have been developed but the previously discussed parametric sampling framework has not. Extending the parametric sampling framework is straightforward because DIF and DTF can be understood as special cases of DBF. Hence, (2.12), (2.13), and (2.15) can be generalized by modifying the indexing function $I(j)$, as demonstrated in Chapter 1. For instance, to focus only on a bundle of items (where $1 < B < J$), the definition for signed DBF is

$$sDBF = \int \left(T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) - T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F)\right) g(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{2.17}$$

In the same spirit, the definition of signed differential item functioning is

$$sDIF = \int \left(S(c|\boldsymbol{\theta}, \boldsymbol{\psi}_R) - S(c|\boldsymbol{\theta}, \boldsymbol{\psi}_F)\right) g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \left(T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) - T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F)\right) g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.18}$$

where the bundle size is $B = 1$. The same reasoning can be applied to the remaining unsigned DTF statistics to form $uDBF$, $dDBF$, $uDIF$, and $dDIF$. Furthermore, the sampling distributions for the differential item and bundle statistics are the same as for the DTF statistics: for signed tests, the statistics follow an approximate z distribution, and for the squared difference measures the statistics follow an approximate χ^2 distribution. Hence, compensatory and non-compensatory DIF, DBF, and

DTF can be well defined within this parametric sampling framework and have similar asymptotic, intepretational, conditional, and graphical properties. Henceforth, the set of statistics that has been described above will collectively be referred to as statistics from the differential response functioning framework, or DRF, and the family of signed, unsigned, and deviation measures can be referenced using the letter R in place of the I, B, and T (e.g., $sDRF$ refers to the measures $sDIF$, $sDBF$, and $sDTF$).

The extensions presented above for DIF and DBF appear straightforward and simple; however, some added benefits arise with respect to the parametric sampling method for DIF. For example, when drawing the parameters to evaluate the DIF statistics, it is easy to evaluate all items simultaneously using the same sampled parameter set; hence, new draws which focus on independent items are not required. This property holds because $\boldsymbol{\Psi}^*$ is a joint realization from the complete parameter distribution space and therefore represents a suitable set for each item independently. This is a very attractive feature because, unlike DIF tests which require independent evaluations (such as the likelihood-ratio approach to testing for DIF, where models must be re-estimated to form nested counterparts), only one model with an estimated parameter covariance matrix need be computed, and only one complete set of $\boldsymbol{\Psi}^*$ need be obtained. This approach has an additional benefit over the CSIBTEST procedure because it need only be computed once, while the required re-sampling technique for CSIBTEST must be applied independently for each item, thereby increasing computational times proportional to the number of items under investigation.

2.3.3 Commonalities with the Wald Test for DIF

The parametric sampling approach in the DRF framework shares many commonalities with the setup for the popular Wald test (Wald, 1943) for DIF in that both approaches can be performed after obtaining suitable $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ terms following ML estimation of multiple-group IRT models. The Wald test can be used for detecting DIF effects by forming suitable contrast matrices to construct linear hypothesis tests with one or more degrees of freedom. Depending on which parameters are tested across groups, the Wald test may be used to detect either uniform DIF (e.g., when only intercepts are tested) or non-uniform DIF (e.g., when slopes are tested, or slopes and intercepts are tested simultaneously). The Wald test statistic has the form

$$W^2 = \mathbf{L}(\hat{\boldsymbol{\Psi}} - \mathbf{c}) \left(\mathbf{L}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})\mathbf{L}' \right)^{-1} (\hat{\boldsymbol{\Psi}} - \mathbf{c})'\mathbf{L}', \tag{2.19}$$

where $\mathbf{L}$ is a $Q \times P$ contrast matrix of design-based constants used to index the suitable parameter estimates from the $P \times 1$ matrix of parameter estimates, and $\mathbf{c}$ is a vector of fixed parameter values defined under H_0 to be tested against (often consisting only of 0s). Under mild regularity conditions, W^2 is approximately χ^2 distributed in large samples with Q degrees of freedom.

When using the Wald test for DIF, the matrix $\mathbf{L}$ need only consist of the values 1 and -1 to test whether the respective slopes and intercepts are equal across groups. For example, say that a researcher is interested in testing non-uniform DIF for a three item multiple-group IRT model consisting of only 2PL models for each item. After estimating the model by treating the first item as an anchor item, the vector of parameter estimates is

$$\hat{\boldsymbol{\Psi}} = (\alpha_1, \beta_1, \alpha_{R2}, \beta_{R2}, \alpha_{F2}, \beta_{F2}, \alpha_{R3}, \beta_{R3}, \alpha_{F3}, \beta_{F3}, \mu_F, \sigma_F^2).$$

The elements α_1 and β_1 are the slope and intercept for the first item which are constrained to be equal across groups, the next eight elements are the parameter estimates for each item indexed by groups (reference or focal), and finally the last two elements are the estimated hyper-parameters for the focal group (the corresponding hyper-parameters in the reference group are fixed to $\mu_R = 0$ and $\sigma_R^2 = 1$). To test whether the second item contains non-uniform DIF, the $\mathbf{L}$ matrix is constructed as

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

which is substituted into Equation 2.19, along with $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$, to form a test statistic with $df = 2$. A non-significant Wald tests implies that there is insufficient evidence to conclude that the parameter sets are unequal across groups; hence, there is no evidence that DIF exists in the respective item.

The parametric sampling approach for the DRF statistics and the Wald test both require that the groups are properly equated and that a suitable $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix has been obtained. In fact, the initial conditions and prior computations for both tests are identical. In situations where it is possible to conduct a Wald test for DIF, it is also appropriate to use the parametric sampling approach with the DRF statistics. The converse, however, is not true because the Wald test requires the parameter sets to be equivalent in both groups so that differences in parameter estimates can be meaningfully tested. For instance, it does not makes sense to compare a 2PL model in one group with an IRT

model fit by a polynomial model in a different group because it is not clear which parameters should be tested for equality.

The DRF approach to DIF does not require the item parameter sets to be commensurate across groups because the measures do not directly relate to the sets of parameters. Instead, the DRF measures focus on the differences between the model-implied response functions, where the item parameter estimates are only used to build these expected response functions. Hence, any class of IRT model may be compared across groups *regardless* of whether they have commensurate parameter sets. It is anticipated, however, that the Wald test will have more power to detect true DIF effects compared to the DRF framework because it constructs the hypothesis test using more efficient parametric information. For these reasons I anticipate that the Wald test will generally outperform the comparable statistics from the DRF framework across a range of conditions.

2.4 Improvements of the DRF Framework Over the SIBTEST and DFIT Frameworks

Before comparing the performance of the DRF framework to the SIBTEST and DFIT frameworks it is worthwhile first to highlight some of the attractive characteristics of the DRF framework compared to DFIT and SIBTEST. To start, unlike DFIT and SIBTEST, which use auxiliary data-driven properties from the observed sample data, the DRF statistics use the expected scoring functions directly (e.g., Equations 1.1 and 1.2) and therefore remain in the metric of the test. The three

proposed DRF measures roughly form statistical analogs of a mean difference (cf. Equation 2.12), mean absolute difference (cf. Equation 2.13), and mean squared deviation (i.e., standard deviation; cf. Equation 2.15) of the difference between response functions across the desired latent trait range. Hence, the statistics are in effect size metrics associated with the scale of the item, bundle, or test, which are readily interpretable by the test developer. For example, if $sDTF = 1$, the researcher could conclude that, on average, the reference group scored 1 full point higher than the focal group across the selected range of $\boldsymbol{\theta}$. In addition, if $dDTF = 2$, then the researcher could conclude that, on average, the response curves deviated by 2 points over the selected integration range. Note that these example effect sizes may not be serious if the test were out of 100 total points, but may be substantial if the test were out of 20. An identical relationship holds for the DIF and DBF statistics; hence, effects sizes with meaningful metrics are available at all levels of the response bias analysis.

Because the parametric sampling method required for the DRF statistics shares a clear lineage with the graphical methods presented by Thissen and Wainer (1990), the DRF statistics naturally have graphical counterparts which are very useful for visually diagnosing the $\boldsymbol{\theta}$ locations with which the bias is greatest. Specifically, visualization is one area where the $sDTF_{\boldsymbol{\theta}}$ family of measures is useful. For example, after creating a range of desired $\boldsymbol{\theta}$ values, each $\widehat{sDRF}_{\boldsymbol{\theta}}$ estimate can be evaluated, along with its respective sampling variability, to form sets of point-wise confidence intervals. Subsequently, these intervals can be depicted graphically to indicate locations of higher or lower sampling variability in the expected difference between the test, bundle, or item response functions. Figure 2.1 demonstrates the use of this diagnostic process after a significant

signed or unsigned DTF effect is detected (or non-significant, in the case of the image on the left), and is a simple remapping of Figure 1.3 using obtained 95% confidence interval estimates for $\widehat{sDTF}_{\boldsymbol{\theta}}$. Compared to non-parametric approaches for constructing expected score plots for bias detection (e.g., Bolt & Gierl, 2006; Ramsay, 2000) these plots implicitly adjust for the differences in the latent trait distributions because of the anchoring technique, thereby demonstrating more appropriate conditional bias effects.

To demonstrate the use of $sDIF_{\boldsymbol{\theta}}$ for interpreting DIF, a small example with six focal items is depicted in Figure 2.2. In this figure, items 1 and 2 (bottom row) have non-significant $sDIF$ and $dDIF$, while item 3 (second row) has non-significant $sDIF$ but significant $dDTF$. After the sampling variability is constructed for the figures on the right, the difference between the two probability functions for items 1 and 2 clearly fall within the 95% confidence intervals across all levels of $\boldsymbol{\theta}$ (i.e., the confidence interval contains the solid horizontal reference line). Items 4, 5, and 6, however, all have significant $sDIF$ and $dDIF$ effects, which is clear in both the expected probability functions as well as the $sDIF_{\boldsymbol{\theta}}$ plots containing the respective sampling variability. Furthermore, although item 3 appears to have a large difference in the upper end of the $\boldsymbol{\theta}$ axis, the associated figure on the right makes it clear that this observed difference is not reliable when considering the sampling variability. Therefore, researchers should be less inclined to interpret such a difference between functions at extreme levels of $\boldsymbol{\theta}$ where sampling variability is clearly larger. Without the parametric sampling approach, researchers may prematurely conclude that

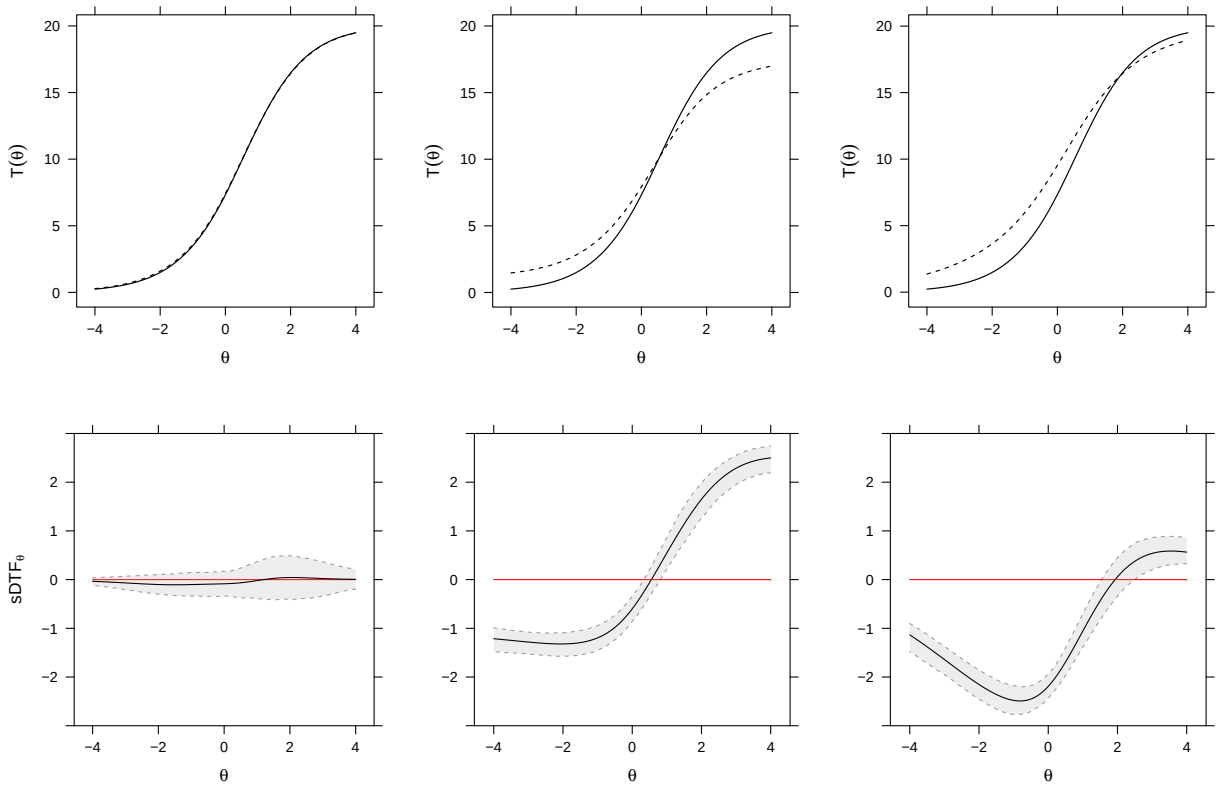

Figure 2.1: A remapping of functions from Figure 1.3 (top row) using the $sDTF_{\boldsymbol{\theta}}$ measure across a range of $\boldsymbol{\theta}$ values. Bottom row of images represent the sampled 95% confidence intervals at each $\boldsymbol{\theta}$ location shaded in gray while the observed $\widehat{sDTF_{\boldsymbol{\theta}}}$ estimates are indicated with the solid black line. The solid horizontal red line is a reference line where $sDTF_{\boldsymbol{\theta}} = 0$.

item 3 contains a large amount of bias at the positive end of $\boldsymbol{\theta}$ when in fact this effect is largely attributed to sampling uncertainty.

There are many other important properties that make the DRF statistics appealing and more flexible than the SIBTEST and DFIT frameworks. For example, because the estimation of the item parameters and $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ are carried out by FIML estimation, and because the statistics themselves do not require reference to any particular aspects of the data, missing values (which are missing at random or missing completely at random; Rubin, 1987) do not negatively impact the validity or consistency of the statistics. Similarly, the FIML-based multiple-group model estimation approach provides optimal results when the sample sizes are unequal, the distributions for $\boldsymbol{\theta}$ are different in each group, and when test-specific properties such as the length of the test, number of response categories, selection of the focal group, and so on are modified. With respect to testing DIF, DBF, and DTF, the DRF framework has a computational advantage in that all the relevant statistical variability can be obtained using the same sampled parameter set. Therefore, computing DIF for larger tests is no more computationally demanding than smaller tests, and the same fitted model can be used for DIF, DBF, and DTF testing without having to fit further IRT models and asymptotic covariance matrices.

Finally, the DRF framework has the capability to measure compensatory and non-compensatory differential functioning for numerous IRT models, where the nature of the measures is not limited by the shape of the response functions (unlike CSIBTEST, for example, which requires monotonically increasing response functions). Furthermore, the measures from the DRF framework are

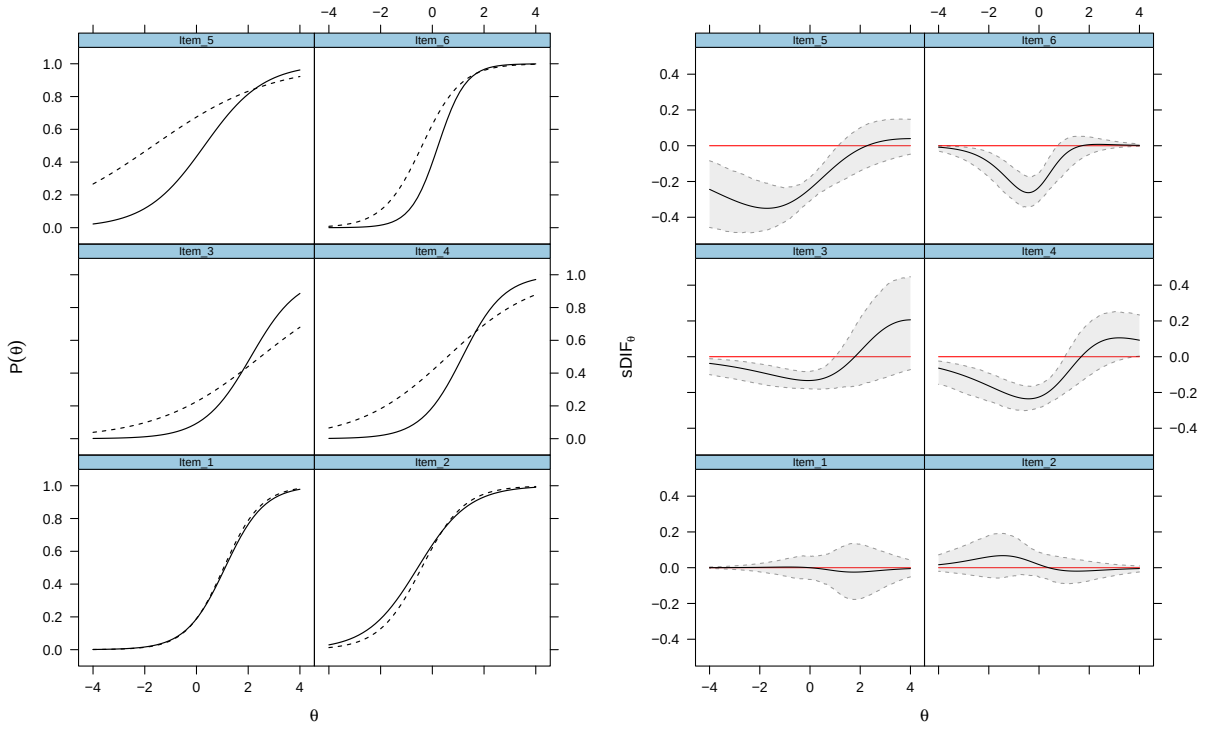

Figure 2.2: Example of six items under DIF investigation. The left block represents the probability response functions for the focal and reference group, while the right block of figures represents the $sDIF_{\boldsymbol{\theta}}$ plots with 95% confidence intervals (shaded in gray) evaluated across 1000 equally spaced $\boldsymbol{\theta}$ locations.

able to accommodate any mix of dichotomous or polytomous IRT models, do not require the use of monotonic or even parametric statistical forms, can be approximated using non-parametric estimation methods (such as bootstrapping), and will naturally accommodate multidimensional IRT tests with complex structures. These and other advanced topics are discussed in greater detail in Chapter 4.

2.5 Summary

Two previously proposed statistical frameworks for detecting differential item, bundle, and test functioning are described in this chapter, and a new approach termed the differential response functioning (DRF) framework is proposed as a solution to many of the limitations inherent in the first two frameworks. The DRF framework extends and improves upon many of the less attractive features of the SIBTEST and DFIT frameworks, and presents a unified approach to investigating differential functioning. DRF provides both compensatory and non-compensatory marginal statistics when testing DIF, DBF, and DTF, and provides an effective approach to diagnose conditional differential functioning (i.e., at specific $\boldsymbol{\theta}$ values) which are straightforward for isolated hypothesis testing as well as for presenting the range of differences between the response functions graphically. Refer to Table 2.1 for a summary of the statistics available for detecting differential functioning via marginal statistics.

On the surface, the DRF framework appears more flexible and promising than either the SIBTEST or DFIT frameworks for computing compensatory and non-compensatory differential functioning

			Compensatory			Non-Compensatory		
Framework	Model-Based	Full-Information[a]	DIF	DBF	DTF	DIF	DBF	DTF
Wald	Yes	Yes	—	—	—	W^2	—	—
SIBTEST	No	No	SIBTEST	SIBTEST	SIBTEST	CSIBTEST	—	—
DFIT	Yes	No	*CDIF*	*CDBF*	*DTF*	*NCDIF*	*NCDBF*	*NCDTF*
DRF	Yes	Yes	*sDIF*	*sDBF*	*sDTF*	*uDIF*/*dDIF*	*uDBF*/*dDBF*	*uDTF*/*dDTF*

Table 2.1: Breakdown of the SIBTEST, DFIT, and DRF framework statistics by type of test and compensatory nature for unidimensional tests consisting only of dichotomous items.

[a]Refers to whether missing response data are handled efficiently and effectively.

statistics. However, the DRF framework may be limited by a number of properties; namely: the quality of the maximum-likelihood solution (e.g., local minima solutions will not behave correctly), the quality and number of the selected anchor items, the number of samples to draw, the range of the $\boldsymbol{\theta}$ integration, and the quality of the parameter covariance matrix estimate. Clearly, if there are too few anchors used, or the anchors selected contain DIF, then the maximum-likelihood scaling of the hyper-parameters may not represent the most optimal scaling of the expected response functions. Note, however, that several of these limitations are present in the DFIT and SIBTEST frameworks, as well as the Wald and likelihood-ratio tests, and may in fact be less problematic in the DRF framework. To investigate the overall performance under comparable conditions, the three frameworks (as well as the Wald test for DIF) are compared in the next chapter using Monte Carlo simulations to evaluate their respective Type I error control and power to detect true differential functioning effects.

3 Monte Carlo Simulations Comparing the Differential Testing Frameworks

This chapter presents a set of extensive Monte Carlo simulation studies which evaluate the performance of the SIBTEST, DFIT, and DRF frameworks with respect to detecting compensatory and non-compensatory DIF, DBF, and DTF. Empirical Type I error rates are obtained in conditions where no DIF is present and in tests for DTF and DBF when DIF is present but the DIF effects completely cancel. Power rates for detecting population-level DIF, DBF, and DTF effects are also investigated using sets of predefined DIF items that are held constant across sections. Additional properties with respect to the newly developed DRF framework will also be investigated to determine the general properties of the parametric sampling scheme not investigated by Chalmers et al. (2016).

3.1 Selecting Conditions for Direct Comparison Between Frameworks

Although the three frameworks have the same general purpose of identifying differential effects, prior Monte Carlo simulation work has generally not been comparable across published studies. This inconsistency is mainly because a) the number of anchor items used in previous simulations has not been commensurate; b) the method of equating the parameters across groups has not been consistent; c) the types of DIF investigated has been discrepant; and d) the variety of simulation conditions investigated has generally been non-overlapping (e.g., different sample sizes, group sizes, latent trait distributions, and so on).

The following is an example of why comparing the DFIT frameworks to the SIBTEST or DRF framework is problematic. Most, if not all, applications of the statistics from the DFIT framework have used a sub-optimal linking method when DIF was present. The linking approaches adopted were based on statistical techniques whereby all the items — including those with known DIF — were used to equate the parameters across the groups. Hence, it was known a priori that the group parameters were not optimally equated and instead were confounded by the magnitude of DIF, the number of items containing DIF, and so on. Because this linking method has generally not been optimal, the statistics from the DFIT framework are re-evaluated after the parameters are equated using the more optimal likelihood-based multiple-group estimation approach. This linking approach should provide a more realistic application of detecting response bias, particularly when DIF is present.

With respect to the SIBTEST procedure, several authors (including Shealy & Stout, 1993) have used — and indeed recommend using — all non-focal items as anchor items *regardless* of whether DIF is suspected. In these situations there are varying degrees of contamination in the linking process which may not behave as optimally as what Shealy and Stout (1993) had previously witnessed[8]. To avoid this problem in the following simulations, the use of a small number of uncontaminated anchor items (five or ten) is selected so that each framework can be compared given identical information regarding group equivalence. That being said, the use of contaminated anchor items is also investigated for the SIBTEST procedures to determine whether Shealy and Stout's recommendation is appropriate for the DIF, DBF, and DTF detection given the predefined DIF items under investigation.

3.2 Global Simulation Details

There is a number of conditions which ultimately affect the ability to detect differential effects, including, but not limited to, the number of test items, sample size, whether the group sizes are equal, latent trait distribution characteristics, number of anchor items used, and, particularly for the DFIT framework, which group is selected as the focal group. Power rates are further influenced by the population DIF characteristics as well, which may include the type of DIF present (uniform

[8]Shealy and Stout (1993) acknowledge the possibility of contaminated anchor items but demonstrated in their simulation study that contaminated anchor items are generally a non-issue. However, this finding is suspicious given the current understanding of contaminated anchor items (e.g., see Millsap, 2011), and therefore will be re-investigated in this chapter.

versus non-uniform), the magnitude of the parameter differences (and consequently the expected response functions), and whether optimal statistics for examining DIF are selected (e.g., using statistics for non-uniform DIF when uniform DIF is present).

In addition, the detection of DBF and DTF is further affected by the number of items containing DIF and their respective propagation effects at different $\boldsymbol{\theta}$ levels. For example, items containing DIF may cause larger (i.e., amplification) or smaller (i.e., cancellation) degrees of separation between the composite response functions across different $\boldsymbol{\theta}$ levels. Furthermore, in the extreme case where the DIF effects completely cancel, nominal Type I error rates should obtained when testing DBF and DTF. Empirical Type I error rates arising from complete cancellation have never been studied directly in any of the literature concerning the frameworks, although the cancellation effect appears to be indirectly present in some studies (e.g., see Chalmers et al., 2016; Nandakumar, 1993; Oshima et al., 1998). Therefore, this chapter also investigates the performance of the detection frameworks in the presence of complete cancellation.

The following Monte Carlo simulations are organized into sections which focus solely on empirical DIF and DBF or DTF detection behavior when investigating unidimensional 2PL models for dichotomous response data. In the first section simulations pertaining to DIF detection are constructed to obtain empirical Type I error rates when no DIF is present, and power rates are investigated when compensatory and non-compensatory DIF effects were present. In the next section regarding DBF and DTF detection, Type I error rates are evaluated in conditions where no DIF is present as well as in situations where there was DIF but the effects were perfectly balanced

(i.e., completely cancel across the item bundles). Power rates are also investigated to determine the effects of compensatory and non-compensatory DIF propagation in the respective bundle and test response functions.

3.2.1 Computational Considerations

Before beginning the simulations, the initial setup for the DFIT and DRF frameworks, as well as the Wald testing approach, requires some special considerations. Because these approaches involve the estimation of IRT models, the question of whether *non-focal* items should be included or omitted from the fitted models must be considered. Non-focal items are items which are included in the modeling process but are neither focal nor anchor items; hence, they are included in the full model because they may include DIF, though the test analysis is unclear whether they do or not. Omitting the non-focal items so that only the anchor and focal items are used results in fewer parameters to estimate because fewer items are required. From a computational standpoint, the DRF and DFIT frameworks, as well as the Wald approach for DIF testing, may benefit when using fewer items because the number of parameters to estimate is considerably smaller. With respect to the DRF framework and Wald test, computing the information matrix also takes considerably fewer computational resources because fewer terms are required to build $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$. On the other hand, omitting non-focal items also discards valuable statistical information, therefore the quality of the $\hat{\boldsymbol{\theta}}$ estimates for the DFIT framework may suffer, the ML parameter estimates may not be as precise, or the variability in the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix may unintentionally grow to be too large.

Investigating the performance of the detection frameworks when only the anchor and focal items are modeled (i.e., leaving the non-focal items unmodeled) is of interest because it may be beneficial to use a smaller yet more reliable subset of items to detect differential response effects. Omitting non-focal items will demonstrate how well the statistics perform in very short tests; however, doing so generally decreases the power of the statistics which are able to use the axillary information from the non-focal items. Furthermore, because the SIBTEST framework is not able to accommodate non-focal items, it is important to include these item subsets so that the properties can be directly compared given the same amount of empirical information. Therefore, where appropriate the inclusion and omission of the non-focal items in the fitted IRT models are investigated for the DFIT, DRF, and Wald test statistics.

3.3 Specific Details of the Simulations

Throughout the following Monte Carlo simulations a number of auxiliary characteristics were kept constant to ensure that the results were comparable, consistent, and easier to interpret across sections. The Monte Carlo simulation factors under investigation within each simulation design were:

- the number of anchors used to equate the groups (5 versus 10),
- test length (20, 30 and 40 items),
- total sample size (900, 1800, and 2700),

- latent trait distribution ($\mathcal{N}(0, 1)$ was used for the reference group while the focal group was either $\mathcal{N}(0, 1)$ or $\mathcal{N}(1/2, 2/3)$), and
- the effect of equal and unequal sample sizes across groups (where the size of focal group either equaled or was half the size of the reference group).

With respect to the statistics within the DFIT framework, the sizes of the reference and focal groups were also switched in each condition to determine the complementary effect of unequal sample sizes when the focal group was larger or smaller than the reference group. Each combination was investigated using 1000 independent Monte Carlo replications.

The simulated datasets were generated and analyzed using estimation functions from the `mirt` package in R (Chalmers, 2012) while the Monte Carlo simulation work-flow was controlled by the `SimDesign` package (Chalmers, 2016b; Sigal & Chalmers, 2016). Models were fitted using `mirt`'s multiple-group estimation engine with the EM algorithm. Model estimation was terminated when all the parameter estimates changed less than |.0001| across successive EM cycles. The reference and focal groups were equated using the likelihood-based procedure described in the previous chapter whereby all parameters within each select anchor item were constrained to be equal across groups.

The $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix required for the DRF statistics and Wald test was estimated using the computationally efficient cross-product approximation approach (Pawitan, 2001). The cross-product approximation is known to be asymptotically equivalent to the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix based on the ob-

served information. Therefore, given the somewhat larger sample sizes under investigation, it was appropriate to use this variant as a proxy to the exact observed information matrix. Expected a posteriori (EAP) estimates obtained from the fitted multiple-group models were used as the required $\hat{\boldsymbol{\theta}}$ values for the DFIT statistics. Using the same fitted multiple-group IRT models that were used to compute the DFIT statistics, the sampling variability for the DRF statistics was obtained after sampling $M = 500$ parameter sets and numerically integrating across the response functions using rectangular quadrature in the range $[-6, 6]$ for $\boldsymbol{\theta}$.

The implementation of the original SIBTEST procedure was identical to Shealy and Stout's (1993) description (e.g., minimum matched score frequency of 2, excluded lowest and highest matched scores, etc.). Where applicable, the SIBTEST was also compared to the uncorrected counterpart (SIBTEST_{UC}; cf. Equation 2.1) to determine the effect of the regression adjustment in conditions where the latent trait distributions were identical. When the latent trait distributions are identical the uncorrected SIBTEST procedure should return nominal Type I error rates. As well, the CSIBTEST statistic followed the same details presented by Li and Stout (1996) for non-compensatory DIF with only one crossing location in the response functions.

Unless otherwise specified, all detection rates were calculated using a nominal α rate of .05. To visualize aberrant empirical detection rates, tables containing Type I errors that were outside of Bradley's (1978) so-called 'liberal' robustness interval definition [.025 .075] (or in specific situations where $\alpha = .01$, [.005, .015]) are emphasized with **bold** font. Detection rates for statistics which were potentially influenced by the use of non-focal items are presented in separate appen-

dices to help conserve space. If the statistics did not contain indirect information from the non-focal items (such as SIBTEST and CSIBTEST with a predetermined number of anchor items and the subset version for the DRF, DFIT, and Wald statistics) then these are displayed in-text. Finally, power rates from the appendices were averaged across the overall test length so that the marginal results could be more easily compared to the detection rates that did not use information from the non-focal items; this process was repeated for Type I error rates as well, but only when the test length was deemed to be an irrelevant condition which could be averaged across.

3.3.1 Stability of the Parametric Sampling Method for the DRF Statistics

3.3.1.1 Number of Imputation Sets

Before performing the simulation studies it is first important to determine whether $M = 500$ draws is sufficient for obtaining stable p-values for the DRF statistics. To determine if the precision of the p-values was adequate, three separate IRT models containing no differential effects were generated and investigated using 500, 1000, 2000, 4000, and 8000, and 16000 draws. In preliminary simulation work, when the average simulated p-values were closer to 0 or 1 the variability tended to decrease, therefore models were selected such that all the observed p-values were within the range of .3 to .7.

Multiple-group IRT models containing 2PL items were generated for tests of length 20, 30, and 40. The latent distribution for the focal group was $\mathcal{N}(1/2, 2/3)$ and the reference group was

		Signed Measures			Deviation Measures		
Type	Number of Draws	20 Items	30 Items	40 Items	20 Items	30 Items	40 Items
DIF	500	.012 (.544)	.013 (.473)	.016 (.404)	.021 (.621)	.019 (.707)	.026 (.602)
	1000	.010 (.544)	.009 (.473)	.010 (.403)	.016 (.624)	.015 (.707)	.017 (.600)
	2000	.006 (.544)	.007 (.473)	.007 (.403)	.011 (.624)	.010 (.708)	.013 (.601)
	4000	.004 (.544)	.004 (.473)	.005 (.402)	.008 (.624)	.007 (.709)	.009 (.602)
	8000	.003 (.544)	.003 (.473)	.004 (.402)	.005 (.624)	.005 (.709)	.006 (.601)
	16000	.002 (.544)	.002 (.473)	.002 (.403)	.004 (.624)	.003 (.709)	.004 (.602)
DBF	500	.010 (.629)	.012 (.526)	.009 (.654)	.024 (.522)	.026 (.351)	.024 (.448)
	1000	.007 (.629)	.010 (.526)	.006 (.655)	.016 (.521)	.019 (.350)	.018 (.449)
	2000	.005 (.630)	.006 (.526)	.005 (.655)	.012 (.522)	.013 (.349)	.012 (.449)
	4000	.004 (.629)	.004 (.526)	.003 (.655)	.008 (.522)	.009 (.349)	.008 (.448)
	8000	.002 (.629)	.003 (.526)	.002 (.655)	.005 (.521)	.006 (.350)	.006 (.448)
	16000	.001 (.629)	.002 (.526)	.001 (.655)	.004 (.521)	.004 (.349)	.004 (.448)
DTF	500	.012 (.577)	.011 (.604)	.012 (.600)	.022 (.638)	.022 (.548)	.021 (.669)
	1000	.008 (.576)	.007 (.605)	.008 (.599)	.015 (.639)	.016 (.549)	.015 (.669)
	2000	.006 (.577)	.005 (.604)	.006 (.599)	.010 (.640)	.011 (.549)	.010 (.669)
	4000	.004 (.577)	.004 (.604)	.004 (.599)	.008 (.639)	.008 (.548)	.007 (.670)
	8000	.003 (.577)	.002 (.604)	.003 (.599)	.005 (.639)	.005 (.549)	.005 (.670)
	16000	.002 (.577)	.002 (.604)	.002 (.599)	.004 (.640)	.004 (.549)	.003 (.670)

Table 3.1: Standard deviation of the parametrically sampled p-values (with the average p-value across 500 independent draws in brackets) for the DRF statistics when increasing the number of draws and test length.

$\mathcal{N}(0, 1)$. The multiple-group IRT models contained five anchor items; thus, a total of ten parameters were constrained to be equal across groups. Each simulation condition was replicated 500 times to determine the relative variability across independent sampled sets. All DTF, DBF (with eight focal items), and DIF statistics from the DRF framework were investigated. Table 3.1 contains the results from this brief simulation study.

As is clear in Table 3.1, increasing the number of draws systematically decreased the variability of the obtained p-values. Furthermore, this appeared to be the only major effect that contributed to the consistency of the observed p-values. For example, the most extreme collapsing of second-

order information appeared in the condition where the test contained 40 items with a total of 152 ($40 \cdot 2 \cdot 2 - 5 \cdot 2 + 2$) freely estimated parameters. However, even with 152 freely estimated parameters the joint variability used to obtain the sampling variability of the DRF statistics did not appear to be any different than the DIF conditions where only four parameters contributed to the variability. The deviation-based DRF statistics appeared to have slightly more variability compared to the signed statistics, but this effect was not dramatic enough to affect any of the subsequent simulation (or empirical analysis) results. This brief simulation suggests that the standard error estimates for DRF statistics will be sufficiently stable with as few as 500 sampled parameter sets.

3.3.1.2 Integration Range

Another important area worth investigating for the parametric sampling approach is whether the integration range has any effect on the Type I error detection rates. Naturally, modifying the integration range to focus on particular $\boldsymbol{\theta}$ regions of interest will affect the power to detect DRF when it exists; for instance, focusing on a wide $\boldsymbol{\theta}$ range when bias only occurs within a small $\boldsymbol{\theta}$ range will be less powerful because the magnitude of the true effect will be lessened due to averaging across a wider function. However, the use of different integration ranges is only justifiable if the Type I error rates are not affected by the integration range. To investigate this effect, integration ranges between $[-2, 2]$, $[-4, 4]$, $[-6, 6]$, $[-8, 8]$, and $[-10, 10]$ across $\boldsymbol{\theta}$ were investigated in tests of length 20, 30, and 40. To remove the likelihood of non-convergent solutions or inaccurate $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrices when using the cross-product approximation, N was set to 20,000.

DRF Measure	Test Length	$[-2, 2]$	$[-4, 4]$	$[-6, 6]$	$[-8, 8]$	$[-10, 10]$
$sDIF$	20	.032	.036	.032	.036	.034
	30	.062	.054	.052	.054	.052
	40	.054	.054	.056	.054	.048
$sDBF$	20	.054	.052	.050	.054	.054
	30	.044	.036	.034	.034	.032
	40	.036	.040	.052	.044	.048
$sDTF$	20	.066	.050	.060	.058	.054
	30	.044	.050	.042	.040	.044
	40	.054	.048	.048	.044	.040
$dDIF$	20	.038	.038	.034	.038	.034
	30	.070	.054	.060	.064	.058
	40	.064	.052	.046	.048	.052
$dDBF$	20	.054	.056	.056	.052	.060
	30	.064	.056	.054	.054	.048
	40	.052	.054	.056	.060	.052
$dDTF$	20	.062	.060	.064	.062	.058
	30	.068	.058	.054	.058	.052
	40	.062	.054	.062	.058	.060

Table 3.2: Type I error rates for the DRF statistics when modifying the integration range across $\boldsymbol{\theta}$.

Each simulation condition was repeated 500 times to obtain reasonable stability in the empirical Type I error rate estimates. Empirical Type I error rates were collected at $\alpha = .05$ for DIF, DBF (where five focal items were selected), and DTF; these results are results are displayed in Table 3.2. The results generally suggest that the Type I error rates were not influenced by the integration range selected for any of the DRF measures. Therefore, the integration range selected when detecting differential effects only appears to affect the power to detect DRF and not the Type I error rates.

3.4 Differential Item Functioning

There are many characteristics that can influence the detection of DIF. For example, with respect to the 2PL model any difference in intercept parameters (δ) will cause a systematic shift in the probability response functions so that one group will systematically have a greater probability of answering correctly. Consequently, statistics such as SIBTEST and $sDIF$ should be effective at detecting this type of DIF because the probability functions will, by definition, have no compensatory effects (i.e., will never cross). However, when the difference in slope parameters (α) is the cause of DIF then the response probability function of one group will be lower than the other at some levels of $\boldsymbol{\theta}$ and higher at other levels of $\boldsymbol{\theta}$. Non-compensatory statistics are more useful for this type of DIF, such as $NCDIF$, CSIBTEST, $dDIF$, and W^2, because these are less affected by non-uniform differences in the response functions. When both of these parameters are different across groups then various compensatory and non-compensatory effects can arise. The following study will focus on testing for DIF by generating differences in the α and δ parameters sets simultaneously for the 2PL model to construct varying degrees of compensatory and non-compensatory effects.

Unless otherwise specified, all slope parameters were drawn from a log-normal distribution, $\alpha \sim \log \mathcal{N}(0.2, 0.2)$, and the intercept parameters were drawn from a normal distribution $\mathcal{N}(0, 1/2)$. The statistics under investigation in the following simulation study were: SIBTEST and SIBTEST$_{UC}$, where the focal set contained only one item, CSIBTEST, $NCDIF$ from the DFIT framework with

subscripts to indicate whether the focal group was larger ($NCDIF_{LF}$) or smaller ($NCDIF_{SF}$) than the reference group, $sDIF$ and $dDIF$ from the DRF framework, and the Wald test (with two degrees of freedom). In the Type I error section all items were constructed to contain no DIF.

3.4.1 Type I Error Rates

Tables 3.3 and 3.4 contain the empirical Type I error rates when only the anchor and focal items were modeled, while the tables in Appendix A contain the Type I error rates for statistics which are capable of including information from non-focal items. Beginning with the $NCDIF$ statistic from the DFIT framework, it was apparent that its Type I error rates were extremely inflated across all simulation conditions. When including the non-focal items in the fitted models (Appendix A), the χ^2 test for $NCDIF$ had an average empirical Type I error rate of .838 ($SD = .032$) when the focal group was equal to or larger than the reference group, and .815 ($SD = .024$) when the focal group was equal to or smaller than the reference group. Omitting the non-focal items from the fitted models (Tables 3.3 and 3.4) resulted in similar detection rates. $NCDIF$'s detection rates increased when lager sample sizes and test lengths were studied, when the group sizes were unequal, and when the latent trait distributions were unequal. Using only a subset of the test items to compute the $\hat{\theta}$ estimates resulted in comparably high Type I error rates which demonstrated similar detection rate patterns; hence, using item subsets does not appear to be a helpful strategy for the DFIT framework.

Anchors	Focal Distribution	Sample Sizes	$dDIF$	$dDIF_M$	Wald	Wald$_M$	CSIBTEST	$NCDIF_{LF}$	$NCDIF_{SF}$	$P(NCDIF_{LF} > .006)$	$P(NCDIF_{SF} > .006)$
5	$\mathcal{N}(0,1)$	450/450	**.023**	.025	.034	.033	**.000**	**.757**	**.754**	**.189**	**.194**
		900/900	.035	.040	.042	.041	**.000**	**.787**	**.785**	**.036**	**.041**
		1350/1350	.033	.047	.055	.047	**.000**	**.822**	**.814**	.007	.006
		600/300	.035	.039	.042	.040	**.000**	**.810**	**.775**	**.056**	**.055**
		1200/600	**.024**	.047	.035	.042	**.000**	**.833**	**.803**	**.000**	**.002**
		1800/900	.044	.051	.050	.049	**.000**	**.853**	**.825**	**.000**	**.000**
	$\mathcal{N}(1/2,2/3)$	450/450	**.021**	**.023**	.036	.032	**.000**	**.809**	**.737**	**.290**	**.218**
		900/900	.029	.039	.047	.043	**.000**	**.827**	**.817**	**.087**	**.032**
		1350/1350	.029	.039	.056	.049	**.000**	**.872**	**.831**	**.030**	.009
		600/300	.028	.034	.035	.042	**.000**	**.862**	**.783**	**.094**	**.054**
		1200/600	.036	.041	.057	.039	**.000**	**.886**	**.809**	.010	.007
		1800/900	.026	.049	.043	.045	**.000**	**.889**	**.824**	**.001**	**.000**
10	$\mathcal{N}(0,1)$	450/450	.033	.026	.033	.035	.064	**.773**	**.778**	**.085**	**.081**
		900/900	.037	.041	.042	.040	.054	**.814**	**.801**	.011	.012
		1350/1350	.046	.045	.050	.043	.063	**.836**	**.829**	**.000**	**.000**
		600/300	.050	.037	.056	.039	.061	**.804**	**.775**	**.016**	.015
		1200/600	.042	.049	.044	.047	.065	**.846**	**.821**	**.001**	**.001**
		1800/900	.039	.051	.040	.048	.054	**.865**	**.841**	**.000**	**.000**
	$\mathcal{N}(1/2,2/3)$	450/450	**.020**	**.022**	.034	.038	.056	**.793**	**.776**	**.138**	**.085**
		900/900	.034	.034	.042	.040	.043	**.856**	**.843**	**.024**	.013
		1350/1350	.049	.039	.054	.042	.066	**.863**	**.848**	.011	.003
		600/300	.034	.033	.049	.036	.054	**.843**	**.806**	**.036**	**.021**
		1200/600	.031	.042	.041	.046	.070	**.887**	**.829**	**.000**	**.000**
		1800/900	.049	.048	.062	.048	.062	**.893**	**.855**	**.000**	**.001**

Table 3.3: Type I error rates for non-compensatory statistics when testing DIF. $dDIF_M$ and Wald$_M$ represent the marginal detection rates after averaging over the number of test items in Appendix A, while the remainder of the statistics used only the information provided by the anchor items and a single focal item. Type I error rates greater than .075 and less than .025 are highlighted in bold.

Focusing on Raju et al.'s (1995) rule-of-thumb approach of using the cutoff of .006 for the observed $NCDIF$ values, the simulation results revealed that the distribution of this cutoff approach was primarily influenced by sample size and the number of anchors selected. As sample size increased, the proportion of $NCDIF$ values greater than the cutoff began to approach zero, experiencing floor and positive skewness effects, while in smaller sample sizes the variability of the $NCDIF$ values was considerably larger. These effects occurred regardless of the test length and number of non-focal items included in the model. When using more anchor items the proportion of values above the cutoff tended more rapidly towards zero as well, further indicating that the magnitude of $NCDIF$ was affected by model fitting properties. The marginal effects in Figure 3.1 demonstrate that this rule-of-thumb is liberal in smaller sample sizes and conservative in larger sample sizes regardless of the nominal α level selected.

Anchors	Focal Distribution	Sample Sizes	$sDIF$	$sDIF_M$	SIBTEST	SIBTEST$_{UC}$
5	$\mathcal{N}(0,1)$	450/450	.034	.029	.058	.055
		900/900	.038	.039	.043	.040
		1350/1350	.050	.045	.061	.049
		600/300	.035	.036	.048	.042
		1200/600	.040	.044	.063	.048
		1800/900	.047	.045	.052	.046
	$\mathcal{N}(1/2,2/3)$	450/450	**.020**	.025	**.077**	–
		900/900	.036	.036	**.084**	–
		1350/1350	.056	.040	**.088**	–
		600/300	.029	.033	.074	–
		1200/600	.044	.041	**.084**	–
		1800/900	.045	.048	.069	–
10	$\mathcal{N}(0,1)$	450/450	**.023**	.028	.045	.045
		900/900	.032	.039	.046	.034
		1350/1350	.055	.040	.049	.056
		600/300	.048	.035	.059	.061
		1200/600	.037	.043	.046	.042
		1800/900	.044	.046	.045	.039
	$\mathcal{N}(1/2,2/3)$	450/450	**.022**	**.023**	.049	–
		900/900	.045	.035	.057	–
		1350/1350	.061	.042	**.076**	–
		600/300	.040	.037	.063	–
		1200/600	.034	.043	.057	–
		1800/900	.053	.047	**.079**	–

Table 3.4: Type I error rates for compensatory statistics testing DIF. $sDIF_M$ represents the marginal detection rates after averaging over the number of test items in Appendix A, while the remainder of the statistics used only the information provided by the anchor items and a single focal item. Type I error rates greater than .075 and less than .025 are highlighted in bold.

SIBTEST, on the other hand, performed considerably better than the DFIT framework. SIBTEST achieved an average empirical detection rate of .066 (SD = .014) and .055 (SD = .011) when using five and ten anchors, respectively, and .047, (SD = .006) when using all non-focal items as

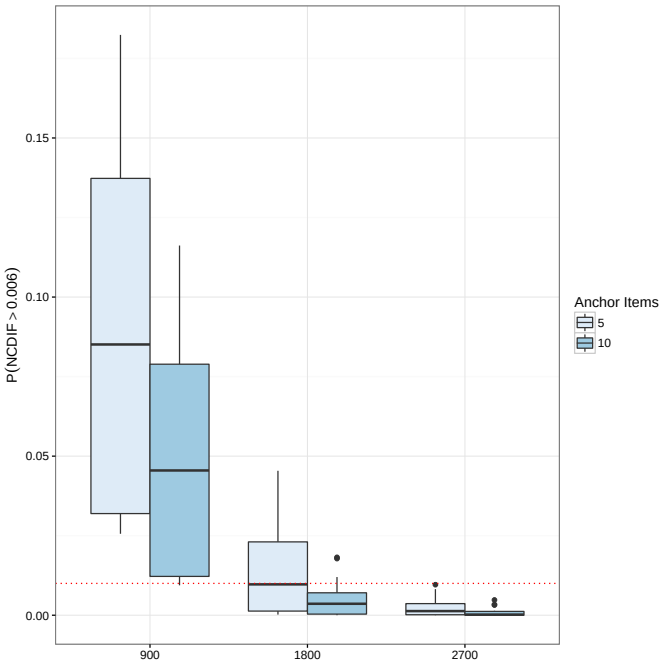

Figure 3.1: Proportion of *NCDIF* values less than the cutoff of .006 for different sample sizes and number of anchors. The dotted red line indicates the nominal rate of $\alpha = .01$, and the darker distributions indicate that 10 anchors were used (lighter distribution has 5 anchor items).

anchors (Appendix A). However, SIBTEST was strongly influenced by the number of anchor items selected and whether the latent distributions were equal. SIBTEST generally demonstrated more liberal Type I error rates when the latent distributions were unequal, and this effect was worse when only five anchor items were selected. For the uncorrected SIBTEST in the conditions where the latent trait distributions were equal, the detection rates were .046 ($SD = .005$) and .046 ($SD = .010$) when five and ten anchors were used, respectively, and .046 ($SD = .004$) when all non-focal items in the test were used as anchors. These results generally indicated that the uncorrected SIBTEST

procedure behaved appropriately and that the regression adjustment used for SIBTEST did not affect the results when the latent trait distributions were equal.

CSIBTEST's Type I error rates unfortunately did not behave as well as SIBTEST's for investigating non-compensatory DIF. In the conditions where only five anchors were used, the CSIBTEST statistic never rejected the true null hypothesis; therefore, all the Type I error rate estimates were exactly equal to 0. This indicates an important limitation in the CSIBTEST statistic when only a small number of anchors are known a priori. When ten anchor items were used the CSIBTEST Type I error rates were more reasonable (.057, $SD = .007$), and when using all non-focal items as anchors (see Appendix A) the rates appeared to be closer to the nominal level (.058, $SD = .005$). Though not studied herein, an improved approach for the CSIBTEST statistic has recently been proposed (Chalmers, in review) to amend many of the inherent issues with Li and Stout's (1996) version of the statistic, including an approach which provides more optimal Type I error control.

Finally, the DIF statistics from the DRF framework, as well as the Wald test, both provided Type I error rates consistently close to the nominal α across all studied conditions when all non-focal items were included in the models (see Appendix A). When omitting the non-focal items from the fitted models, the Wald test continued to behave optimally; however, the DRF statistics occasionally had slightly conservative Type I error rates. Primarily, the DRF statistics become more conservative in smaller sample sizes and when only five anchor items were used, generally indicating that the quality of the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ estimate, as well as the precision of the ML parameter estimates, negatively affected the results. Therefore, using the cross-product approximation to

compute $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ in very short tests, or when small subsets of items are used instead of the full item set, the DRF statistics should be regarded as slightly conservative[9].

3.4.2 Power Rates

To study the detection rates from the three differential response frameworks when DIF is present, a small selection of items was generated with a predetermined amount of response bias. The Monte Carlo simulation results presented below pertain to the power to detect DIF when five separate items contained DIF within each test. The simulation factors under investigation were carried over from the previous section on Type I error rates to determine how these properties affect power rates. However, because of the extremely poor Type I error control and lack of unconditional interpretability of the *NCDIF* statistics the DFIT framework was not included in the following Monte Carlo simulations.

The five items containing DIF were generated from the parameter sets $\delta = [1, 0.5, 0, -.5, -1]$ and $\alpha = [1, 1.25, 1.5, 1.75, 2]$ in the reference group while the sets $\delta = [0.7, 0.2, -0.3, -0.8, -1.3]$ and $\alpha = [0.5, 0.75, 1, 1.25, 1.5]$ were used for the focal group. The expected probability functions for these parameter sets can be seen in Figure 3.2, where the dashed lines indicate the expected probability functions of the focal group. These DIF items were organized to elicit different magnitudes of compensatory and non-compensatory DIF effects across the range of $\boldsymbol{\theta}$. As is apparent from Figure 3.2, DIF Item 1 demonstrates a relatively large amount of cancellation across the full

[9]The use of alternative $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ estimates was investigated but is not included in this chapter. See Chapter 4 for further discussion.

range of $\boldsymbol{\theta}$, while DIF Item 5 contains very little cancellation. Therefore, it is anticipated that compensatory detection statistics such as *sDIF* and SIBTEST will be less effective at detecting DIF for Item's 1 and 2 compared to the other DIF items.

Tables containing the power rates when the number of non-focal items in the test was a factor are included in Appendix B, while the remaining rates are included in Tables 3.5 through 3.7. The general trend across all statistics was that, as should be expected, increasing the sample size and number of anchor items resulted in higher rejection rates. As well, when the group sizes were unequal, or the latent trait distributions differed across groups, all statistics resulted in higher rejection rates. In the designs where non-focal items were included in the fitted models the rejection rates for *sDIF*, *dDIF*, and Wald tests were generally higher than the related designs where the non-focal items were excluded. This result indicated that the non-focal items added extra information for determining the ML parameter estimates and provided a less variable $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix. Furthermore, the compensatory statistics were more powerful for the DIF items that demonstrated smaller cancellation effects across the $\boldsymbol{\theta}$ range (e.g., DIF items 3, 4, and 5) and less powerful for the items with larger cancellation effects (e.g., DIF item 1 and 2). Finally, the non-compensatory statistics generally displayed more power than the compensatory statistics across all DIF items studied.

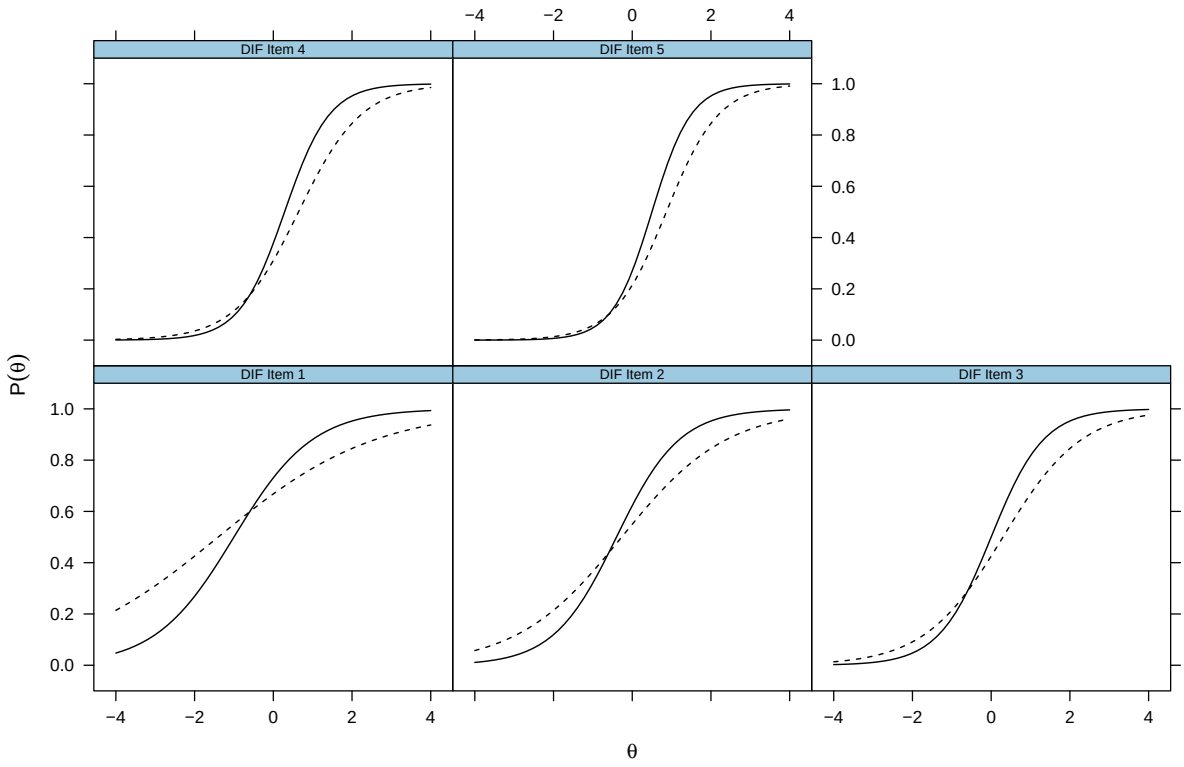

Figure 3.2: Probability functions for population-level DIF. DIF items are organized to have progressively smaller cancellation effects, where Item 1 has the most cancellation and Item 5 the least.

				Compensatory			Non-compensatory				
Anchors	DIF Item	Focal Distribution	Equal Sample Sizes	*sDIF*	$sDIF_M$	SIBTEST	*dDIF*	$dDIF_M$	Wald	Wald_M	CSIBTEST
5	1	$\mathcal{N}(0,1)$	Yes	.099	.104	.238	.190	.292	.529	.995	.000
			No	.194	.218	.394	.584	.855	.785	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.087	.068	.407	.073	.091	.629	.998	.000
			No	.103	.104	.683	.303	.615	.843	1.000	.000
	2	$\mathcal{N}(0,1)$	Yes	.110	.096	.339	.225	.451	.560	.993	.000
			No	.152	.178	.513	.682	.921	.789	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.082	.079	.500	.069	.238	.625	.998	.000
			No	.132	.140	.749	.554	.894	.862	1.000	.000
	3	$\mathcal{N}(0,1)$	Yes	.427	.439	.427	.317	.555	.537	.984	.000
			No	.791	.803	.627	.770	.938	.806	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.443	.477	.585	.269	.578	.646	.993	.000
			No	.783	.803	.835	.792	.956	.870	1.000	.000
	4	$\mathcal{N}(0,1)$	Yes	.503	.601	.490	.320	.561	.540	.935	.000
			No	.873	.922	.736	.772	.918	.836	.997	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.659	.709	.676	.358	.668	.647	.962	.000
			No	.933	.951	.895	.823	.944	.876	1.000	.000
	5	$\mathcal{N}(0,1)$	Yes	.413	.600	.567	.316	.510	.536	.682	.000
			No	.841	.921	.792	.737	.881	.838	.934	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.577	.728	.713	.319	.629	.613	.750	.000
			No	.890	.956	.918	.776	.928	.860	.939	.000
10	1	$\mathcal{N}(0,1)$	Yes	.117	.099	.242	.301	.332	.656	.999	.279
			No	.220	.217	.424	.799	.883	.910	1.000	.435
		$\mathcal{N}(1/2,2/3)$	Yes	.080	.056	.473	.092	.083	.696	1.000	.301
			No	.145	.118	.769	.563	.674	.931	1.000	.435
	2	$\mathcal{N}(0,1)$	Yes	.107	.105	.342	.449	.515	.670	1.000	.335
			No	.178	.183	.564	.891	.946	.895	1.000	.468
		$\mathcal{N}(1/2,2/3)$	Yes	.098	.094	.558	.217	.293	.769	1.000	.330
			No	.148	.154	.841	.854	.922	.938	1.000	.474
	3	$\mathcal{N}(0,1)$	Yes	.535	.515	.506	.575	.658	.666	.992	.358
			No	.841	.847	.738	.911	.959	.890	1.000	.497
		$\mathcal{N}(1/2,2/3)$	Yes	.509	.507	.653	.577	.673	.764	.998	.401
			No	.839	.833	.924	.957	.973	.951	1.000	.514
	4	$\mathcal{N}(0,1)$	Yes	.677	.684	.604	.580	.670	.662	.955	.359
			No	.957	.950	.858	.924	.952	.917	.999	.489
		$\mathcal{N}(1/2,2/3)$	Yes	.782	.806	.744	.682	.780	.757	.974	.381
			No	.978	.980	.959	.961	.978	.946	.999	.539
	5	$\mathcal{N}(0,1)$	Yes	.721	.699	.689	.606	.617	.703	.718	.357
			No	.955	.965	.895	.904	.949	.903	.938	.497
		$\mathcal{N}(1/2,2/3)$	Yes	.805	.821	.781	.700	.752	.759	.743	.363
			No	.971	.981	.963	.947	.967	.941	.954	.576

Table 3.5: DIF Power rates when $N = 900$. Statistics either used information from anchor and focal items only (SIBTEST, CSIBTEST, $sDIF$, $dDIF$, Wald) or were marginalized over the complete test length ($sDIF_M$, $dDIF_M$, Wald_M).

				Compensatory			Non-compensatory				
Anchors	DIF Item	Focal Distribution	Equal Sample Sizes	*sDIF*	$sDIF_M$	SIBTEST	*dDIF*	$dDIF_M$	Wald	Wald_M	CSIBTEST
5	1	$\mathcal{N}(0,1)$	Yes	.169	.202	.437	.530	.813	.877	1.000	.000
			No	.332	.408	.637	.930	.996	.988	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.088	.087	.678	.181	.465	.922	1.000	.000
			No	.158	.212	.896	.753	.981	.991	1.000	.000
	2	$\mathcal{N}(0,1)$	Yes	.188	.187	.555	.648	.913	.872	1.000	.000
			No	.320	.340	.801	.972	.999	.986	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.135	.137	.760	.423	.870	.919	1.000	.000
			No	.230	.266	.966	.957	.999	.998	1.000	.000
	3	$\mathcal{N}(0,1)$	Yes	.814	.813	.714	.775	.937	.869	1.000	.000
			No	.988	.984	.897	.977	.999	.984	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.763	.811	.864	.775	.971	.945	1.000	.000
			No	.976	.984	.983	.990	.999	.997	1.000	.000
	4	$\mathcal{N}(0,1)$	Yes	.887	.942	.804	.765	.924	.877	1.000	.000
			No	.994	.999	.953	.981	.998	.991	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.966	.971	.913	.857	.976	.947	1.000	.000
			No	1.000	.999	.991	.995	.999	.999	1.000	.000
	5	$\mathcal{N}(0,1)$	Yes	.850	.941	.836	.711	.903	.869	.962	.000
			No	.987	.998	.972	.970	.996	.982	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.927	.979	.942	.817	.958	.912	.975	.000
			No	.998	1.000	.995	.984	1.000	.992	.999	.000
10	1	$\mathcal{N}(0,1)$	Yes	.231	.198	.428	.752	.856	.939	1.000	.456
			No	.417	.428	.679	.986	.998	.999	1.000	.664
		$\mathcal{N}(1/2,2/3)$	Yes	.098	.084	.752	.365	.503	.975	1.000	.513
			No	.209	.233	.976	.958	.987	.999	1.000	.583
	2	$\mathcal{N}(0,1)$	Yes	.208	.205	.630	.897	.940	.947	1.000	.509
			No	.344	.363	.860	.997	1.000	.997	1.000	.680
		$\mathcal{N}(1/2,2/3)$	Yes	.157	.162	.823	.771	.908	.973	1.000	.514
			No	.271	.287	.989	.995	.999	1.000	1.000	.621
	3	$\mathcal{N}(0,1)$	Yes	.866	.853	.787	.931	.966	.950	1.000	.545
			No	.991	.991	.958	.999	1.000	.999	1.000	.697
		$\mathcal{N}(1/2,2/3)$	Yes	.859	.821	.939	.960	.985	.985	1.000	.589
			No	.987	.992	1.000	1.000	1.000	1.000	1.000	.668
	4	$\mathcal{N}(0,1)$	Yes	.953	.963	.888	.932	.964	.943	.999	.513
			No	1.000	.999	.986	1.000	1.000	1.000	1.000	.695
		$\mathcal{N}(1/2,2/3)$	Yes	.984	.984	.957	.974	.986	.977	1.000	.618
			No	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.727
	5	$\mathcal{N}(0,1)$	Yes	.965	.972	.913	.920	.958	.947	.970	.539
			No	.999	1.000	.995	.997	.999	.999	.999	.686
		$\mathcal{N}(1/2,2/3)$	Yes	.989	.992	.983	.968	.987	.978	.976	.634
			No	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.738

Table 3.6: DIF Power rates when $N = 1800$. Statistics either used information from anchor and focal items only (SIBTEST, CSIBTEST, *sDIF*, *dDIF*, Wald) or were marginalized over the complete test length ($sDIF_M$, $dDIF_M$, Wald_M).

With respect to the compensatory statistics, SIBTEST was more powerful for DIF Items 1 and 2 than both the marginalized version of *sDIF* ($sDIF_M$) and the version of *sDIF* which did not include the non-focal items. However, for DIF items 4 and 5, the *sDIF* statistics were more compa-

				Compensatory			Non-compensatory				
Anchors	DIF Item	Focal Distribution	Equal Sample Sizes	*sDIF*	$sDIF_M$	SIBTEST	*dDIF*	$dDIF_M$	Wald	Wald_M	CSIBTEST
5	1	$\mathcal{N}(0,1)$	Yes	.244	.296	.575	.797	.976	.969	1.000	.000
			No	.478	.577	.811	.991	1.000	.998	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.129	.135	.856	.407	.847	.989	1.000	.000
			No	.235	.321	.983	.964	.999	1.000	1.000	.000
	2	$\mathcal{N}(0,1)$	Yes	.265	.269	.734	.888	.994	.965	1.000	.000
			No	.468	.474	.932	.997	1.000	1.000	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.187	.200	.918	.822	.989	.994	1.000	.000
			No	.290	.369	.993	.998	1.000	1.000	1.000	.000
	3	$\mathcal{N}(0,1)$	Yes	.943	.947	.851	.945	.993	.971	1.000	.000
			No	.998	1.000	.986	.998	1.000	.998	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.910	.932	.960	.960	.996	.990	1.000	.000
			No	.998	1.000	.999	.999	1.000	.999	1.000	.000
	4	$\mathcal{N}(0,1)$	Yes	.988	.993	.913	.936	.992	.979	1.000	.000
			No	1.000	1.000	.997	.998	1.000	1.000	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.998	.996	.980	.970	.999	.990	1.000	.000
			No	1.000	1.000	.997	1.000	1.000	1.000	1.000	.000
	5	$\mathcal{N}(0,1)$	Yes	.980	.993	.945	.915	.985	.979	.997	.000
			No	.999	1.000	.998	.999	1.000	1.000	1.000	.000
		$\mathcal{N}(1/2,2/3)$	Yes	.987	.998	.982	.964	.995	.985	.997	.000
			No	.999	1.000	1.000	.999	1.000	1.000	1.000	.000
10	1	$\mathcal{N}(0,1)$	Yes	.277	.306	.620	.951	.981	.995	1.000	.598
			No	.539	.619	.835	1.000	1.000	1.000	1.000	.745
		$\mathcal{N}(1/2,2/3)$	Yes	.110	.137	.894	.706	.879	.995	1.000	.591
			No	.290	.329	.998	.999	1.000	1.000	1.000	.667
	2	$\mathcal{N}(0,1)$	Yes	.273	.274	.780	.984	.995	.993	1.000	.605
			No	.493	.494	.968	1.000	1.000	1.000	1.000	.749
		$\mathcal{N}(1/2,2/3)$	Yes	.221	.221	.959	.978	.992	.998	1.000	.667
			No	.352	.393	1.000	1.000	1.000	1.000	1.000	.679
	3	$\mathcal{N}(0,1)$	Yes	.967	.966	.945	.991	.997	.993	1.000	.641
			No	.999	1.000	.996	1.000	1.000	1.000	1.000	.781
		$\mathcal{N}(1/2,2/3)$	Yes	.954	.943	.981	.996	.999	.999	1.000	.710
			No	1.000	.999	1.000	1.000	1.000	1.000	1.000	.760
	4	$\mathcal{N}(0,1)$	Yes	.995	.996	.974	.989	.996	.992	1.000	.641
			No	1.000	1.000	1.000	.999	1.000	.999	1.000	.768
		$\mathcal{N}(1/2,2/3)$	Yes	1.000	.999	1.000	.999	1.000	.999	1.000	.735
			No	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.790
	5	$\mathcal{N}(0,1)$	Yes	.995	.999	.982	.991	.997	.994	.996	.628
			No	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.776
		$\mathcal{N}(1/2,2/3)$	Yes	1.000	1.000	.999	1.000	1.000	1.000	.999	.750
			No	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.833

Table 3.7: DIF Power rates when $N = 2700$. Statistics either used information from anchor and focal items only (SIBTEST, CSIBTEST, $sDIF$, $dDIF$, Wald) or were marginalized over the complete test length ($sDIF_M$, $dDIF_M$, Wald_M).

rable in terms of power, and *sDIF* often displayed higher rejection rates than the SIBTEST procedure. Furthermore, the DRF statistics tended to become more powerful than SIBTEST when group sizes were unequal, and generally acquired power rates closer to 1 more quickly than SIBTEST when ten anchor items were selected. DIF item 3 did not favor any particular compensatory statistic and seems to represent the approximate parameter transition combination when the DRF framework began to outperform SIBTEST. Coupled with the fact that SIBTEST tended to display nominal to liberal detection rates, while the *sDIF* tended to display nominal to conservative rates, it is clear that overall *sDIF* performed better than SIBTEST for items which had little cancellation effects.

The non-compensatory statistics, on the other hand, demonstrated a different trend than the compensatory statistics. CSIBTEST continued to have difficulty obtaining p-values less than the nominal α when only five anchors were used, indicating a severe limitation to CSIBTEST when the number of anchor items selected is too small. Furthermore, both variants of the *dDIF* statistics performed much better than the CSIBTEST statistics when five or ten anchor items were selected regardless of whether the non-focal items were included. This result demonstrates that the DRF family of non-compensatory statistics clearly performed better than CSIBTEST with respect to both Type I error control and statistical power when the same number of anchors were used. As well, contrary to the results presented by Li and Stout (1996), CSIBTEST did not perform better than SIBTEST for detecting DIF for these particular non-uniform DIF items studied.

Finally, the Wald test was the most powerful of all non-compensatory tests investigated, though

$dDIF_M$ did outperform the item subsetted version of the Wald test in many of the conditions studied. Nevertheless it was clear that the Wald test had the most power of all the statistics investigated, especially when non-focal items were included in the model. Combined with the consistent Type I error rates in the previous section the Wald test was clearly the most optimal DIF detection statistic studied in these Monte Carlo simulation conditions. This finding also implies that likelihood-ratio variants for detecting DIF should demonstrate similar properties because the Wald test is asymptotically equivalent to the likelihood ratio test.

3.4.3 Anchor Contamination in the SIBTEST Statistics

The following simulation investigated the effects of using anchor items that contained compensatory and non-compensatory DIF effects with the SIBTEST and CSIBTEST statistics. The purpose was to emulate the 'all non-focal items as anchors' approach that Li and Stout (1996) and Shealy and Stout (1993) recommended, even when DIF is present, to determine whether nominal Type I error rates could be achieved. Using the same five DIF item combinations from the previous section on DIF power rates all non-focal items were included as anchor items. Similar simulation factors were investigated as before, including sample size, whether the group sample sizes were equal, test length (and incidentally the number of anchor items), and whether the latent distributions were identical. Table 3.8 contains the Type I error detection rates for SIBTEST, SIBTEST$_{UC}$ (when the latent trait distributions were equal), and CSIBTEST.

Table 3.8 demonstrates that the Type I error rates for SIBTEST, the uncorrected version of

		$\mathcal{N}(0,1)$			$\mathcal{N}(1/2,2/3)$	
Sample Sizes	Test Length	SIBTEST	SIBTEST$_{UC}$	CSIBTEST	SIBTEST	CSIBTEST
450/450	20	.075	.069	**.087**	**.102**	**.099**
	30	.052	.048	.063	.063	.073
	40	.041	.041	.057	.049	.063
600/300	20	**.106**	**.086**	**.138**	**.152**	**.135**
	30	.070	.064	**.090**	**.093**	**.088**
	40	.053	.050	**.095**	**.079**	**.085**
900/900	20	**.109**	**.087**	**.132**	**.171**	**.137**
	30	**.076**	.071	**.080**	**.090**	**.089**
	40	.062	.058	.073	.061	**.077**
1200/600	20	**.171**	**.132**	**.198**	**.248**	**.200**
	30	**.095**	**.087**	**.118**	**.134**	**.111**
	40	**.078**	.072	**.101**	**.097**	**.094**
1350/1350	20	**.148**	**.118**	**.172**	**.217**	**.172**
	30	**.086**	**.077**	**.094**	**.114**	**.109**
	40	.068	.064	**.076**	.073	**.084**
1800/900	20	**.230**	**.174**	**.263**	**.341**	**.271**
	30	**.116**	**.102**	**.145**	**.181**	**.147**
	40	**.091**	**.085**	**.114**	**.116**	**.110**

Table 3.8: Type I error rates under contamination effects for the SIBTEST procedures when five anchor items contained DIF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

SIBTEST, and CSIBTEST were negatively affected by the inclusion of contaminated anchor items. Although the contamination occurred across a number of the simulation conditions it was largest in the conditions where the total test length was only 20 (which is where the proportion of anchor items that contained DIF was the highest). The error rates rose as high as .341 and .271 for SIBTEST and CSIBTEST, respectively, and became progressively worse as the sample sizes increased. Furthermore, because the uncorrected SIBTEST statistic became inflated when the latent trait distributions were equal, it was clear that anchor contamination did not solely influence the

regression correction for the SIBTEST and instead fundamentally influenced the procedure itself. Finally, the Type I error rates tended to be even more liberal when the latent distributions were not equal.

From this brief simulation study, it is clear that the SIBTEST family of statistics do not always perform well when adopting the 'all non-focal items as anchors' strategy. When including contaminated items in the linking process, the Type I error rates become more akin to the behavior of power rates in that increasing sample size results in higher rejection rates. Therefore, the use of the SIBTEST family of statistics with this anchor selection strategy should be avoided.

The IRT analogue of considering all non-focal items as anchor items was not considered for computational and theoretical reasons. To achieve the same number of anchor items in the IRT frameworks the fitted IRT models would have to be estimated such that all the IRT parameters — excluding those from the focal item — are constrained to be equal across groups. Following the estimation of these models, the DRF framework would then require a parameter covariance matrix to be estimated for each respective model, and the parametric sampling procedure would then be required. Although this approach is certainly possible, and indeed may result in much higher power due to the reduced sampling variability, it was considered too computationally intensive to explore. Furthermore, this anchoring strategy generally does not reflect good practice in the presence of known contamination effects (Millsap, 2011). Use of the likelihood-ratio approach to DIF would follow the same line of reasoning (although computation of $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ for each model would not be required); however, again this approach would result in including contaminated anchor items,

which is a practice that should be avoided.

3.4.4 Summary of DIF Simulations

After examining the differential functioning frameworks across a range of conditions that varied sample size, equal and unequal group sizes, test lengths, latent distributions, number of anchor items, inclusion of non-focal items, and various DIF effects, a few trends are apparent:

- The DFIT framework was unable to obtain Type I error rates close to nominal α rate and was influenced by several design factors. Hence, based on these results the DFIT detection statistics should not be used to detect DIF. Furthermore, the observed $NCDIF$ values do not demonstrate behavior which would be useful for quantifying the magnitude of DIF effects; therefore, use of $NCDIF$ as an effect size is also not recommended. This combination of findings renders the DFIT framework largely unsuitable for detecting or quantifying DIF.

- SIBTEST, $sDIF$, $dDIF$, and the Wald test provided reasonable Type I error control across the conditions studied, including whether non-focal items were fitted (where applicable). Therefore, the simulation results suggest that these statistics are justifiable for detecting DIF effects.

- SIBTEST demonstrated slightly liberal Type I error rates when too few anchor items were selected and the latent trait distributions were unequal. $sDIF$ and $dDIF$ demonstrated slightly conservative Type I error rates when the non-focal items were omitted from the fitted models

and when the sample sizes were smaller.

- Of the compensatory statistics, SIBTEST appeared to be effective for the DIF items 1 and 2 (see Figure 3.2). However, the *sDIF* family of statistics performed better for DIF items 4 and 5. This finding suggests that *sDIF* generally performs better than SIBTEST when DIF is more uniform, while SIBTEST performs better when the DIF is more non-uniform.

- Of the non-compensatory statistics, the Wald test demonstrated the highest power rates and best Type I error control across the conditions studied. Of the three frameworks studied, however, *dDIF* performed the best in terms of Type I error control and DIF detection rates. CSIBTEST appeared to be of no use when only five anchor items were selected and only obtained nominal Type I error rates when 10 or more anchors were selected. Overall, CSIBTEST did not show an improvement in power over the standard SIBTEST procedure.

- With respect to SIBTEST and CSIBTEST, when non-focal items containing DIF were used as anchor items, the Type I error rates increased to unacceptably liberal levels. Therefore, the routine use of using all non-focal items as anchors is generally not recommended.

In addition to these general conclusions from the Monte Carlo simulations, it was apparent that SIBTEST was not consistently the most well performing of the compensatory statistics, even when nominal Type I errors were achieved. In fact, simulation results show that *sDIF* generally outperforms SIBTEST when studying items with little to no cancellation effects. For simpler IRT models which necessarily have no cancellation effect (such as models from the Rasch (1960) family where

only intercepts are fitted), the *sDIF* measures are likely to outperform SIBTEST across a wide variety of empirical settings. For Rasch models in particular, *sDIF* has the additional benefit that it will contain less sampling variability compared to 2PL models because the estimation of the slope parameters is not required; hence, even higher detection rates will occur, further suggesting that *sDIF* will outperform SIBTEST in simpler IRT models. Because the DRF framework explicitly capitalizes on the type of item response models that have been fitted it will have greater detection rates and stability as the sampling variability decreases. This particular property is not shared with SIBTEST, for example, because no model-implied information about the respective items is included in the formulation of the statistics.

The DFIT statistics greatly lacked fidelity for detecting DIF. As well, it also appears to be the case that the observed *NCDIF* values should not be used as effect size measures. Because the *NCDIF* values are highly influenced by sample size and the number of anchors its usefulness as an effect size measure is problematic. Varying sample size and the number of anchor items should generally *not* affect the *NCDIF* values if they are to be used as an effect size measure. For example, in smaller sample sizes it is clear that the likelihood of obtaining a large *NCDIF* value is very high, even when no DIF is present. Therefore, it is difficult to ascertain whether a value of $\widehat{NCDIF} = .1$ is truly 'large' when this value is obtained in smaller samples.

In light of the simulation results presented, it appears that the DFIT framework is of very limited use in both detecting and quantifying DIF in large and small sample sizes. Although the Type I error rates for the *NCDIF* allegedly can be corrected by simulating ad-hoc null hypothesis

distributions (e.g., see Oshima et al., 2006) the problem of interpretation the observed $NCDIF$ values remains because they must be interpreted in concert with other empirical characteristics; namely, sample size, number of anchor items, estimator used to obtain latent trait values, the size of the focal group relative to the reference group, and so on.

Finally, of the three frameworks investigated this simulation study suggested that the DRF framework provided the most consistent Type I error control and demonstrated the greatest potential for detecting compensatory and non-compensatory DIF effects. In situations where it is apparent that non-uniform DIF is present, the $dDIF$ statistics will be of the most effective of the three frameworks at detecting DIF. In situations where items demonstrate uniform DIF in their expected response curves, the $sDIF$ statistics will become useful (and should be more powerful than $dDIF$ for completely uniform DIF because the test has fewer degrees of freedom). Couple this advantage with the ability to investigate DIF at particular integration ranges along $\boldsymbol{\theta}$, the ability to represent variability in the response curves at independent $\boldsymbol{\theta}$ levels graphically, the flexibility to include information about non-focal items, support for any class of IRT model (whether these be for polytomous or dichotomous data), and so on, and it is clear that, of the three frameworks studied, the DRF framework offers the most comprehensive and effective set of tools for studying DIF.

3.5 Differential Bundle and Test Functioning

The statistics investigated in this section pertaining to DBF and DTF were SIBTEST, DTF and DBF from the DFIT framework (again, with the subscripts LF and SF to indicate whether the focal group was larger or smaller than the reference group), and $sDTF$, $sDBF$, $dDTF$, and $dDBF$ from the DRF framework. The CSIBTEST statistic was not studied because it is unable to account for multiple crossing locations that are likely to occur in composite response functions. As was the case with the previous simulation regarding DIF, to ensure that the SIBTEST procedure behaved correctly, and to determine the effect of the regression adjustment, SIBTEST was compared to the unadjusted variant of the statistic and the Type I error behavior was investigated when the DIF items were included in the matched set of items.

The data were again generated from a multiple-group IRT model where each item had a 2PL structure. The slope parameters were drawn from a log-normal distribution, where $\alpha \sim \log \mathcal{N}(0.2, 0.2)$, and the intercept parameters were drawn from a normal distribution, $\delta \sim \mathcal{N}(0, 1/2)$. In the conditions where no DIF was present these parameters were set to be equal across groups. When investigating DBF, the sizes of the focal item bundles were organized into sets of three and five items. DTF was tested by treating all non-anchor items as a complete focal item bundle; therefore, the size of the bundle used to compute the DTF statistics is obtained by subtracting the number of anchor items from the length of the test.

3.5.1 Type I Error Rates

Tables 3.9 and 3.10 contain the Type I error rates for DBF when three and five focal items were investigated, respectively. These tables pertain to the false detection rates when information about the non-focal items was omitted from the fitted models. Appendix C contains the Type I error rates for DTF and DBF when including the non-focal items in the fitted IRT models.

Anchors	Focal Distribution	Sample Sizes	*sDBF*	*dDBF*	SIBTEST	SIBTEST$_{UC}$	DBF_{LF}	DBF_{SF}
5	$\mathcal{N}(0,1)$	450/450	.039	.037	**.076**	.046	**.732**	**.743**
		900/900	.047	.050	.067	.060	**.798**	**.809**
		1350/1350	.045	.060	.072	.049	**.813**	**.812**
		600/300	.040	.040	.073	.059	**.807**	**.764**
		1200/600	.034	.047	.063	.055	**.799**	**.771**
		1800/900	.048	.056	**.077**	.055	**.853**	**.824**
	$\mathcal{N}(1/2,2/3)$	450/450	.025	.040	**.100**	–	**.780**	**.749**
		900/900	.031	.027	**.090**	–	**.809**	**.792**
		1350/1350	.039	.049	**.094**	–	**.835**	**.813**
		600/300	.047	.043	**.085**	–	**.847**	**.765**
		1200/600	.041	.048	**.088**	–	**.883**	**.801**
		1800/900	.053	.049	**.100**	–	**.889**	**.823**
10	$\mathcal{N}(0,1)$	450/450	.046	.055	.066	.063	**.755**	**.753**
		900/900	.038	.054	.046	.037	**.781**	**.789**
		1350/1350	.044	.063	.063	.049	**.807**	**.807**
		600/300	.038	.041	.063	.053	**.831**	**.790**
		1200/600	.038	.052	.053	.050	**.857**	**.837**
		1800/900	.048	.066	.061	.053	**.878**	**.851**
	$\mathcal{N}(1/2,2/3)$	450/450	.034	.035	.070	–	**.819**	**.759**
		900/900	.040	.037	**.079**	–	**.809**	**.811**
		1350/1350	.028	.047	**.077**	–	**.826**	**.824**
		600/300	.043	.043	**.077**	–	**.857**	**.809**
		1200/600	.046	.048	.066	–	**.883**	**.806**
		1800/900	.056	.055	.063	–	**.884**	**.839**

Table 3.9: Type I error rates for DBF testing with three focal items when all non-focal items are omitted from the fitted models. Type I error rates greater than .075 and less than .025 are highlighted in bold.

Beginning with the DBF and DTF statistics from the DFIT framework, it is apparent that

Anchors	Focal Distribution	Sample Sizes	*sDBF*	*dDBF*	SIBTEST	SIBTEST$_{UC}$	DBF_{LF}	DBF_{SF}
5	$\mathcal{N}(0,1)$	450/450	.043	.048	**.079**	.050	**.744**	**.740**
		900/900	.045	.046	.075	.062	**.772**	**.778**
		1350/1350	.038	.048	**.078**	.045	**.805**	**.803**
		600/300	.027	.049	.048	.042	**.753**	**.711**
		1200/600	.061	.062	**.083**	.050	**.814**	**.786**
		1800/900	.050	.058	.074	.043	**.853**	**.826**
	$\mathcal{N}(1/2,2/3)$	450/450	.028	.037	**.114**	–	**.783**	**.754**
		900/900	.049	.048	**.107**	–	**.814**	**.781**
		1350/1350	.041	.041	**.119**	–	**.813**	**.810**
		600/300	.045	.051	**.106**	–	**.855**	**.786**
		1200/600	.056	.043	**.112**	–	**.872**	**.800**
		1800/900	.041	.044	**.104**	–	**.878**	**.807**
10	$\mathcal{N}(0,1)$	450/450	.031	.046	.064	.048	**.757**	**.756**
		900/900	.041	.065	.057	.045	**.770**	**.771**
		1350/1350	.038	.063	.058	.047	**.812**	**.811**
		600/300	.040	.052	.069	.055	**.811**	**.778**
		1200/600	.051	.055	.068	.063	**.873**	**.834**
		1800/900	.045	.062	.060	.048	**.860**	**.835**
	$\mathcal{N}(1/2,2/3)$	450/450	.029	.042	**.076**	–	**.786**	**.756**
		900/900	.031	.039	.067	–	**.818**	**.787**
		1350/1350	.042	.046	.067	–	**.837**	**.782**
		600/300	.033	.043	.070	–	**.824**	**.764**
		1200/600	.045	.043	.060	–	**.838**	**.815**
		1800/900	.052	.059	**.080**	–	**.885**	**.827**

Table 3.10: Type I error rates for DBF testing with five focal items when all non-focal items are omitted from the fitted models. Type I error rates greater than .075 and less than .025 are highlighted in bold.

the Type I error rates were again extremely inflated across all conditions, regardless of whether DTF or DBF was tested. Overall, the empirical Type I error rate for the DTF simulation was .800 (SD = .048) when the focal group size was equal to or larger than the reference group, and .778 (SD = .038) when the focal group was equal to or less than the reference group. As sample size and the number of items increased the Type I error rate also steadily increased. Increasing the number of anchors had the effect of decreasing the error rate, while differences in the latent

distributions caused the Type I error rates to increase. There was a similar pattern of inflation in the DBF simulations when using item bundles of size three and five, and when omitting the non-focal items to compute the statistics. A full discussion of these effects is unnecessary given that the error rates were so poor; therefore, the DTF statistics and DBF are not discussed further.

The SIBTEST framework, on the other hand, had more promising results than the DFIT framework. However, SIBTEST also demonstrated inflated error rates that were primarily influenced by the size of the focal bundle and number of anchor items selected, consequently causing a large majority of rates for falsely detecting compensatory DTF (as well as compensatory DBF, though to a lesser extant) to be unacceptably liberal. Overall, SIBTEST demonstrated a mean Type I error rate of .073 ($SD = .013$) when investigating DBF with three focal items, .078 ($SD = .020$) when investigating DBF with five focal items, and .102 ($SD = .027$) when investigating DTF (with false detection rates reaching as high as .165). The results also suggest that SIBTEST was negatively influenced by the latent trait distributions. When unequal latent traits were combined with five anchors, the false detection rates become more inflated, especially when investigating DTF. Increasing the total number of anchor items generally appeared to help reduce the Type I error rates; however, more anchor items would be required before the rates could reach the nominal α level.

SIBTEST's Type I error rates demonstrated other interesting features as well, especially when compared to the uncorrected variant (SIBTEST_{UC}). As anticipated for the uncorrected SIBTEST, when the latent trait distributions were exactly equal the error rates were close to the nominal α rate regardless of the size of the focal item bundle or number of anchors used (see Tables 3.9

and 3.10, as well as Appendix C). This result suggests that the regression correction procedure for SIBTEST requires a larger amount of linking information from the anchor items when testing item bundles for compensatory DBF and DTF effects. In general, it appears that Shealy and Stout's (1993) statistics for DBF and DTF require a larger matched set of items for the Type I error rates to behave close to the nominal level; otherwise, the regression adjustment results in inflated error rates, even when the latent trait distributions are equal.

Finally, the DRF statistics all demonstrated reasonable Type I error rates regardless of the simulation conditions investigated and size of the focal bundles. When all non-focal items were included in the fitted models the DTF and DBF statistics were slightly conservative when the sample size was only 900. However, error rates became closer to the nominal α level as sample size increased, which is the same observation reported by Chalmers et al. (2016) (see the associated on-line appendix for more specific details). When non-focal items were omitted from the analyses the rates all demonstrated effective Type I error control. Hence, compared to DFIT and SIBTEST only the DRF framework provided sufficient Type I error control when studying DTF and DBF.

3.5.2 Power Rates From DIF Amplification

DTF and DBF share a number of detection properties with DIF in that their rates are all influenced by empirical characteristics such as sample size, test length, number of anchor items, as so on. However, DTF and DBF are also influenced by the magnitude of the individual DIF effects and how these effects propagate at different $\boldsymbol{\theta}$ levels. Generally speaking, the more items which display

DIF, and the larger these DIF effects are, the more likely their effects will propagate within the composite response functions, thereby causing larger DBF and DTF effects.

The power analysis in this section was organized to determine the effect of so-called 'DIF amplification', or the compound effects of DIF which combine to form larger response differences in the composite response function (Shealy & Stout, 1993). Two different sets of DIF effects were included: DIF items 1, 3, and 5 from the previous power analysis section relating to DIF were used when three of the test items contained DIF, and DIF items 1 through 5 were used when five of the test items contained DIF. The composite effects at the respective bundle and test levels can be seen in Figure 3.3. The left images in Figure 3.3 indicate the DTF effect in a 20 item test, while the right graphics demonstrate the respective bundles of size three and five. The expected differences between the test and bundle response functions are in fact identical because the response differences are only caused by the items containing DIF, not by the items which do not contain DIF. For instance, at $\boldsymbol{\theta} = 1$ the expected difference between both response curves is 0.17 (i.e., $14.02 - 13.85 \equiv 2.11 - 1.94$). Therefore, if the IRT parameters were known a priori the DTF and DBF measures from the DRF framework (when the DBF bundle contains all the items with DIF) will provide the exactly same effect size values. However, in practice these estimates will slightly differ because the DTF statistics contain more sampling variability than the DBF measures.

To conserve space, tables containing the simulation results where the non-focal items were included in the fitted models are in Appendix D. Appendix D also contains the results for DTF detection because these rates varied as a function of the length of the tests. DBF rejection rates

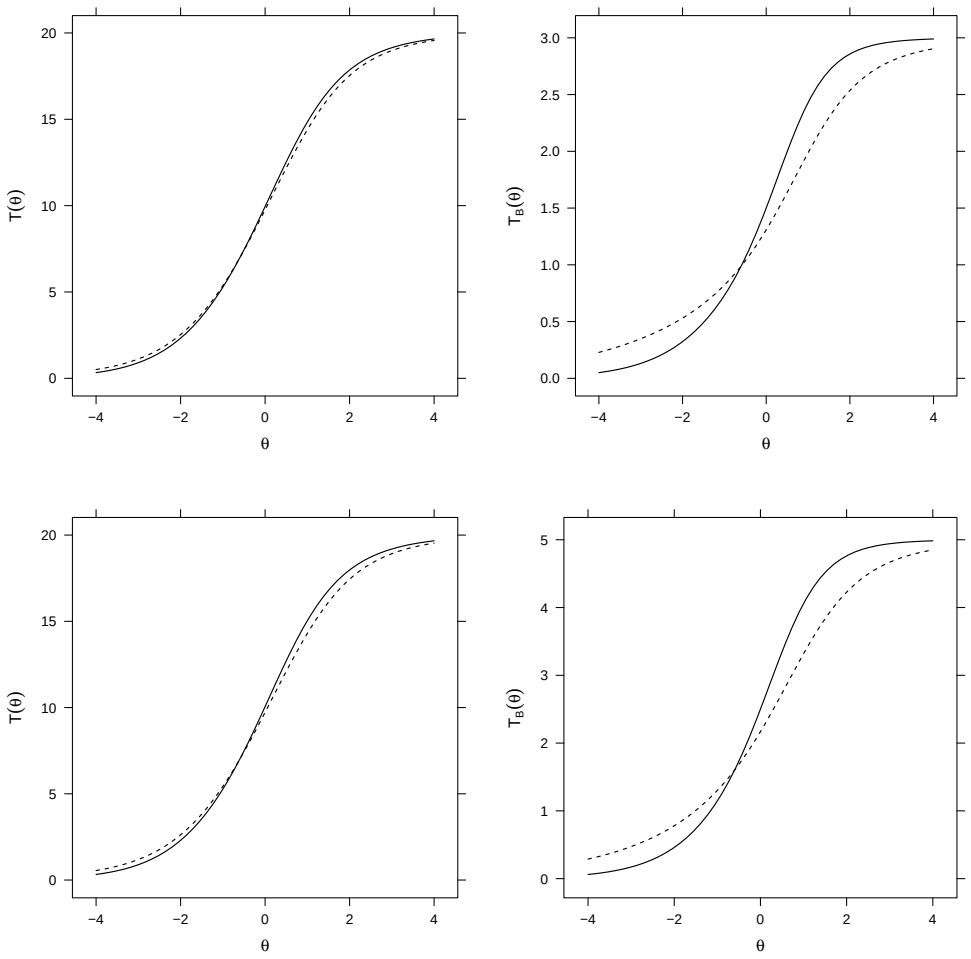

Figure 3.3: DTF and DBF response functions generated from the DIF response curves. Leftmost graphs pertain to the test response function for a 20 item test, while the rightmost graphs contain only the items demonstrating DIF (i.e., a bundle). The function in the top two figures are based on three DIF items while the bottom two figures are based on five DIF items.

when the non-focal items were omitted from the fitted models can be seen in Table 3.11. For convenience, Table 3.11 also includes the marginalized results from the DBF rates in Appendix D to help gauge the effect of including information when modeling non-focal items. Finally, because of the extremely ineffective Type I error control from the DFIT framework the compensatory DTF and DBF statistics were omitted from the following power analysis.

				Three DIF Items					Five DIF Items				
Sample Size	Anchors	Focal Distribution	Equal Group Sizes	*sDBF*	$sDBF_M$	SIBTEST	*dDBF*	$dDBF_M$	*sDBF*	$sDBF_M$	SIBTEST	*dDBF*	$dDBF_M$
900	5	$\mathcal{N}(0,1)$	Yes	.149	.150	.763	.780	.915	.393	.366	.882	.928	.981
			No	.253	.301	.936	.993	1.000	.700	.693	.981	1.000	1.000
		$\mathcal{N}(1/2,2/3)$	Yes	.131	.123	.905	.654	.850	.271	.267	.975	.918	.978
			No	.219	.208	.991	.990	.998	.515	.540	.999	1.000	1.000
	10	$\mathcal{N}(0,1)$	Yes	.175	.157	.843	.950	.973	.464	.462	.964	.998	.997
			No	.349	.321	.983	.999	1.000	.778	.792	.999	1.000	1.000
		$\mathcal{N}(1/2,2/3)$	Yes	.157	.134	.964	.885	.932	.312	.308	.995	.995	.998
			No	.234	.229	1.000	1.000	1.000	.607	.631	1.000	1.000	1.000
1800	5	$\mathcal{N}(0,1)$	Yes	.293	.305	.955	.987	.998	.696	.706	.995	.999	1.000
			No	.514	.563	.998	1.000	1.000	.936	.954	.999	1.000	1.000
		$\mathcal{N}(1/2,2/3)$	Yes	.218	.207	.992	.992	.999	.470	.529	.998	1.000	1.000
			No	.335	.402	1.000	1.000	1.000	.818	.863	1.000	1.000	1.000
	10	$\mathcal{N}(0,1)$	Yes	.328	.336	.989	1.000	1.000	.797	.803	1.000	1.000	1.000
			No	.593	.614	1.000	1.000	1.000	.977	.984	1.000	1.000	1.000
		$\mathcal{N}(1/2,2/3)$	Yes	.240	.239	.999	1.000	1.000	.550	.586	1.000	1.000	1.000
			No	.384	.425	1.000	1.000	1.000	.885	.903	1.000	1.000	1.000
2700	5	$\mathcal{N}(0,1)$	Yes	.404	.444	.992	1.000	1.000	.855	.873	.998	1.000	1.000
			No	.694	.742	1.000	1.000	1.000	.993	.991	1.000	1.000	1.000
		$\mathcal{N}(1/2,2/3)$	Yes	.274	.291	1.000	1.000	1.000	.644	.692	1.000	1.000	1.000
			No	.493	.547	1.000	1.000	1.000	.931	.966	1.000	1.000	1.000
	10	$\mathcal{N}(0,1)$	Yes	.469	.479	.999	1.000	1.000	.927	.925	1.000	1.000	1.000
			No	.767	.783	1.000	1.000	1.000	.998	.998	1.000	1.000	1.000
		$\mathcal{N}(1/2,2/3)$	Yes	.294	.330	1.000	1.000	1.000	.735	.745	1.000	1.000	1.000
			No	.567	.585	1.000	1.000	1.000	.978	.976	1.000	1.000	1.000

Table 3.11: Power rates for DBF testing with three and five focal items. Marginalized rates represented by $sDBF_M$ and $dDBF_M$ were obtained by averaging the detection rates across the total number of items from Appendix C.

As was the case with the prior DIF simulation, a number of systematic trends occurred across all the statistics investigated. For example, increasing the sample size led to higher detection rates, unequal group conditions resulted in higher rejection rates than equal group conditions, models with five DIF items contained more power than models with three DIF items, and increasing the number of anchor items led to higher detection rates. Furthermore, including information about the non-focal items in the fitted models led to higher rejection rates than when these items were omitted; a result that also was observed when studying DIF. Finally, the effect of DIF amplification when investigating more than one item with DIF simultaneously is clear. When only DIF items were included in the bundle of focal items, the power rates were considerably larger than the DTF detection rates as well as the rates based on the individual DIF items found in the previous section.

With respect to the DTF statistics in Appendix D, the power rates for $sDTF$, $dDTF$, and SIBTEST decreased as the length of the test increased. This effect was expected because including items without DIF in the composite response functions will necessarily add additional sampling variability to the respective statistical estimates, thereby making it more difficult to detect true DTF effects (cf., Chalmers et al., 2016). Although the magnitude of the DTF measures in the DRF framework are asymptotically equivalent to the respective DBF counterparts when the focal bundles contain only the items with DIF, the sampling variability will necessarily be smaller when the focal bundle includes only items that contain DIF (i.e., including non-DIF items in the focal bundle only adds additional sampling variability to the test statistics).

Among the compensatory statistics, SIBTEST had the most power to detect DTF and DBF for

the select DIF items compared to the signed statistics from the DRF framework. Unfortunately, however, the inflated Type I error rates complicates the interpretation of the SIBTEST power rates in that the majority of the rates were systematically inflated. Moreover, because SIBTEST was negatively influenced by the size of the focal bundle, the power rates for DTF were especially difficult to interpret. Therefore, there is a trade-off when using SIBTEST for DBF and DTF detection. If liberal Type I error rates can be tolerated, and the size of the focal bundle can be kept as small as possible, then SIBTEST may be a useful tool for detecting DBF effects.

The non-compensatory $dDBF$ statistics, on the other hand, generally demonstrated the most power across the conditions studied. This advantage was especially apparent when non-focal items were included in the fitted models, where the smallest power rate was in the $N = 900$, three focal item conditions. These rates were .840 and .923 when five and ten anchor items were investigated, respectively, generally indicating that detecting non-compensatory DBF was very effective across all conditions. The reason these statistics demonstrated such large power compared to the compensatory statistics was because of the cancellation effects in the composite response functions. While the cancellation in the response functions generally diminished the magnitude of the signed DRF statistics and SIBTEST, the non-compensatory nature of the deviation left the non-compensatory family of the DRF statistics largely unaffected. This is also the reason that the $dDTF$ statistic was the most powerful of all the relevant DTF statistics. Given that the deviation statistics from the DRF framework also demonstrated reasonable Type I error rates, it is clear that this family of detection statistics is optimal for detecting DBF or DTF effects, especially in settings where the

composite response functions contain cancellation effects across the range of the latent trait.

3.5.3 Anchor Contamination in the SIBTEST Procedures

Similar to Section 3.4.3, the following simulation investigated the effect of including contaminated anchor items with compensatory and non-compensatory DIF for SIBTEST when testing for DBF[10]. The purpose of this section was to evaluate whether the 'all non-focal items as anchors' approach that Shealy and Stout (1993) recommended for DIF would also be problematic when studying DBF. Therefore, in the following simulation all non-focal items are included as anchor items and the same five DIF item combinations from the previous section on DIF power rates are included as the contaminated items. The simulation investigated the effects of varying sample sizes, equal or unequal sample sizes, test lengths, focal bundle sizes, and whether the latent trait distributions were equal. Table 3.12 contains the results of this Type I error study.

Similar to the results from the anchor contamination simulation when investigating DIF, SIBTEST became more liberal when contaminated anchors were used. However, DBF Type I error control was much worse than when testing DIF. Type I error rates climbed to extremely liberal levels, as high as .761 when three focal items were investigated, and .949 when five focal items were investigated. Furthermore, the uncorrected SIBTEST procedure was negatively affected by the contaminated anchors because the error rates in the designs where the latent distributions were equal also demonstrated inflated error rates. This effect largely suggests that the contamination was indepen-

[10]Investigating the effect of contaminated anchors for DTF was not relevant because all non-anchor items are, by definition, already included in the focal bundle.

		Three Focal Items			Five Focal Items		
		$\mathcal{N}(0,1)$		$\mathcal{N}(1/2,2/3)$	$\mathcal{N}(0,1)$		$\mathcal{N}(1/2,2/3)$
Sample Sizes	Test Length	SIBTEST	SIBTEST$_{UC}$	SIBTEST	SIBTEST	SIBTEST$_{UC}$	SIBTEST
450/450	20	**.153**	**.121**	**.206**	**.263**	**.179**	**.350**
	30	.069	.067	.067	**.122**	**.102**	**.107**
	40	.064	.057	.044	.066	.066	.047
600/300	20	**.232**	**.170**	**.358**	**.390**	**.266**	**.611**
	30	**.108**	**.082**	**.165**	**.172**	**.146**	**.275**
	40	**.082**	**.077**	**.077**	**.120**	**.112**	**.103**
900/900	20	**.265**	**.203**	**.362**	**.422**	**.276**	**.599**
	30	**.141**	**.119**	**.138**	**.201**	**.155**	**.227**
	40	**.090**	**.084**	.056	**.118**	**.104**	.069
1200/600	20	**.427**	**.316**	**.613**	**.671**	**.501**	**.871**
	30	**.198**	**.157**	**.305**	**.334**	**.269**	**.491**
	40	**.133**	**.114**	**.135**	**.190**	**.171**	**.232**
1350/1350	20	**.368**	**.271**	**.506**	**.620**	**.456**	**.752**
	30	**.185**	**.156**	**.214**	**.293**	**.234**	**.360**
	40	**.128**	**.116**	**.085**	**.190**	**.169**	**.120**
1800/900	20	**.524**	**.387**	**.761**	**.803**	**.648**	**.949**
	30	**.268**	**.225**	**.452**	**.450**	**.376**	**.683**
	40	**.166**	**.152**	**.232**	**.292**	**.249**	**.361**

Table 3.12: Contamination effects for the SIBTEST procedures when five anchor items contained DIF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

dent of the SIBTEST regression correction. Finally, as was the case when investigating DIF with contaminated anchor items it is clear that including all non-focal items as anchors is a suboptimal strategy when using SIBTEST; moreover, this strategy is especially bad when SIBTEST is used for testing bundles of focal items of nearly any size.

3.5.4 Type I Error Rates From DIF Cancellation

This section presents a Monte Carlo simulation to demonstrate how effective the three statistical frameworks were able to achieve consistent Type I error rates in the presence of complete DIF cancellation effects. Conditions under investigation were identical to the conditions in the previous

section on Type I errors when no DIF was present, however one additional condition relating to the number of items containing completely balanced DIF effects (two versus four) was included. The results were expected to be similar to the previous DTF and DBF Type I error rate simulations. Finally, to conserve space the simulation tables which included axillary information from the non-focal items, as well as the results relating to DTF, are available in Appendix E.

The item parameters used to generate the DIF effects were constructed as follows: For the reference group, the slope parameters for the items containing DIF were evenly spaced between 0.5 and 1, while the intercept parameters were evenly spaced between 0.5 and -0.5; the opposite trend was used for the focal group, where slopes were evenly spaced between 1 and 0.5 and the intercepts were evenly spaced between -0.5 and 0.5. The expected probability plots for these DIF effects are in Figures 3.4. DIF items 1 and 4 were used in the simulation where two items contained DIF, and all four DIF items were used in the investigation where four DIF items were included. When the expected test and bundle scoring functions are generated the contribution of the DIF effects completely cancels out; hence, the reference and focal groups generate identical expected response functions across all levels of θ.

The results from Tables 3.13 and 3.14, as well as the tables in Appendix E, indicate that the Type I error rates for the $sDTF$ and $dDTF$ were very similar to the detection rates when there were no items containing DIF. Overall, it was clear that the DRF framework demonstrated nominal to slightly conservative Type I error rates, where the $dDBF$ measures appeared to be the most conservative. The SIBTEST procedure, on the other hand, was again negatively affected by the

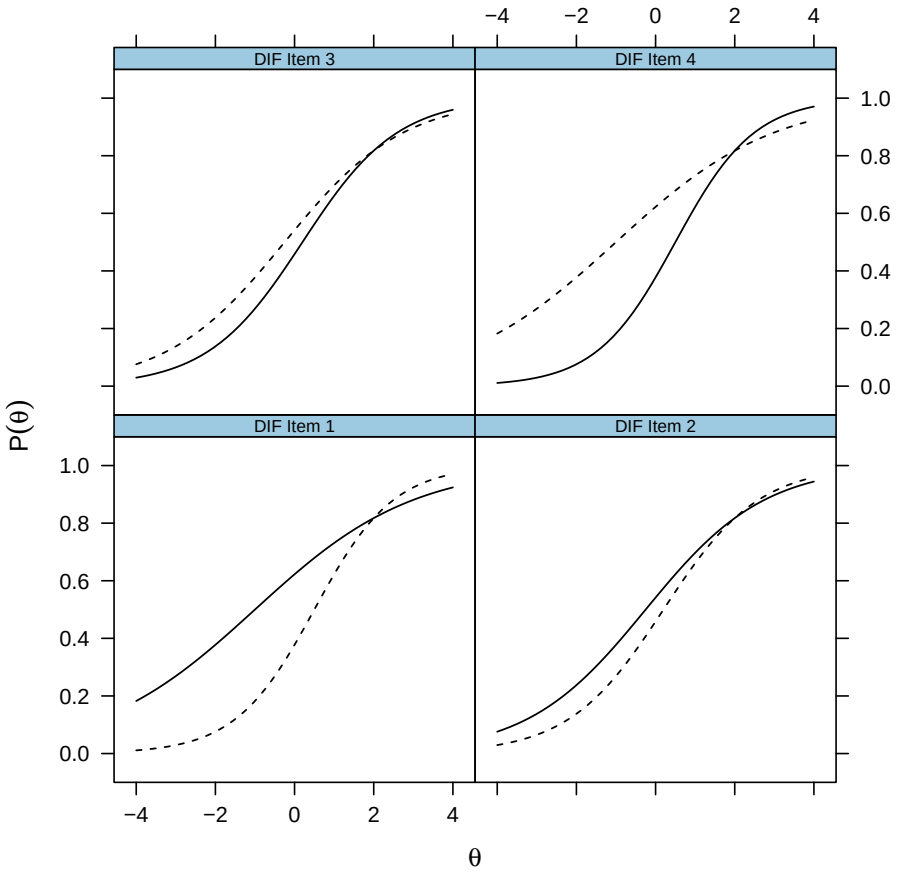

Figure 3.4: Probability response functions in the Monte Carlo cancellation-effect design when the number of items containing DIF is four. Notice the mirroring effect across the response functions, where the item pairs 1-4 and 2-3 are identical but exactly opposite within each group.

size of the focal bundles and demonstrated nearly identical behavior as the previous DTF and DBF Type I error simulation. Therefore, it is clear that both frameworks were generally not substantially affected by the presence of DIF cancellation.

Anchors	Focal Distribution	Sample Sizes	*sDBF*	*dDBF*	SIBTEST	SIBTEST$_{UC}$	DBF_{LF}	DBF_{SF}
5	$\mathcal{N}(0,1)$	450/450	.035	**.020**	.053	.053	**.658**	**.677**
		900/900	.045	**.017**	.040	.037	**.700**	**.710**
		1350/1350	.047	**.023**	.053	.050	**.690**	**.703**
		600/300	.037	.025	.063	.060	**.732**	**.658**
		1200/600	.048	.038	.052	.045	**.743**	**.707**
		1800/900	.047	.027	**.076**	.068	**.725**	**.725**
	$\mathcal{N}(1/2,2/3)$	450/450	.028	.025	.073	–	**.760**	**.715**
		900/900	.032	.035	.075	–	**.777**	**.738**
		1350/1350	.053	**.023**	**.077**	–	**.765**	**.705**
		600/300	.040	**.015**	**.087**	–	**.810**	**.678**
		1200/600	.040	.025	.055	–	**.822**	**.683**
		1800/900	.047	.025	**.083**	–	**.810**	**.727**
10	$\mathcal{N}(0,1)$	450/450	.047	**.020**	.047	.043	**.705**	**.705**
		900/900	.047	**.015**	.038	.042	**.772**	**.730**
		1350/1350	.037	**.022**	.050	.047	**.768**	**.742**
		600/300	.047	**.023**	.062	.052	**.772**	**.733**
		1200/600	.043	.032	.048	.045	**.775**	**.733**
		1800/900	.043	.040	.038	.042	**.773**	**.740**
	$\mathcal{N}(1/2,2/3)$	450/450	.027	**.013**	.057	–	**.788**	**.733**
		900/900	.045	.028	.063	–	**.777**	**.747**
		1350/1350	.048	.037	.063	–	**.830**	**.757**
		600/300	.050	**.015**	.063	–	**.818**	**.728**
		1200/600	.053	**.023**	.058	–	**.807**	**.763**
		1800/900	.040	**.020**	.047	–	**.825**	**.755**

Table 3.13: Type I error rates for DBF testing with two completely balanced focal items containing DIF when all non-focal items were omitted from the fitted models. Type I error rates greater than .075 and less than .025 are highlighted in bold.

The DFIT framework, on the other hand, continued to display extremely inflated Type I error rates across every condition. Furthermore, the DFIT framework was influenced by the number of items demonstrating DIF, suggesting that the cancellation effect was not captured at all. In general,

Anchors	Focal Distribution	Sample Sizes	*sDBF*	*dDBF*	SIBTEST	SIBTEST$_{UC}$	DBF_{LF}	DBF_{SF}
5	$\mathcal{N}(0,1)$	450/450	.035	**.017**	.072	.040	**.655**	**.670**
		900/900	.047	.027	.037	.035	**.652**	**.670**
		1350/1350	.052	.028	.062	.052	**.675**	**.697**
		600/300	.040	**.023**	.068	.048	**.702**	**.665**
		1200/600	.055	.033	.073	.055	**.732**	**.632**
		1800/900	.040	**.017**	.067	.052	**.723**	**.647**
	$\mathcal{N}(1/2,2/3)$	450/450	.053	.028	**.097**	–	**.740**	**.685**
		900/900	.048	.023	**.090**	–	**.785**	**.698**
		1350/1350	.045	.027	**.080**	–	**.760**	**.702**
		600/300	.038	**.018**	.067	–	**.802**	**.630**
		1200/600	.047	.030	.068	–	**.767**	**.663**
		1800/900	.050	.038	**.078**	–	**.805**	**.667**
10	$\mathcal{N}(0,1)$	450/450	.035	**.020**	.045	.042	**.690**	**.695**
		900/900	.030	**.023**	.043	.040	**.702**	**.700**
		1350/1350	.045	.037	.065	.047	**.718**	**.753**
		600/300	.030	.030	.060	.062	**.710**	**.715**
		1200/600	.055	.043	.060	.055	**.723**	**.698**
		1800/900	.052	.038	.045	.045	**.740**	**.723**
	$\mathcal{N}(1/2,2/3)$	450/450	**.023**	**.015**	.052	–	**.725**	**.737**
		900/900	.048	.037	**.078**	–	**.775**	**.733**
		1350/1350	.028	.030	.053	–	**.773**	**.712**
		600/300	.045	**.017**	.062	–	**.825**	**.710**
		1200/600	.047	.037	.063	–	**.828**	**.703**
		1800/900	.038	.033	**.077**	–	**.808**	**.733**

Table 3.14: Type I error rates for DBF testing with four completely balanced focal items containing DIF when all non-focal items were omitted from the fitted models. Type I error rates greater than .075 and less than .025 are highlighted in bold.

when more items contained DIF the tests resulted in higher Type I error rates, and in turn this was complicated by a large number of interaction effects with the other simulation design conditions. Given the large Type I error rates, and the influence of the number of items containing DIF, it is clear that the DFIT framework is not suitable for any type of DTF or DBF detection analyses, even when the DIF items create a complete cancellation effects.

3.5.5 Summary of DTF and DBF Simulations

A few general patterns were observed after examining the differential functioning frameworks across a wide range of conditions, varying sample size, equal and unequal group sizes, test lengths, latent distributions, number of anchor items, use of non-focal items, bundle sizes, and various DIF amplification and cancellation effects:

- The DFIT framework demonstrated extremely liberal Type I error rates, even in the cancellation simulation study, and was influenced by all the simulation conditions under investigation. From these results it appears that the DFIT statistics should not be used for detecting DBF or DTF.

- SIBTEST often demonstrated liberal Type I error rates and was influenced by factors such as the latent distribution, number of anchors, and size of the focal item bundles. In order for SIBTEST to obtain acceptable Type I error rates, more than 10 anchor items are required, especially when investigating larger focal bundles (e.g., DTF).

- The inclusion of contaminated anchor items caused the SIBTEST Type I error control to become unacceptably liberal. Therefore, based on the Monte Carlo results presented above, the general strategy to use all non-focal items as anchors is not recommended. This conclusion is the same conclusion which was reached in previous section on DIF. However, in concert with the inflated error rates caused by the size of the focal bundle, the observed inflation effect appeared to be considerably more severe when testing for DBF and DTF.

- The sign and deviation-based statistics from the DRF framework consistently demonstrated Type I error rates that were at, or slightly below, the nominal α level across all simulation conditions studied. Hence, of the three frameworks investigated only the DRF framework provided a consistent set of tools for detecting DBF and DTF.

- Power rates improved with increasing sample size, groups had unequal sample sizes (compared to equal sample sizes), increasing the number of anchor items, increasing the number of DIF items in the focal bundle, decreasing the number of items without DIF in the focal bundle, and, with respect to the DRF statistics, increasing number of non-focal items in the fitted models.

- The empirical power rates showed that the *dDRF* family of statistics provided the highest power in the item bundles studied, while SIBTEST closely followed. However, in light of its inflated Type I error rates the interpretation of the SIBTEST power rates is more problematic. Of all the compensatory statistics studied, the *sDRF* family demonstrated the lowest power rates for the type of composite response bias generated.

- Type I error rates from the complete cancellation analysis mirrored the error rates in the previous Type I error study, thereby validating the use of SIBTEST and DRF when DIF items display complete cancellation effects. The DFIT framework again failed to provide sufficient Type I error rates, and therefore its use as a detection tool should be discouraged.

Although the results from the DIF Monte Carlo simulations were mixed when comparing SIBTEST

to the DRF framework, the results from the DBF and DTF simulations clearly demonstrate that the DRF framework is superior in terms of Type I error control and power to detect DTF and DBF. Type I error rates were controlled considerably better than SIBTEST and DFIT, the DRF statistics were minimally influenced by the factors investigated, and the respective rejection rates were the highest when studying non-compensatory differential response functioning. As was the case with the DIF analyses, I conjecture that $sDBF$ and $sDTF$ would outperform the SIBTEST framework when the response functions contain little to no cancellation effects. However, because this simulation study did not investigate this effect, future simulations should attempt to test this assertion. Nevertheless, given the control, flexibility, reliability, and post-hoc tools provided by the DRF framework it was evident the DRF family of statistics is the optimal choice for investigating DBF or DTF effects.

4 Further Topics Regarding the Differential Response Functioning Framework

This chapter focuses on extensions and special considerations for the DRF framework. Primarily, the topics addressed relate to the sampling variability of the DRF statistics, the relationship of the DRF measures with effect size measures that have been proposed in the response bias detection literature, and more general extensions of the framework for equivalence testing, multidimensional differential response functioning, and conditional tests for DRF.

4.1 DRF Measures as Effect Size Estimates

Regardless of the statistics chosen for investigating differential functioning, the process of detecting DIF, DBF, or DTF though the use of p-values will rarely inform the investigator about the magnitude or severity of the response bias. Instead, detection methods typically provide evidence about the likelihood that the associated null hypothesis is true, largely as a function of how much empirical information is available. In order to express the magnitude and practical consequences

of the differential effects, an investigator instead must adopt measures that reflect how and where along the θ continuum the differential effects have occurred. In this respect, after detecting DIF, DBF, or DTF the use of *effect sizes* are important because their general goal is to express the magnitude of the response bias in a metric which is practically and substantively meaningful.

To be useful in practice, effect sizes generally have two properties: 1) the measures should generally be unaffected by varying the sample size, and 2) they should be in a metric that is meaningful to the investigator (Kelley & Preacher, 2012). This is one of the reasons that standardized effect sizes have become popular in the social science literature (e.g., see Cohen, 1990). However, the routine use of standardized effect sizes is not required, and in some cases not recommended (Baguley, 2009). In particular, unstandardized effect sizes can be used when the selected measure provides meaningful values to researchers familiar with the subject matter. For example, following a significant p-value in an independent t-test an estimate as simple as the observed mean difference corresponds to one type of unstandardized effect size because 1) it is not generally affected by changes in sample size (the observed difference does not systematically increase or decrease with N) and 2) it is in a meaningful metric to the investigator.

With respect to IRT models, additional considerations are required to express the response bias in a meaningful metric. Because differences between response curves generally vary across different θ levels, even in the very simple Rasch family of IRT models, effect size measures should be either a) expressed in terms which are conditioned on specific values of θ, or b) expressed as marginalized estimates across some desired range of θ. Marginal and conditional effect size

estimates have different strengths and weaknesses for expressing the type and magnitude of bias present. More specifically, marginal estimates attempt to summarize response bias in terms of simple scalar values for each item separately, whereas conditional estimates provide a more fine-tuned level of inspection because they relate to how the bias affects particular $\boldsymbol{\theta}$ locations. However, the conditional estimates may in fact provide too much information, particularly for multidimensional IRT models (see below). Both marginal and conditional effect size applications for the DRF framework, as well as their relationship to existing approaches for quantifying response bias, are discussed in this section.

4.1.1 Comparison of Marginal Effect Sizes

Compared to conditional effect size estimates marginal effect sizes help conceptualize the magnitude of bias in much simpler terms, which may be easier to compare across items, scales, and populations. However, marginal effect size measures are complicated by several additional characteristics and therefore require additional discussion. Each of the measures explored in the Chapter 3 simulations from the three detection frameworks (excluding the W^2 test) can be considered a marginal effect size for differential functioning. The measures are marginal effect size estimates because they collapse all available information (either implicitly or explicitly) across different values of $\boldsymbol{\theta}$ into scalar measures based on compensatory or non-compensatory rules. Therefore, the respective measures found in these three frameworks may serve as useful effect size measures if they are both interpretable and unaffected by sample size.

Unfortunately, as was clear in Chapter 3, the compensatory and non-compensatory measures from the DFIT framework were highly affected by sample size, suggesting that they are not optimal measures for quantifying marginal bias. In fact, a wide variety of ad-hoc IRT-based effect sizes which rely on $\hat{\boldsymbol{\theta}}$ estimates are inherently adversely affected by sample size and testing properties (e.g., see Meade, 2010). This limitation is generally not surprising because the use of $\hat{\boldsymbol{\theta}}$ estimates is known to be influenced by several sampling characteristics (Mislevy, Beaton, Kaplan, & Sheehan, 1992). As was clear in Chapter 3, relying on secondary point estimates for $\hat{\boldsymbol{\theta}}$ is problematic in that the associated statistics may not behave in the way that practitioners expect. Moreover, the statistics based on secondary estimates rarely lend themselves to rigorous statistical theory due to the problem of shrunken or overly variable estimates (Bollen, 1989; Mislevy et al., 1992). While statistics based on $\hat{\boldsymbol{\theta}}$ may be useful for diagnostics within a given sample of individuals (indeed, their relative contributions in the respective sample likely are important; see Meade, 2010), their generalization and interpretation outside the sample from which they are obtained is highly problematic and generally should be avoided (Chalmers et al., 2016).

On the other hand, the SIBTEST and DRF frameworks did behave well according to the simulation results in Chapter 3, therefore it is plausible that these frameworks could be used as effect size measures provided that their metrics are interpretable. Shealy and Stout (1993) note that the observed SIBTEST values (see Equation 2.3) have an approximate relationship to the Mantel-Haenszel (MH) statistic (Mantel & Haenszel, 1959) when testing for DIF (Holland & Thayer, 1988). The MH DIF statistic is a simple log-linear detection test based on examining the interac-

tion between the 2×2 contingency table containing frequencies about the group and item responses after controlling for stratification effects based on the unweighted sum score. Shealy and Stout note that the SIBTEST values are approximately -15 times smaller than the MH statistics; therefore, the classification schemes previously proposed for the MH statistics may be relevant for SIBTEST as well (e.g., Dorans & Holland, 1993).

Considering the DRF framework now, all the defined measures in Chapter 2 have a clear effect size interpretation regardless of whether DIF, DBF, or DTF is under investigation. If the signed statistics are used then these represent the average difference between the expected response functions over some specified range of $\boldsymbol{\theta}$, where cancellation effects are possible if the functions happen to cross at one or more locations. With respect to the unsigned or deviation based statistics, these represent the average area-difference or deviation between the response curves across the desired $\boldsymbol{\theta}$ range. Furthermore, the signed and unsigned measures are always in the metric of the expected item, bundle, or test scores; hence, they are always in a form that is familiar to the test analyst. This property makes the interpretation of the DRF measures as marginal effect size estimates very appealing, especially after significant differential effects have been detected. Given that the DRF measures are also not positively or negatively influenced by different sample sizes it is clear that the DRF measures qualify as useful marginal effect size estimates. Note that the underlying density of the latent trait values are omitted from these area-based statistics; however, research which extends these DRF statistics to include this missing component is already underway (Chalmers, submitted).

4.1.2 Relationship to Area-based Effect Size Measures

The DRF measures are intimately related to the topic which Millsap (2011) referred to as area-based effect size measures. Millsap notes that there are three important considerations when using area-based measures: 1) the range of $\boldsymbol{\theta}$ over which to construct the measures, 2) whether numerical integration or closed-form integration solutions should be adopted, and 3) whether signed or unsigned measures should be formed. Millsap's second consideration primarily relates to issues when computing integrals for IRT models which do not offer a closed-form solution over the complete $\boldsymbol{\theta}$ space (such as the difference between 3PL probability functions when the lower bound parameters are unequal); however, this problem is not relevant when definite integration bounds are utilized.

The DRF framework avoids many problems associated with closed-form integration solutions by using numerical integration across a specified $\boldsymbol{\theta}$ range, which addresses Millsap's (2011) first and second points. The DRF framework also includes different measures for signed and unsigned differential response functioning; therefore, Millsap's third point is also addressed. The metric of the DRF measure's is always interpreted as the *average difference* between the response curves rather than the *observed area* obtained through direct integration (cf. Raju, 1988); hence, the DRF weighted approach has a more natural interpretation (e.g., it is difficult to conceptualize what an absolute area-between-the-curves integral of 1 would look like in a 2PL model, but easier to understand a value of $uDIF = 0.1$ over the range $\boldsymbol{\theta} = [0, 2]$). Finally, the DRF measures are not limited to DIF analyses for dichotomous items (unlike the methods area proposed by Raju, 1988)

and are conceptually equivalent regardless of the size of the focal bundle or use of polytomous items.

The application of a weighting function $g(\boldsymbol{\theta})$ is the only important difference between the observed and average area effect size approaches. That being said, the DRF measures are not unique in their use of a weighting function to construct marginal effect size measures from expected response functions. When focusing only on DIF, the DRF measures are closely related to the *impact* measures proposed by Wainer (1993). Wainer's $T(1)$ and $T(3)$ measures of impact are in fact identical to $sDIF$ and the squared version of $dDIF$. However, the impact measures differ in their application because they use a Gaussian distribution for $g(\boldsymbol{\theta})$ instead of a uniform distribution with weights determined by $\mathcal{N}(\mu_F, \sigma_F^2)$ over the integration range $-\infty$ and ∞.[11] The purpose of using a Gaussian-based weighting function, as Millsap (2011) notes, is because "area measure[s] should focus on regions in which most examinees appear." (p. 217). This is the same reasoning applied by Raju et al. (1995) when constructing the DFIT framework, which is not overly surprising because the DFIT measures are in fact two-step approximations of Wainer's impact measures wherein $\hat{\boldsymbol{\theta}}$ estimates are used instead of the hypothetical $\boldsymbol{\theta}$ values.

[11] Wainer's (1993) $T(2)$ and $T(4)$ measures are not included in this discussion because they include information about the size of the focal group. Therefore, these measures are sample dependent in that they are influenced by increasing sample sizes, and therefore do not qualify as effect sizes in the general sense. Because $T(2)$ and $T(4)$ can be obtained by simply multiplying $T(1)$ and $T(3)$ by N_F, respectively, the discussion of these statistics is largely redundant.

4.1.3 Correcting the Impact Measures Density Function

With respect to the proposed DRF measures, whether to apply $g(\boldsymbol{\theta})$ to create a non-uniform weight function across the latent trait distribution to more heavily weight the areas in which most individuals propagate in the $\boldsymbol{\theta}$ space is an important consideration for constructing marginal effect size measures. However, I do not believe that this approach should be the default when constructing marginal effect sizes for reasons that I elaborate on below. Furthermore, I demonstrate that the current density proposed by Wainer (1993) (and later adopted by Raju et al. (1995) in the DFIT framework) does not effectively achieve the goal of weighting by the expected density of $\boldsymbol{\theta}$. Finally, I close by arguing that the inclusion of an unequally-weighted $g(\boldsymbol{\theta})$ function is prone to several empirical problems and potentially unrealistic assumptions which generally do not arise when $g(\boldsymbol{\theta})$ is given equal weight across the response functions.

The question as to why Wainer (1993) and Raju et al. (1995) set $g(\boldsymbol{\theta})$ to be the density provided by the focal group is the topic that should be investigated first. Allegedly, the purpose of including this particular density function was to weight the differences in the response functions by the underlying density of the latent trait. However, this particular density function only weights the difference by the density of the focal group and entirely ignores the density in the reference group; hence, the measures potentially down-weight important differences which would affect the reference group population at a rate proportional to how far the μ_F parameters are from μ_R (typically, $\mu_R = 0$ by convention). This down-weighting becomes exceptionally problematic if the reference

group has a larger sample size than the focal group because the majority of the true underlying density of $\boldsymbol{\theta}$ is masked — largely negating the purpose of using weights to capture the DIF in the most dense $\boldsymbol{\theta}$ regions. Furthermore, the current form of Wainer's impact measures is not invariant to the selection of the focal and reference groups in that changing which group is considered the focal group can dramatically change the value of the impact measures, despite the model fit being the same. For further discussion on this topic and improvements for the empirical density function based on model-implied elements from the EM algorithm see Chalmers (submitted).

Instead of including only information from the focal group to construct the density function for the latent trait distribution a more appropriate density in the multiple-group IRT model is to combine the expected focal and reference group densities as a mixture distribution proportional to their respective sample sizes (Bock & Zimowski, 1997). Generally, this goal can be achieved by creating a mixture density function

$$g(\boldsymbol{\theta}) = \frac{N_R \cdot g(\boldsymbol{\theta}|\mu_R, \sigma_R^2) + N_F \cdot g(\boldsymbol{\theta}|\mu_F, \sigma_F^2)}{\int \left[N_R \cdot g(\boldsymbol{\theta}|\mu_R, \sigma_R^2) + N_F \cdot g(\boldsymbol{\theta}|\mu_F, \sigma_F^2)\right]}, \tag{4.1}$$

where in practice (4.1) can be evaluated using quadrature nodes in place of $\boldsymbol{\theta}$. This particular density function forms a (potentially bimodal) mixture distribution by combining two Gaussian distributions, which then displays higher density in areas where larger sample sizes are present. Therefore, this weighting scheme achieves the original goal that Wainer (1993) attempted to express.

Equation 4.1 has the added property of being invariant to the selection of the focal group because it appropriately weights the underlying theoretical density in a manner which is proportional

to the sample sizes. If the goal is to weight the differential response function by the underlying latent trait densities, and assuming that both groups are in fact Gaussian distributed, then Equation 4.1 should be adopted instead of simply weighting by a Gaussian distribution based on information only from the focal group. Alternatively, and perhaps more realistically if the assumption of Gaussian distributions is less plausible, $g(\boldsymbol{\theta})$ can be selected such that it reflects the underlying density of $\boldsymbol{\theta}$ by empirically estimating the density from sample information (Bock & Aitkin, 1981; Mislevy, 1984). In this case, these integration-based measures (including those from DRF) become closely related to the metric depicted by the SIBTEST family of statistics; in other words, these statistics will reflect the difference between the response functions weighted by an estimate of the density of $\boldsymbol{\theta}$. However, estimating $g(\boldsymbol{\theta})$ adds additional uncertainty to the process of determining reasonable weights because there are many ways to approximate this density function (e.g., empirical histograms, splines, Monte Carlo sampling, etc), and generally loses the more natural metric compared to using a uniform distribution for $g(\boldsymbol{\theta})$.

4.1.4 Current Limitations of the Impact Measures

Although the above amendments address the original intentions of the impact measures there are still other issues present which make their general use problematic. The impact measures can be viewed as a special case of the DRF measures for DIF when the item responses are dichotomous and there is a need to systematically weight the response function differences based on some type of prior importance. By default, the DRF measures assume that differences in the response

functions are of equal importance across the entire $\boldsymbol{\theta}$ range because an item displaying DIF will ultimately affect *some* individuals, even if these individuals are unlikely to be included in the given sample. This is the general philosophy used in numerous well-established statistical technique used to detect DIF, especially those from which belong to the likelihood family (i.e., the score test, likelihood ratio, and Wald test).

According to the original presentation, Wainer's (1993) impact measures place the greatest importance around the mean of the focal distribution while symmetrically generating less importance around the differences in the response curves as $\boldsymbol{\theta}$ deviates from μ_F. Hence, extremely large response differences in latent trait areas which are less likely (e.g., $\boldsymbol{\theta} = -3$) are viewed as of little consequence to the impact measures, regardless of their magnitude (i.e., compare the weights $g(0) = .5$ and $g(-3) = .001$). These non-uniform weights have benefits in tests that must make accurate inferences at specific $\boldsymbol{\theta}$ ranges, particularly where the inferences around the mean of the focal group are the most important. However, outside this type of application this weighting scheme is generally *not* a reasonable strategy. If the entire $\boldsymbol{\theta}$ distribution is of interest to the test analyst then the use of a non-uniform $g(\boldsymbol{\theta})$ is highly questionable and potentially very misleading.

An alternative view of $g(\boldsymbol{\theta})$, as was previously alluded to, is to treat the integration density as a *weighted measure of importance* rather than as a quasi-correction for the latent trait densities. This type of interpretation has a type of "Bayesian prior" flavor, whereby investigators define $g(\boldsymbol{\theta})$ based on where they believe the difference between the response functions along $\boldsymbol{\theta}$ are most important. For instance, if the test requires accurate inferences to be made between a particular range of $\boldsymbol{\theta}$,

perhaps for classification-based inferences, then applying a suitable weight function which focuses on the area of interest may be justified. The use of a Gaussian density then may not be required or even recommended, where clearly the function certainly has little to do with the distribution of the focal group. Only when the area of interest is coincidentally at the mean of the focal distribution should the weights implied by the impact measures be considered, and only when the dispersion happens to coincide with the variability of Gaussian distribution (with variance equal to σ_F^2) should the focal distribution variability be adopted. From this perspective, the default weights used in the impact measures are quite arbitrary and rarely results in the weighting scheme which test analysts find optimal for their specific applications. Note that a similar interpretation was made for the extra weight function included in Douglas et al.'s (1996) smoothed SIBTEST extension (cf. Equation 8, p. 338).

The $g(\boldsymbol{\theta})$ used in the impact measures is also closely related to restricting the integration range of the DRF statistics by focusing on smaller areas. Restricting the integration range is equivalent to supplying weights of 0 to areas outside the range of interest (Chalmers et al., 2016); therefore, restricting the integration range of $\boldsymbol{\theta}$ achieves nearly the same goal as the impact measures. Hence, interpreting weighted composites can (and in my opinion, should) be avoided in favor of specifying a different integration range of interest. Fortunately, it turns out that the impact measures have the same definition for conditional effect sizes as the DRF framework in that as the integration range tends to 0 at some $\boldsymbol{\theta}$ location then $sDRF_{\boldsymbol{\theta}}$ and the conditional impact measures will become equivalent. Hence, the conditional DRF figures in Chapter 2, along with their associated variability due

to the parametric sampling procedure, apply equally well when inspecting conditional differential effects with the impact measures.

4.1.5 Conditional Effect Sizes for Observed Response Patterns

As indicated in Chalmers et al. (2016) and in Chapter 2, conditional effect size estimates, and their respective sampling variability, are built into the DRF framework because they are evaluated by forming the expected difference in the desired response functions given isolated $\boldsymbol{\theta}$ locations. This approach lends itself naturally to graphical depictions of the conditional bias effect, visually demonstrating the expected point-wise differences between the respective response functions and their associated sampling variability (see Figures 2.2, for example). However, test analysts will often be interested in how these conditional DRF effect size measures relate to particular response patterns in their data so that the response bias can be quantified for each participant.

The presentation of the DRF framework explicitly relates to population level $\boldsymbol{\theta}$ values rather than the predicted $\hat{\boldsymbol{\theta}}$ values for each response pattern. This is because the imprecision borne from the $\hat{\boldsymbol{\theta}}$ estimates is entirely avoided, thereby allowing the respective DRF statistics to achieve more desirable sampling characteristics. However, in order to relate the expected bias given the conditional sampling information from the DRF measures to any given response pattern two forms of sampling variability must be considered: the variability of the expected difference between the response curves at some $\boldsymbol{\theta}$ level (characterized by the family of *sDRF* measures), and the associated sampling variability of the $\hat{\boldsymbol{\theta}}$ estimates.

Considering both sources of sampling information in practice is easy to present as a simple example. Note that although the following example uses DTF to demonstrate how to consider multiple sources of sampling variability the example is equally applicable to DIF and DBF applications. Suppose that a test with 20 dichotomous items is administered to two groups with $N = 500$ participants in each group. After equating the groups via the estimation approach required for the DRF framework (Chapter 2) a conditional DTF plot can be constructed to resemble Figure 4.1. The graphic on the left depicts the expected total scores for both groups and (näively) appears to demonstrate response bias at different $\boldsymbol{\theta}$ levels. The graphic on the right of Figure 4.1, on the other hand, demonstrates the expected differences between the response functions while also including 95% CIs to better account for the sampling variability of the item parameter estimates. As well, in this figure vertical lines are included to demonstrate a specific estimate of $\boldsymbol{\theta}$, as well as its associated CIs, given a specific response pattern under investigation; in this example $\hat{\boldsymbol{\theta}} = 1.25$ with an associated 95% CI of [0.75, 1.75].

Beginning with the left plot in Figure 4.1, if an analyst were to assume that the expected total scores were exactly equal to the population response functions (i.e., $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$) then they could claim the population generating $\boldsymbol{\theta}$ value is affected by the observed response bias. More specifically, because all $\boldsymbol{\theta}$ values within the entire 95% CI range (vertical bars) demonstrate a non-zero $sDRF_{\boldsymbol{\theta}}$ value (solid line on the right graphic) it is likely that, whatever the true $\boldsymbol{\theta}$ value is, the participant is affected by the known response bias. Unfortunately, as is clear from the right image in Figure 4.1, the uncertainty about $\boldsymbol{\theta}$ and the item parameters in concert makes it more difficult to conclude

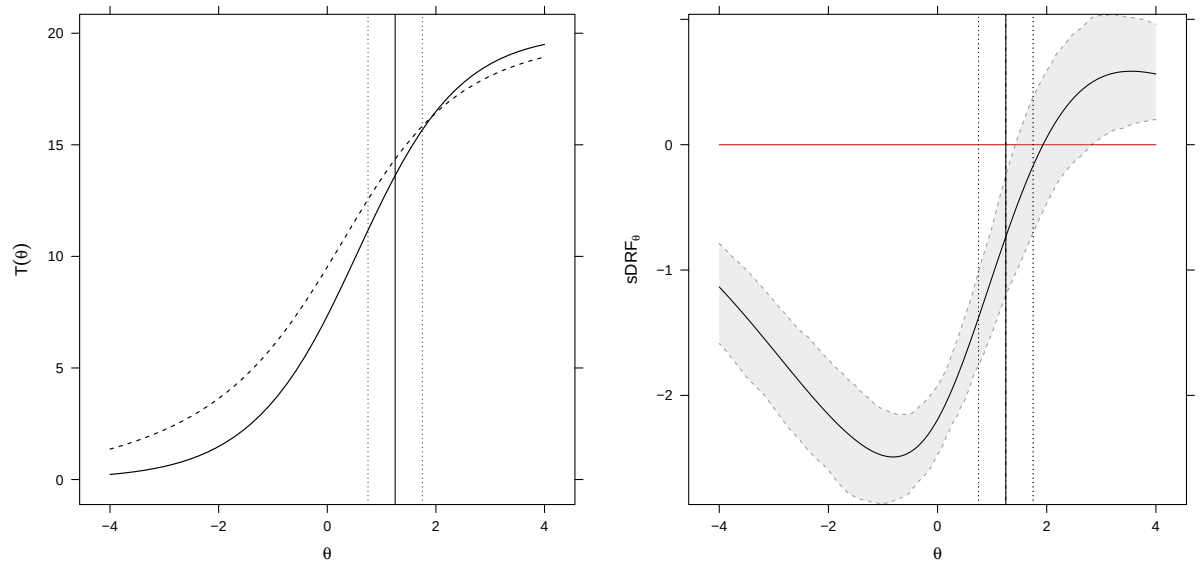

Figure 4.1: Expected total score function (left) and the associated $sDTF_{\boldsymbol{\theta}}$ values with 95% CIs (right) for a hypothetical 20 item test. Vertical lines are included to denote a particular $\hat{\boldsymbol{\theta}}$ estimate (solid) and its respective CIs (dotted).

whether the given individual is influenced by the observed response bias.

After considering both forms of sampling variability, the prediction estimate, as well as the lower bound estimate suggest that the generating $\boldsymbol{\theta}$ value is influenced by the test bias. This is seen visually because the first two vertical lines from the left do not contain either the red horizontal line or the gray 95% CI around the $sDRF_{\boldsymbol{\theta}}$ function. However, at the upper CI of $\boldsymbol{\theta}_{.975} = 1.75$ there is insufficient evidence that bias exists because values within the last two vertical lines include gray areas which overlap with the red horizontal line. Therefore, given the uncertainty in both the item parameter estimates and $\boldsymbol{\theta}$ there is insufficient evidence to conclude that the given response pattern

is influenced by the observed response bias.

To help determine whether a given respondent is affected by the presence of response bias either the measurement precision of the $\hat{\boldsymbol{\theta}}$ estimate, or the precision of the item parameters, can be improved. Improving the measurement precision of $\hat{\boldsymbol{\theta}}$ will only be achieved by administering more items to the given individual, while improving the precision of the item parameter estimates will primarily be improved by obtaining larger sample sizes (see Chapter 3 for other characteristics which influence parameter estimate uncertainty). Continuing again with the visual example in Figure 4.1, the former approach to improving the power to detect bias will result in vertical lines which are closer together, indicating less uncertainty about the location of $\boldsymbol{\theta}$, while the latter approach will result in the gray shaded 95% confidence intervals which are closer to the $sDRF_{\boldsymbol{\theta}}$ function.

4.2 Computational Considerations when Obtaining Sampling Variability

This section discusses specific computational considerations when investigating the statistical variability of the DRF measures. Primarily, use of alternative $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ estimators is discussed, and alternative forms of obtaining sampling variability are reviewed; namely, the bootstrap method (Efron & Tibshirani, 1998) and Markov chain Monte Carlo estimation (Metropolis, Rosenbluth, Teller, & Teller, 1953).

4.2.1 Alternative Estimates of the Parameter Covariance Matrix

In Chapter 3, and in the work by Chalmers et al. (2016), the cross-product method for computing $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ was adopted primarily because this approximation requires very few CPU cycles and RAM. This feature makes the cross-product approximation very appealing for longer tests, in IRT models which contain a large number of parameters, or in samples with a large number of participants. The cross-product estimate of $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ is asymptotically equivalent to the inverse of the observed-data information matrix, and in IRT models generally performs well in larger sample sizes (Paek & Cai, 2014; Pawitan, 2001). However, several other estimators are possible following convergence with the EM algorithm, and a small selection of these estimators were explored in pilot studies for Chapter 3. This section summarizes the pilot studies which adopted alternative $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ estimators.

4.2.1.1 Numerical Approximations

To begin, the use of numerical approximations based on evaluating the observed-data log-likelihood may be useful following the convergence of the multiple-group IRT model. Numerical methods only require the objective function (i.e., observed-data log-likelihood) to be evaluated after slight perturbations have been made to the parameter estimates, thereby allowing finite approximations of the gradient and Hessian terms. Simple numerical methods such as the forward or central difference are possible and generally are not too expensive to compute. However, if more accurate nu-

merical approximations are required then the use of Richardson's (1911) extrapolation will become important. Using numerical approximations following convergence of the EM algorithm is generally considered a good strategy primarily because of the complicated nature of the observed-data log-likelihood, but also because computation of the analytical Hessian is often infeasible (which is why the EM was adopted in the first place; Pawitan, 2001).

Unfortunately, there is a small number of concerns about using numerical derivatives for multiple-group IRT models. First, the numerical approximations may not contain enough precision to support the DRF statistics when a larger number of parameters is modeled. Inaccuracies in the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ elements may propagate in the sampling process required to evaluate the DRF measures, especially when sampling a large number of parameters for the focal bundles (as is required when testing DTF).[12] Second, and perhaps more practically important, accurate numerical approximations may be too computationally demanding due to the difficulty of evaluating the observed-data log-likelihood and the sheer number of elements in $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$. In preliminary simulation results, $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ was obtainable using Richardson extrapolation in the simulations where no focal items were modeled because these contained fewer items and parameters. In the simulations where non-focal items were included, the computations via Richardson's extrapolation were simply infeasible, often taking hours to compute for each model.

In the pilot study simulations where non-focal items were omitted, the detection rates obtained when adopting the Richardson extrapolation were very similar to the cross-product approximation.

[12]In light of the performance with the supplemented-EM algorithm when studying DBF and DTF this concern certainly appears warranted. See below for further details.

However, the DRF Type I error rates were on average slightly less conservative and closer to the nominal α level when using the Richardson extrapolation in comparison to the cross-product approximation in smaller sample sizes. As well, the Wald test for DIF was always within the tolerable Type I error interval, generally indicating that the numerical approximation was well behaved. Therefore, whenever feasible I would recommend using the Richardson extrapolation method (or the analytical Louis (1982) approach, which contains analogous computational issues for larger tests) because the error rates will be slightly closer to the nominal α rate for the DRF statistics. However, when the computations become too burdensome, as they were when non-focal items were included in the simulation, other estimators should be adopted instead.

4.2.1.2 Supplemented EM Algorithm

Another $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ estimator that has become more popular in the IRT literature is the supplemented-EM algorithm (S-EM) because it can be formed using only code from the EM algorithm (Cai, 2008; Meng & Rubin, 1991). The S-EM algorithm uses information from the EM parameter history by applying 'forced' EM updates at different locations along the iteration history to numerically approximate the proportion of missing information. The estimated proportion of missing information is then combined with the complete-data information matrix, which is readily available from the EM algorithm, to remove the effect of the missing data, leaving only the observed information matrix estimate. When inverted, the observed information matrix provides an estimate of the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix which is based upon differentiating the observed-data log-likelihood.

The S-EM algorithm was also investigated in a pilot study for Chapter 3 to determine its overall performance for estimating $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$. When studying DIF, the S-EM algorithm behaved well and returned nominal Type I error rates for the DRF and Wald tests. Hence, the S-EM estimator is a good candidate for investigating DIF effects in empirical work when using the Wald test or DRF measures. However, for the DBF and DTF measures a less optimistic pattern arose. In these scenarios, the DRF statistics had progressively inflated Type I error rates as the size of the focal bundle increased, climbing as high as .15 when detecting DTF in a 40 item test. This result likely occurred because of the imprecision borne from the numerical approximations when building the Jacobian for the missing data information. The error rates improved when more anchors were used, as well as when the sample size increased. However, even for the $N = 2700$ conditions, the error rates were still mainly inflated. The instability of the S-EM algorithm in high-parameter settings, as is the case in the studied multiple-group IRT models, has been previously noted by other researchers (e.g., see Baker, 1992; Segal, Bacchetti, & Jewell, 1994); therefore, this result should not be overly surprising. Finally, the S-EM algorithm was considerably more demanding computationally than the cross-product approach, so again it appears that the cross-product estimator should be the preferred default estimator in larger sample sizes and tests.

In the future, I recommend investigating the method proposed by Oakes' (1999) for multiple-group IRT models because it may provide a more accurate estimate of $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$, and the more computationally efficient forward and central difference approaches should also be investigated to determine their overall performance. When investigating shorter tests with a smaller number of pa-

rameters, the use of accurate numerical approximations via Richardson extrapolation, or the exact method proposed by Louis (1982), should be adopted if the computations are not too demanding. However, based on the simulation results in Chapter 3, as well as the pilot simulation studies, it appears that the cross-product approximation may be considered a reasonable default estimator for $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$, especially if slightly conservative detection rates can be tolerated in smaller sample sizes.

4.2.2 Alternative Forms of Sampling Variability

The parametric sampling procedure previously described in Chapter 2 and by Chalmers et al. (2016) provides a relatively efficient mechanism to generate sampling variability for the DRF measures using only information from the estimated parameter variance-covariance matrix. However, there are other statistical mechanisms to estimate sampling variability of the DRF measures, the most well-known being the bootstrap (Efron & Tibshirani, 1998) and Markov-Chain Monte Carlo (Metropolis et al., 1953) estimation approaches. This section briefly describes how these approaches can be used to obtain the sampling variability for the DRF estimates.

4.2.2.1 Bootstrap Sampling

The bootstrap technique is a general computer-driven approach to obtaining empirical sampling variability. The bootstrap typically obtains sampling variability non-parametrically by re-sampling available data (with replacement) and reanalyzing each of the newly sampled datasets. Hence, the non-parametric bootstrap builds an empirical sampling distribution by generating M indepen-

dently re-sampled datasets and computing the statistics of interest in each respective sample. The justification for this approach is relatively intuitive: because a given sample is assumed to be a random sample from a given population of interest, replacement sampling from a given dataset can be considered a quasi-random sample from the same population (Efron & Gong, 1983).

Although not as common in IRT applications, bootstrapping is nevertheless a viable option for obtaining suitable sampling variability for respective IRT parameter estimates. Given a random sample of response patterns $\mathbf{Y}$, where $\mathbf{Y}$ is an $N \times J$ matrix of responses, we could obtain independent subsamples by randomly selecting N row vectors with replacement from $\mathbf{Y}$ to form a new dataset $\mathbf{Y}^*$. Given $\mathbf{Y}^*$, the IRT model would then be re-estimated to obtain a new set of IRT parameter estimates $\boldsymbol{\Psi}'$, and this set would be stored for later use. Repeating this re-sampling and re-estimating procedure over M independent occasions and collecting the independent $\boldsymbol{\Psi}'$ sets into a complete row-stacked $M \times P$ matrix, $\boldsymbol{\Psi}^*$, allows us to form empirical variability characteristics for each value in the original $\hat{\boldsymbol{\Psi}}$. Given each column in $\boldsymbol{\Psi}^*$, computing the standard deviation of the respective elements provides a computationally-derived standard error estimate of the original $\hat{\boldsymbol{\Psi}}$ estimates. Additionally, empirical confidence intervals can be obtained from $\boldsymbol{\Psi}^*$ by sorting the respective column elements and locating the approximate $\alpha/2$ and $1 - \alpha/2$ values in each vector (Efron & Tibshirani, 1998).

Building and storing the matrix $\boldsymbol{\Psi}^*$ has another benefit with respect to the DRF framework in that the sampling characteristics of each measure of DIF, DBF, and DTF can be obtained directly from the row vectors stored in $\boldsymbol{\Psi}^*$. The parallel between the parametric sampling methodology in

Chapter 2 and the bootstrap approach should now be clear. Instead of drawing $\boldsymbol{\Psi}^*$ from the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix the parameters can be obtained via the non-parametric bootstrap re-sampling methodology. These two methods share the important property that each row in $\boldsymbol{\Psi}^*$ can be obtained independently; hence, forming the full $\boldsymbol{\Psi}^*$ matrix can be efficiently built in pieces after distributing the computations across a number of independent computing resources.

Unfortunately, obtaining $\boldsymbol{\Psi}^*$ using the bootstrap approach has a number of disadvantages compared to the parametric Monte Carlo sampling approach. First, the bootstrap technique takes considerably more CPU and RAM resources to obtain the complete parameter set because M independent IRT models have to be refit to the re-sampled data. The refitting also can, and often will, suffer other estimation-based issues. For instance, when response categories are rarely endorsed it is likely that they will be completely omitted in the new subsample (in which case a new dataset would have to be redrawn), models may fail to converge in the re-sampled datasets and must be fitted again to a different drawn samples (requiring even more CPU cycles), or worse yet the models may converge to local minimum locations, thereby biasing the overall bootstrap estimates. Similar problems can occur when investigating the parametric bootstrap method as well (Hope, 1968).

Nevertheless, bootstrapping has a number of implicit advantages over the parametric sampling method in that it does not require the computation of any $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrices, can be used when the parameter estimates are bounded (where $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ would be less appropriate due to violations of the regularity conditions), is effective when the expected score functions are complex and unpredictable (e.g., Kernel-smoothing IRT models; Mazza, Punzo, & McGuire, 2014), often performs

better than large-sample approximations in smaller samples (Efron & Tibshirani, 1998), and, unlike some estimation approaches to obtain $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ (e.g., estimates based on Louis's (1982) methodology, including the cross-product and sandwich covariances estimates), implicitly supports the use of Bayesian prior parameter distributions when estimating the IRT models. Therefore, bootstrapping may be a preferred option compared to the parametric sampling approach in situations where the computations are not overly demanding or when the use of $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ is not appropriate.

4.2.2.2 Markov Chain Monte Carlo Sampling

Another alternative to obtaining parameter estimates and their associated variability is Markov Chain Monte Carlo (MCMC) estimation (Metropolis et al., 1953). MCMC is a general purpose computer-driven estimation methodology whereby model estimates and their associated sampling variability are formed using sequential Markov Chains based on comparing different probability distribution states, typically using a random walk method. MCMC methodology has been adopted by researchers invested in Bayesian approaches to fitting models where prior information regarding the distribution of parameters is incorporated into the parameter estimation process; however, including prior distributions is not required because MCMC can be used for obtaining ML estimates as well. Recently, MCMC estimation has become a popular estimation method in IRT applications, due in part by the early work of Albert (1992; see also Patz & Junker, 1999a, 1999b).

The simplest and arguably most intuitive of the MCMC algorithms has been termed the Metropolis-Hasting sampler, named after the seminal work by Metropolis et al. (1953) and Hastings (1970).

For ease of exposition I will only present the Metropolis-Hasting sampler for a single scalar parameter; however, the technique generalizes to vectors of parameters as well. The Metropolis-Hasting sampler begins with some parameter value, ψ_0, and some 'proposal' parameter, ψ_1, which has been obtained through a random sampling process (often simply by jittering the value ψ_0 according to some probability sampling distribution). The Metropolis-Hasting sampler then evaluates these two parameters by comparing whether ψ_1 fits the data better than ψ_0 according to the model-implied probability density function. The sampler does this by finding the ratio of the respective posterior distributions (or likelihood distributions, if no priors were defined), and using this ratio determines whether ψ_1 is a more likely value than ψ_0. If ψ_1 provides a better fit to the data then it is selected, stored, and used in the next iteration of the chain; otherwise, it is rejected at a rate of $Q(\psi_1)/Q(\psi_0)$, where Q is the posterior distribution function[13]. After accepting or rejecting ψ_1, the process is repeated using a new proposal value, ψ_2, and again the Metropolis-Hasting sampler evaluates whether this value should be accepted or rejected. This process is repeated many times until a sufficient number of iterations has been completed.

Once the MCMC sampler has been terminated, the history of the parameter values can be used to obtain estimates of the respective parameters through statistical techniques such as finding the respective mean of each freely estimated parameter chain. Furthermore, variability of these parameter estimates in the form of 'posterior standard errors' can be obtained by computing the standard

[13]This rejection ratio is technically only true when the proposal distributions are symmetric, which is the result determined by Metropolis et al. (1953). Hastings (1970) generalized the sampler to support asymmetric proposal distributions.

deviation of the parameter iteration history, and indeed more intricate analyses of the distribution of the parameter estimates are possible because the iteration history provides an estimate of the entire posterior parameter space (Albert, 2009).

It is clear from the description of the Metropolis-Hasting sampler that MCMC has a sequential dependency issue. Because each new proposal set of estimates is evaluated relative to the previously accepted set, the MCMC iteration history has an inherent auto-correlation between each adjacent estimate. Therefore, the MCMC history is often 'thinned' to help remove this auto-correlation effect by selecting a smaller subset of the complete iteration history by selecting samples which are farther apart in the chain (e.g., selecting every 10th estimate in the chain). However, even after thinning has been performed, MCMC diagnostics are still used to determine whether the auto-correlation remains too high for each estimate, and whether the chain has reached a stable equilibrium (Albert, 2009). Convergence can be inspected by plotting the iteration history of each respective estimate, which also helps evaluate whether further iterations are required.

Focusing now on applications of the DRF framework, after estimating a multiple-group IRT model via MCMC the parameter iteration history will, in fact, provide a suitable stand-in for the $\mathbf{\Psi}^*$ matrix. The MCMC iteration history contains all relevant information about sampling variability of each respective estimate. In turn, the MCMC history can be used to represent sampling variability of the DRF measures in a manner which is analogous to the bootstrap and parametric sampling methodologies. MCMC is potentially useful for the DRF framework because $\mathbf{\Psi}^*$ is obtained as a necessary consequence of the MCMC estimation process.

Unfortunately, however, MCMC estimation has a large disadvantage compared to the parametric sampling and bootstrap methods in that the thinned MCMC history often takes considerably longer to obtain than both methods. Furthermore, the bootstrap and parametric sampling approaches naturally lend themselves to completely parallelized computational frameworks, thereby potentially decreasing the amount of time it takes to build $\boldsymbol{\Psi}^*$ by a factor proportional to the number of independent computing cores available. MCMC estimation, on the other hand, typically must be performed in serial on a single core. Given that modern computing resources often have multiple cores, even in most personal computers and laptops, the ability to distribute computations across independent resources is becoming increasingly important to help reduce computational demands of modern statistical methods. On the other hand, MCMC estimation has an advantage over the bootstrap and parametric sampling approaches in that it completely characterizes the posterior distribution of the model parameters (including the posterior for the DRF measures), $\boldsymbol{\Psi}^*$ is available immediately after the chain has converged, and MCMC estimation naturally supports a wide array of prior parameter distributions for including subjective beliefs about the distribution of population parameters.

The MCMC, bootstrap, and parametric sampling approaches for obtaining $\boldsymbol{\Psi}^*$ clearly have their own strengths and weaknesses. On one hand, the parametric sampling approach is typically the most readily available for a number of simpler IRT models; however, it is problematic when a suitable $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ is difficult to compute, or when estimation of the model becomes less accurate (e.g., multidimensional IRT models where numerical integration becomes a concern). The boot-

strap approach, on the other hand, is a non-parametric approach which does not require $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ to be computed. However, the bootstrap is more computationally intensive than the sampling method, is similarly limited to the family of IRT models where the EM algorithm is effective, and is prone to non-convergence and other sampling-based issues. Finally, MCMC has the advantage that $\boldsymbol{\Psi}^*$ is readily available upon completion, is generally very flexible and generalizable, and behaves well for multidimensional IRT models due to the inherent sampling approach for numerical integration. However, the Markov chain itself may be very computationally demanding and generally does not benefit from computational parallelization. Nevertheless, each approach may be useful depending on the research context and datasets sampled; therefore, these methods should be explored in future empirical studies and simulation work.

4.3 General Extensions of the DRF Framework

This section presents three extensions for the DRF framework: the use of conditional tests at different levels of $\boldsymbol{\theta}$, applications in equivalence testing, and extensions for multidimensional IRT models.

4.3.1 Conditional Testing Approach to Detecting Differential Response Functioning

The differential response functioning measures that were examined in Chapter 3 were all based on marginal estimates for detecting response bias, specifically between the $\boldsymbol{\theta}$ range $[-6, 6]$. However, an alternative approach could have been used to detect differential effects which relate to the

individual or conditional $\boldsymbol{\theta}$ components. The conditional approach is what is used to build the respective $sDRF_{\boldsymbol{\theta}}$ measures that previously were manipulated to construct post-hoc graphics, such as those seen in Figure 2.2. However, this approach has the potential to be used for more formal testing of bias in the response functions and, unlike the marginal estimates previously explored, relates to the exact definitions of DIF, DBF, and DTF presented in Chapter 2.

Testing statistical significance of any given $\boldsymbol{\theta}$ location is in fact no different than testing any of the marginal $sDRF$ measures because $sDRF_{\boldsymbol{\theta}}$ is simply the $sDRF$ measures evaluated at some precise $\boldsymbol{\theta}$ location[14]. Therefore, obtaining suitable p-values for a given $\boldsymbol{\theta}_i$ value is no more complex than evaluating any of the $sDRF$ measures. Practitioners may evaluate $sDRF_{\boldsymbol{\theta}}$ over a wide range of $\boldsymbol{\theta}$ values (say, $Q = 1000$ or more) to determine whether any DRF effects are statistically present instead of using visual inspection after constructing plots. However, one of the issues with this approach is that independently evaluating a moderate to large number of $sDRF_{\boldsymbol{\theta}}$ values will potentially lead to inflated family-wise Type I error rates due to repeated significance testing.

The general correction when investigating a large number of statistical tests which supply p-values is to employ some false-discovery rate control mechanism, such as the approach proposed by Benjamini and Hochberg (1995). Unfortunately, however, even these false-discovery techniques may be too conservative for the $sDRF_{\boldsymbol{\theta}}$ tests due to the fact that closely related $\boldsymbol{\theta}$ locations generally contain a large amount of correlated information. This concern may be more apparent with a

[14]To ensure that the $sDRF_{\boldsymbol{\theta}}$ measures have appropriate coverage rates, the simulation conditions in Subsection 3.3.1.2 were evaluated at 21 equally spaced $\boldsymbol{\theta}$ points between -10 and 10. The results are presented in Appendix F. The estimated coverage rates suggest that $sDRF_{\boldsymbol{\theta}}$ provides appropriate coverage when testing DIF, DBF, and DTF across a number of simulation conditions.

simple example: if a value of $\boldsymbol{\theta} = 1$ is significant with $p < .0001$, then adjacent values of $\boldsymbol{\theta} = 1.01$ and $\boldsymbol{\theta} = 0.99$ are likely to have p-values nearly as small because it is unlikely that the response functions have deviated much from the expected values at $\boldsymbol{\theta} = 1$. The issue of correlated p-values also appears to be prevalent in other empirical testing applications based on spatial detection applications (Bennett, Baird, Miller, & Wolford, 2010). Therefore, spatial detection research may offer greater insights into how to deal with this issue of correlated hypothesis tests.

If it were possible to assume that the $\boldsymbol{\theta}$ locations were independent, perhaps by choosing a wide separation between the values of interest, then an alternative approach to investigating all individual p-values for the $sDRF_{\boldsymbol{\theta}}$ test is to form a composite detection statistic whereby all Q conditional $\boldsymbol{\theta}$ tests are combined. This approach is similar to the non-compensatory DRF equations (cf. 1.10), however instead of averaging across the response function (thereby providing an intuitive effect size measure) the individual components are summed to form a more obvious χ^2 variate. The following evaluates the hypothesis $H_0 : \ sDRF_{\boldsymbol{\theta}_1} = sDRF_{\boldsymbol{\theta}_2} = \cdots = sDRF_{\boldsymbol{\theta}_Q} = 0$ using the form

$$X^2_{sDRF} = \sum_{q=1}^{Q} \left(\frac{sDRF_{\boldsymbol{\theta}=\boldsymbol{\theta}_q}}{\hat{\sigma}(\mathbf{m}_{sDRF_{\boldsymbol{\theta}=\boldsymbol{\theta}_q}})} \right)^2, \tag{4.2}$$

where X^2_{sDRF} has a χ^2 distribution with Q degrees of freedom. This hypothesis test would require Q distinct $\boldsymbol{\theta}$ values to be evaluated across the integration range of interest to obtain the Q independent $sDRF_{\boldsymbol{\theta}}$ values. As well, the Q standard error terms required in the denominator can be approximated using one of the three stochastic sampling methods discussed above (where the parametric sampling approach should be the most efficient when the likelihood function is well approximated). However, it is clear that if the $\boldsymbol{\theta}$ values are too close together then $df = Q$ may result in an overly

conservative test due to the amount of correlated information present.

Future work should investigate these conditional hypothesis testing approach for their empirical performance, particularly in comparison to the proposed methods for detecting marginal response bias as well as other methods for testing conditional response bias (e.g., see Moses, Miao, & Dorans, 2010). Furthermore, the anticipated conservative nature of Equation 4.2 when $Q > 1$ should be amended by determining a lower df value based on the amount of correlated information among the $sDRF_{\boldsymbol{\theta}}$ values, though at the present it is not clear what the most optimal strategy to adjusting the df is. Alternatively, ubiquitous corrections may be possible through the use of confidence envelopes (Pek & Chalmers, 2015; Pek, Chalmers, Kok, & Losardo, 2015) or Scheffé corrections (Scheffé, 1959) to all conditional values simultaneously.

4.3.2 Testing for DRF Equivalence

Due to the natural and intuitive interpretation of marginal effect sizes for DRF measures, and because the framework has been organized within a general null hypothesis testing paradigm, the DRF hypothesis testing framework can be modified to investigate group *equivalence* rather than differences. In order to test equivalence, a researcher must first establish a window of tolerance between the response curves, such as allowing for average unsigned difference of .1 over the desired integration area. This tolerance window is used to establish some practically non-significant difference between the response curves, implying that a small amount of bias can be tolerated between the focal and reference group. In what follows, the topic of equivalence testing is introduced using

large sample statistical theory to establish equivalence between the desired response curves when DIF, DBF, or DTF is present but practically inconsequential. Before that, however, the concepts required for equivalence testing are first introduced.

We begin with one of the simplest contexts in which equivalence tests have been popular: testing mean equivalence between two independent samples. In a typical two-sided independent t-test scenario the null hypothesis is expressed as $H_0 : \mu_1 - \mu_2 = D$, where D is typically taken to be 0. The alternative hypothesis which can be concluded when H_0 is rejected is $H_1 : \mu_1 - \mu_2 \neq D$, indicating that the mean difference is not exactly equal to the constant D. The empirical information for the independent t-test information is constructed by forming the ratio

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - D}{s_{\bar{x}_1 - \bar{x}_2}}, \tag{4.3}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of groups 1 and 2, respectively, and $s_{\bar{x}_1 - \bar{x}_2}$ is the pooled standard error (where $N = n_1 + n_2$)

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Following the computation of Equation 4.3, the value T is compared to a hypothetical t-distribution with degrees of freedom $N - 2$ to determine whether the observed ratio provides evidence against the null hypothesis.

The issue with testing against the constant D in Equation 4.3 is that any deviations from D, however minute, will ultimately be detected when the sample sizes are large enough. Hence, even trivial differences will result in the rejection of the null hypothesis, which is generally unfavorable

when researchers are interested in establishing that between-group differences are negligible. To circumvent this issue, while still remaining in a statistical testing framework, equivalence tests have been proposed to invert the null and alternative hypotheses so that increasing the sample size does not bias towards rejecting some exact value in D. This objective is achieved by defining an *equivalence region*, which is established by setting minimum and maximum thresholds on the statistic of interest, and then testing whether the observed results jointly reject two one-sided null hypothesis tests. This is the basic setup for the two one-sided test proposed by Schuirmann (1987).

Schuirmann (1987) argued that to establish equivalence between two means the use of two unidirectional t-tests should be inspected instead. These directional tests correspond to the threshold values by which a meaningful equivalence region can be established. For instance, in the previous example if the mean difference could be tolerated within the values D_U and D_L then the associated null hypotheses are expressed as $H_0 : \mu_1 - \mu_2 \geq D_U$ and $H_0 : \mu_1 - \mu_2 \leq D_L$, where the alternative hypothesis after rejecting both null hypotheses is then $H_1 : D_L < \mu_1 - \mu_2 < D_U$. Hence, rejecting both null hypotheses ultimately leads to the conclusion that the means are within the defined tolerance interval. Notice that to organize an equivalence test there must be some meaningful metric by which equivalence can be expressed; in this case, the tolerance is expressed in the metric of the observed mean difference.

The setup for the independent t-test is akin to how the signed DRF statistics behave in that they test against the constant of 0. However, the DRF tests implement large sample z and χ^2 distributions instead of the sample-size adjusted t and F distributions. The use of the t and F distributions for

the DRF tests is typically only beneficial when the number of parameter samples is small. Given the relative ease of obtaining parametric samples though, the use of 500 or more parameter sets appears to be well approximated by the large sample analogues (see Chapter 3). That said, the reliance of the large sample asymptotic parameter covariance matrix likely limits the statistics to large sample approximations only; hence, even when a small number of parametric samples are obtained these statistics should still be treated as large sample approximations. For these reasons, large sample based equivalence measures are developed below.

Recall from Chapter 2 that for any signed DRF test

$$z = \frac{\widehat{sDRF} - D}{\sqrt{\hat{\sigma}^2(\mathbf{m}_{sDRF})}}, \tag{4.4}$$

where $\widehat{sDRF}$ is used in place of $\widehat{sDIF}$, $\widehat{sDBF}$, or $\widehat{sDTF}$. However, in Equation 2.14 we have incorporated the constant D into the numerator term; previously, D was assumed to be 0. Modifying D to some value other than 0 allows the DRF framework to adopt the two one-sided hypothesis testing strategy, where $H_0 : sDRF \geq D_U$ and $H_0 : sDRF \leq D_L$ with the alternative hypothesis $H_1 : D_L < sDRF < D_U$. Hence, given reasonable upper and lower bounds to indicate the tolerance interval, a large sample test of equivalence is available for all signed-based DRF measures.

The $dDRF$ statistics have a slightly different form than the two one-sided test because this family of statistics is based on the χ^2 distribution. However, because the χ^2 distribution already is a one-sided test, this generalization is also straightforward. Analogous to the generalization for the $sDTF$ to $dDTF$ statistics in Chapter 2, we can use the squared version of (4.4) to form a suitable χ^2 test

$$X^2 = \frac{(\widehat{dDRF} - D)^2}{\hat{\sigma}^2(\mathbf{m}_{dDRF})} \tag{4.5}$$

to achieve the same goal. We now wish to test the null hypothesis $H_0 : dDRF \geq D$ with the associated alternative hypothesis $H_1 : dDRF < D$ to determine whether the population $dDRF$ measures fall within the desired tolerance. Measures based on the $dDRF$ family of statistics are likely more desirable for test analysts because they are based on the overall response differences without the possibility of cancellation effects; however, (4.4) can be useful in response curves which theoretically have no cancellation effects (such as the Rasch model) or when the $\boldsymbol{\theta}$ interval has been modified to include a smaller range.

Finally, the DRF statistics could be adopted within more modern equivalence testing frameworks other than the two one-sided testing methodology; however, the formal development of this area is outside the scope of this dissertation. Nevertheless, it is clear that the measures outlined above provide meaningful and important statistical tools which link equivalence testing methodology to differential response functioning techniques. This link is largely due to the fact that the DRF measures offer effect size interpretations which can be used to establish reasonable tolerance criteria. This topic appears to be an important avenue of research to explore for test analysts who are willing to tolerate certain amounts of differential response effects in their items, bundles, and tests.

4.3.3 Multidimensional Differential Functioning

Throughout this dissertation, the IRT models investigated were assumed to be unidimensional. However, multidimensional IRT (MIRT) models may be more appropriate in a number of empirical testing situations because items may be influenced by more than one latent trait. Due in part to computational and methodological advances, MIRT models have been becoming more popular in applied settings and generally offer a more flexible model fitting methodology compared to unidimensional IRT models (Wirth & Edwards, 2007).

Similar to unidimensional IRT, multidimensional DIF effects can be tested using similar likelihood-based techniques (e.g., Wald, LR, and Lagrange tests). However, due to the complexity and difficulty of estimating multidimensional IRT models the area of multidimensional differential response functioning in general has been highly under-represented in the psychometrics literature (Reckase, 2009). This section provides a brief overview of how the DRF measures can be adopted for multidimensional differential response functioning, although some important caveats about their application will be noted along the way.

The definition of bias for MIRT models is very similar to the definition of bias for unidimensional IRT models (i.e., Equation 1.8) in that it has the form

$$\forall \boldsymbol{\theta} : \; T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_R) = T_B(C|\boldsymbol{\theta}, \boldsymbol{\Psi}_F), \tag{4.6}$$

where changing the bundle size results in the definitions for DIF, DBF, and DTF. The difference between Equation 4.6 and 1.8 is simply that the expected score function is based on a $\boldsymbol{\theta}$ vector

which has more than one element; therefore, the expected scores become multidimensional surfaces rather than simple one-dimensional response curves. See Figure 4.2 for an example of the expected test scoring functions, with respect to each group, and the difference between these functions which is used to compute DTF. Finally, MIRT models have an additional consideration in that interdependencies among the latent traits can also be defined, most commonly assuming that the traits are related via a multivariate Gaussian distribution (Bock et al., 1988).

Multiple-group IRT estimation can again be used to equate the groups after a sufficient number of anchor items has been chosen. However, it is important to note that each respective latent trait must contain a sufficient number of anchor items to uniquely identify the model. Following the estimation of this IRT model, $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ can again be obtained to perform Wald tests; otherwise, LR tests can be obtained by comparing suitable nested models with constrained parameter sets. Because the Wald configuration allows for the possibility to perform DIF testing, the DRF framework may also be used by obtaining parametric sampling set for $\boldsymbol{\Psi}^*$ (alternatively, bootstrapped or MCMC sets may also be obtained; see above). Therefore, the DRF framework follows exactly the same setup for MIRT models as it did for standard IRT models, and properties for conditional differential functions are still applicable (cf. Equation 2.1). However, the conditional and marginal DRF measures relate to joint locations in the $\boldsymbol{\theta}$ space, which may make interpretations slightly more difficult due to the multidimensional nature of the models.

One of the main difficulties in estimating MIRT models occurs because the numerical integration methods required to evaluate the observed-data log-likelihood (or the E-step in the EM

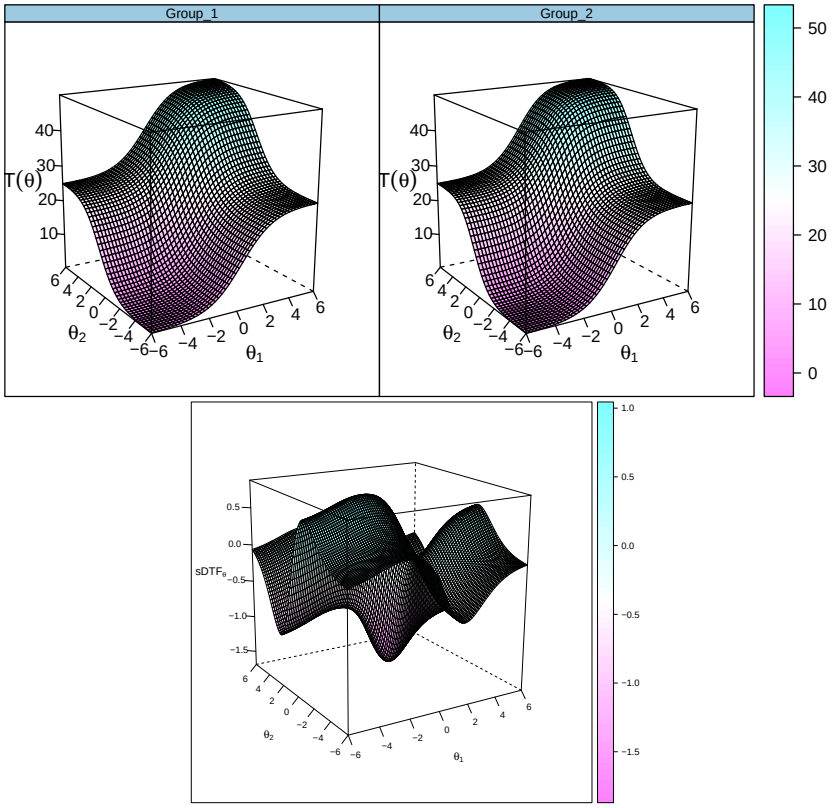

Figure 4.2: Expected test scoring surfaces for a two-dimensional 40 item test (top), generated from a 'complete simple structure' factor loading pattern, and the difference between these functions in the form of the $sDTF_{\boldsymbol{\theta}}$ measure (bottom).

algorithm) become increasingly less tractable as the number of latent traits increases (Bock et al., 1988). Unsurprisingly, the DRF framework shares a similar integration-based issue when evaluating the respective response bias; however, the following examples demonstrate that this integration problem is situation specific. Consider the following pattern of multidimensional slopes in a particular four-dimensional test, where each row corresponds to an item and each column is associated with a unique latent trait:

$$\mathbf{A}_{bifactor} = \begin{pmatrix} \alpha_1 & \alpha_{a1} & 0 & 0 \\ \alpha_2 & \alpha_{a2} & 0 & 0 \\ \alpha_3 & \alpha_{a3} & 0 & 0 \\ \alpha_4 & 0 & \alpha_{b1} & 0 \\ \alpha_5 & 0 & \alpha_{b2} & 0 \\ \alpha_6 & 0 & \alpha_{b3} & 0 \\ \alpha_7 & 0 & 0 & \alpha_{c1} \\ \alpha_8 & 0 & 0 & \alpha_{c2} \\ \alpha_9 & 0 & 0 & \alpha_{c3} \end{pmatrix}.$$

For this particular test structure there are nine items that are influenced by the first latent trait and three additional latent traits which each affect three distinct items. This particular pattern of slopes has been termed a *bifactor model* because each respective item is only affected by two latent traits (Gibbons et al., 2007; Gibbons & Hedeker, 1992). In order to evaluate DIF for any given item using the DRF framework, an integration grid with two dimensions will be required. However, when more than one item is included in the focal bundle (such as when testing for DTF) an integration grid which uses up to four dimensions will be necessary to capture the joint variation contributed by each distinct trait. This type of complex integration often requires special high-dimensional integration techniques such as Monte Carlo or quasi-Monte Carlo integration because

these integrals are generally difficult to compute effectively and efficiently (Caflisch, 1998).

On the other hand, if the test's slope configuration follows what Thurstone (1947) termed a 'complete simple structure' (Gibbons & Hedeker, 1992), whereby each latent trait has only one slope

$$\mathbf{A}_{simple} = \begin{pmatrix} \alpha_{a1} & 0 & 0 \\ \alpha_{a2} & 0 & 0 \\ \alpha_{a3} & 0 & 0 \\ 0 & \alpha_{b1} & 0 \\ 0 & \alpha_{b2} & 0 \\ 0 & \alpha_{b3} & 0 \\ 0 & 0 & \alpha_{c1} \\ 0 & 0 & \alpha_{c2} \\ 0 & 0 & \alpha_{c3} \end{pmatrix},$$

then tests for DIF only require one-dimensional integration. Furthermore, bundle-based DRF tests can also be simplified for the unidimensional item bundles by using isolated one-dimensional integrals. This type of structure may be extremely cost effective for computing the required integrals numerically, and should result in more accurate numerical results than when using multidimensional integration techniques. That being said, there likely is little reason to test DTF in such configurations because a single composite score is generally inappropriate to interpret; therefore, distinct DBF bundles should be explored instead.

Finally, it is likely clear to the reader at this point that including a weighted density function (such as those used in the impact measures of Wainer, 1993) will be even more difficult to interpret. Generally speaking, it is not entirely clear how multidimensional generalizations of Wainer's (1993) impact measures, whereby the focal group generalization provides all the information about

the weighting function, should be interpreted. This confusion arises because the focal group contains a multivariate prior density function focused around the centroid of the focal group as an m-dimensional Gaussian distribution. Interpreting the difference between multidimensional expected surface functions is difficult enough for practitioners to comprehend, and the inclusion of a multidimensional weighting function based on theoretical density distributions again largely obscures the psychological meaning of the response surface differences. As before, however, the inclusion of weights can be used to create regions of ad-hoc importance around particular ranges of $\boldsymbol{\theta}$, and the majority of the properties and amendments discussed in Section 4.1.2 still apply, but the meaningfulness of the resulting measures become even more difficult to interpret. With respect to MIRT models, the default density approach used in the DRF framework likely should be preferred for investigating differential effects, and the use of weights should be introduced only when ad-hoc weighting is desired and justified.

4.4 Summary

This chapter discussed several important areas related to the DRF framework including the interpretation of the DRF measures as effect sizes compared to existing approaches, the generation of different sets of $\boldsymbol{\Psi}^*$ via bootstrapping and MCMC methodology, and provided information about optimally selecting a suitable $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ estimator. Various extensions of the DRF framework were presented relating to conditional detection of DRF given different levels of $\boldsymbol{\theta}$, equivalence testing methods for establishing tolerable levels of bias, and multidimensional differential response

functioning for DIF, DBF, and DTF. This chapter conveyed the generality of the DRF measures, suggesting that there is a number of important applications and extensions of the framework which remain to be explored. The topics reviewed in this chapter, as well as other areas not presented, should be further investigated to determine the overall usefulness of the DRF framework in a much wider variety of contexts than were explored in Chapters 2 and 3.

5 Discussion

This dissertation presented a new detection and quantification methodology termed the differential response functioning (DRF) framework to investigate response bias in items, bundles, and tests. The framework was developed in Chapter 2, and further elaborated and extended in Chapter 4, while Chapter 3 evaluated the detection properties of the new measures to analogous statistics from the SIBTEST (Shealy & Stout, 1993) and DFIT (Raju et al., 1995) frameworks through Monte Carlo simulations. The general conclusion from the Monte Carlo simulations in Chapter 3 was that the DRF framework performed as well or often considerably better than the previously proposed statistical frameworks. Given the arguments and extensions in Chapters 2 and 4, the DRF framework is also able to avoid many of the statistical and conceptual pitfalls present in the SIBTEST and DFIT frameworks. Overall, it was argued that the DRF framework is not only a useful tool for studying DRF, but in fact is conceptually, theoretically, and empirically superior to the two competing frameworks.

Although the sampling variability of the DRF measures can be obtained using bootstrapping or MCMC methodology, the presentation and simulations investigated in this dissertation used

a parametric sampling approach that capitalized on information from the estimated parameter variance-covariance matrix. This particular approach demonstrates a rather attractive feature of the DRF statistics with respect to currently available IRT software; namely, any software package capable of estimating multiple-group IRT models with a suitable $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix can be used as a basis for constructing the DRF measures and their respective sampling variability. This property is beneficial because most of the high-quality software available also include Wald-based tests for DIF which, as discussed in Chapter 2, are the only tools required to setup the parametric sampling method. Because the Wald test and DRF statistics have the same basic setup the results from any software package capable of performing Wald tests for DIF can be used to construct the associated estimates and respective variability of the DRF measures from the exported parameter estimates and $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix.

As a rule of thumb, the DRF measures and their associated parametric sampling variability are valid in situations where the Wald test is theoretically reasonable. This implies that the DRF framework may even be suitable under varying degrees of misspecification because the traditional $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ estimates may be replaced with more robust variance such as the sandwich covariance estimator (White, 1982; Yuan, Cheng, & Patton, 2013). However, when the sampling mechanism used to build a reasonable sample from the posterior distribution of the estimated parameters ($\boldsymbol{\Psi}^*$) is not appropriate, then more appropriate sampling techniques should be adopted instead. In particular, I conjecture that the bootstrap approach will perform better than the parametric sampling approach in datasets with smaller sample sizes, particularly at maintaining nominal Type I error

rates, because the implicit assumption that the likelihood function can be sufficiently approximated by a quadratic function is not required by the bootstrap (this assumption is the basis for using the $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ matrix to form standard errors as well; Chalmers, Pek, & Liu, accepted; Pawitan, 2001). Modifying various IRT models through transformations may help to make $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Psi}}|\mathbf{Y})$ behave better (such as after applying a logit transformation to the lower-bound parameter of the 3PL model). However, if the parameter estimates are too unstable, then this strategy may not completely fix the problem.

In addition to the topics presented in Chapter 4, the DRF measures may also be useful under numerous other IRT models not studied in this body of work, including, but not limited to, polytomous IRT models for items with unordered response categories, non-linear or less-predictable a priori response functions (such as those derived from Kernel-smoothing or spline techniques), models which include latent regression effects (e.g., Adams et al., 1997), and models where test design characteristics are constructed (e.g., Chalmers, 2015). Because the only components required to build the DRF measures are the expected score functions and a reasonable set of plausible parameter estimates to compute the sampling variability, the framework lends itself to a wide variety of IRT modeling applications. This level of generality cannot be overstated because focusing on expected response functions is the reason for using probabilistic response models for categorical data in the first place.

5.1 Conclusion

The DRF framework provides a promising, powerful, and flexible analysis framework for investigating item, bundle, and test bias effects. The area-based DRF measures offer natural interpretations of bias in meaningful metrics which are intimately related to the theoretical definitions of response bias, and these measures can be adopted to a very wide variety of test analysis contexts. The framework promises a number of additional application areas which were not directly studied in this body of work but have important implications for quantifying and controlling test bias. Finally, the measures presented and examined in this dissertation will hopefully be an important contribution to the field of psychometrics, and I hope that future researchers will see the utility and flexibility of the DRF framework in their bias quantification applications.

Bibliography

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *American Educational Research Association*, *17*(3), 251–269.

Albert, J. H. (2009). *Bayesian Computation with R* (2nd ed.). Springer.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617.

Baker, S. G. (1992). A simple method for computing the observed information matrix when using the EM algorithm. *Journal of Computational and Graphical Statistics*, *1*, 63–76.

Beck, A., Steer, R. A., & Brown, O. K. (1996). *Beck Depression Inventory manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*(57),

289-300.

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2010). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction. *Journal of Serendipitous and Unexpected Results*, *1*(1), 1–5.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*(2), 179–197.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer-Verlag.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, *43*(4), 313–333.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometerics*. Cambridge: Cambridge University Press.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*,

31, 144–152.

Caflisch, R. E. (1998). *Monte carlo and quasi-monte carlo methods* (Vol. 7). Cambridge University Press.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309–329. doi: 10.1348/000711007X249603

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*(1), 33–57. doi: 10.1007/S11336–009–9136–X

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement*, *52*(2), 200–222. doi: 10.1111/jedm.12072

Chalmers, R. P. (2016a). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–38. doi: 10.18637/jss.v071.i05

Chalmers, R. P. (2016b). SimDesign: Structure for Organizing Monte Carlo Simulation Designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=SimDesign` (R package version 1.4)

Chalmers, R. P. (in review). Improving the Crossing-SIBTEST statistic for detecting non-uniform

DIF. *Psychometrika*.

Chalmers, R. P. (submitted). Model-based measures for detecting and quantifying response bias. *Psychometrika*.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114–140. doi: 10.1177/0013164415584576

Chalmers, R. P., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement*, *38*(5), 339–358. doi: 10.1177/0146621614520958

Chalmers, R. P., & Ng, V. (in press). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement*.

Chalmers, R. P., Pek, J., & Liu, Y. (accepted). Profile-likelihood confidence intervals in item response theory models. *Multivariate Behavioral Research*.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*(3), 333–353.

Cohen, J. (1990). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*,

39(1), 1–38.

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

Douglas, J. A., Stout, W., & DiBello, L. V. (1996). A kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics*, *21*(4), 333–363. doi: 10.3102/10769986021004333

Efron, B., & Gong, G. (1983). A leisurely look at the boostrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36–48.

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. New York: Chapman & Hall.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

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

Gibbons, R. D., Darrell, R. B., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ... Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255–282.

Hastings, W. K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

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

Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society. Series B (Methodological)*, *30*(3), 582–598.

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics*, *23*(4), 291–322.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate stastistical analysis*. Pearson Prentice Hall.

Kelley, K., & Preacher, K. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647–677.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores*. Reading, MA: Addison-Wesley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society – Series B*, *44*, 226–233.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Maydeu-Olivares, A., Hernández, A., & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research*, *41*(4), 445–471.

Mazza, A., Punzo, A., & McGuire, B. (2014, 6 30). Kernsmoothirt: An r package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*(6), 1–34. Retrieved from `http://www.jstatsoft.org/v58/i06`

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahawah, NJ: Lawerence Erlbaum Associates.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728–743.

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*(416), 899–909.

Metropolis, N., Rosenbluth, A. W., Teller, A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1091.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

Moses, T., Miao, J., & Dorans, N. J. (2010). A comparison of strategies for estimating conditional DIF. *Journal of Educational and Behavioral Statistics*, *35*(6), 726–743.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, *30*(4), 293–311.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *61*(2), 479–482.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of ltems and tests. *Journal of Educational Measurement*, *34*(3), 253–272.

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*(4), 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), 1–17.

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional IRT modeling. *Educational and Psychological Measurement*, *74*, 58–76.

Patz, R. J., & Junker, B. (1999a). Applications and extensions of MCMC in IRT: Multiple item types. missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342–366.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. New York: Oxford University Press.

Pek, J., & Chalmers, R. P. (2015). Diagnosing nonlinearity with confidence envelopes for a semiparametric approach to modeling bivariate nonlinear relations among latent variables. *Structural Equation Modeling*, *22*(2), 288–293.

Pek, J., Chalmers, R. P., Kok, B. E., & Losardo, D. (2015). Visualizing confidence bands for semi-parametrically estimated nonlinear relations among latent variables. *Journal of Educational and Behavioral Statistics*, *40*(4), 402–423.

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

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368.

Ramsay, J. O. (2000). Testgraf: A program for the graphical analysis of multiple choice test and questionnaire data [Computer software manual]. Retrieved from `http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html`

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

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer-Verlag.

Richardson, L. (1911). The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society A*, *210*, 307–357.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons.

Scheffé, H. A. (1959). *The analysis of variance*. New York: Wiley.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657–680.

Segal, M. R., Bacchetti, P., & Jewell, N. P. (1994). Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *Journal of the Royal Statistical Society – Series B*, *56*, 345–352.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194.

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education*, *24*(3), 136–156. doi: 10.1080/10691898.2016.1246953

Stout, W., Li, H.-H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentially two-dimensional. *Applied Psychological Measurement*, *21*(3), 195–213.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Development and applications* (pp. 43–75). New York: Taylor & Francis.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using

the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational Statistics*, *15*(2), 113–128.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Erlbaum.

Wald, A. (1943). Test of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(426–482).

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in IRT scale scores. *Educational and Psychological Measurement*, *72*(2), 264–290.

Yuan, K.-H., Cheng, Y., & Patton, J. (2013). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, *79*(2), 232–254.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128.

Appendix A Type I Error Rates for DIF Simulations

		$\mathcal{N}(0,1)$			$\mathcal{N}(1/2,2/3)$	
Sample Sizes	Test Length	SIBTEST	SIBTEST$_{UC}$	CSIBTEST	SIBTEST	CSIBTEST
450/450	20	.047	.047	.055	.048	.051
	30	.049	.050	.068	.055	.057
	40	.049	.046	.065	.059	.058
600/300	20	.045	.046	.060	.051	.060
	30	.055	.054	.067	.057	.060
	40	.050	.049	.068	.057	.057
900/900	20	.046	.047	.057	.039	.053
	30	.045	.045	.058	.040	.060
	40	.046	.046	.055	.044	.055
1200/600	20	.054	.051	.064	.047	.053
	30	.047	.045	.059	.045	.050
	40	.046	.045	.058	.050	.055
1350/1350	20	.039	.038	.053	.032	.048
	30	.043	.042	.051	.038	.061
	40	.050	.049	.061	.039	.056
1800/900	20	.042	.042	.065	.045	.060
	30	.056	.056	.068	.051	.058
	40	.049	.050	.060	.049	.057

Table 1: Type I error rates for SIBTEST procedures for detecting DIF when all non-focal items are included as anchor items.

Anchors	Focal Distribution	Sample Sizes	Test Length	*sDIF*	*dDIF*	Wald	$NCDIF_{LF}$	$NCDIF_{SF}$	$P(NCDIF_{LF} > .006)$	$P(NCDIF_{SF} > .006)$
5	$\mathcal{N}(0, 1)$	450/450	20	.033	.032	.039	**.767**	**.770**	**.140**	**.137**
			30	.030	**.022**	.036	**.779**	**.779**	**.123**	**.124**
			40	.025	**.020**	.025	**.793**	**.794**	**.132**	**.133**
		600/300	20	.039	.041	.047	**.816**	**.782**	**.032**	**.031**
			30	.037	.040	.039	**.824**	**.788**	**.029**	**.030**
			40	.032	.035	.033	**.831**	**.802**	**.028**	**.026**
	$\mathcal{N}(1/2, 2/3)$	450/450	20	.026	.027	.035	**.793**	**.766**	**.182**	**.139**
			30	**.024**	**.023**	.031	**.799**	**.787**	**.174**	**.132**
			40	**.024**	**.018**	.030	**.809**	**.801**	**.178**	**.138**
		600/300	20	.038	.038	.045	**.852**	**.790**	**.047**	**.038**
			30	.029	.033	.043	**.856**	**.797**	**.043**	**.032**
			40	.031	.030	.037	**.856**	**.803**	**.042**	**.035**
10	$\mathcal{N}(0, 1)$	450/450	20	.031	.031	.044	**.768**	**.764**	**.079**	**.079**
			30	.030	.025	.033	**.775**	**.778**	**.076**	**.076**
			40	**.024**	**.021**	.027	**.775**	**.774**	**.070**	**.069**
		600/300	20	.038	.038	.038	**.819**	**.783**	.010	.010
			30	.033	.039	.040	**.819**	**.786**	.010	.009
			40	.035	.034	.038	**.834**	**.803**	.011	.011
	$\mathcal{N}(1/2, 2/3)$	450/450	20	.026	.026	.045	**.794**	**.767**	**.116**	**.087**
			30	**.022**	**.019**	.040	**.800**	**.782**	**.106**	**.073**
			40	**.019**	**.021**	.030	**.798**	**.794**	**.113**	**.082**
		600/300	20	.040	.036	.041	**.849**	**.780**	**.022**	**.020**
			30	.034	.032	.040	**.857**	**.790**	**.018**	.013
			40	.037	.032	.028	**.852**	**.808**	**.020**	.015

Table 2: Empirical Type I error rates for detecting DIF when $N = 900$ and all items are included in the fitted model. Type I error rates greater than .075 and less than .025 are highlighted in bold.

Anchors	Focal Distribution	Sample Sizes	Test Length	*sDIF*	*dDIF*	Wald	$NCDIF_{LF}$	$NCDIF_{SF}$	$P(NCDIF_{LF} > .006)$	$P(NCDIF_{SF} > .006)$
5	$\mathcal{N}(0, 1)$	900/900	20	.045	.043	.044	**.803**	**.798**	**.023**	**.022**
			30	.035	.039	.037	**.816**	**.812**	**.016**	**.017**
			40	.037	.037	.041	**.827**	**.824**	**.023**	**.023**
		1200/600	20	.046	.050	.047	**.856**	**.823**	**.001**	**.001**
			30	.041	.044	.043	**.863**	**.835**	**.000**	**.000**
			40	.047	.048	.035	**.860**	**.837**	**.001**	**.001**
	$\mathcal{N}(1/2, 2/3)$	900/900	20	.040	.042	.045	**.833**	**.818**	**.045**	**.024**
			30	.039	.037	.048	**.830**	**.817**	**.038**	**.025**
			40	.029	.037	.036	**.837**	**.831**	**.035**	**.019**
		1200/600	20	.043	.042	.045	**.870**	**.826**	**.002**	**.002**
			30	.039	.042	.036	**.880**	**.820**	**.003**	**.001**
			40	.040	.040	.036	**.872**	**.834**	**.003**	**.002**
10	$\mathcal{N}(0, 1)$	900/900	20	.041	.035	.043	**.795**	**.797**	.007	.006
			30	.039	.047	.038	**.810**	**.813**	.007	.006
			40	.038	.040	.037	**.810**	**.814**	.007	.007
		1200/600	20	.043	.046	.049	**.845**	**.817**	**.000**	**.000**
			30	.044	.048	.046	**.849**	**.819**	**.000**	**.000**
			40	.042	.054	.046	**.857**	**.829**	**.000**	**.000**
	$\mathcal{N}(1/2, 2/3)$	900/900	20	.039	.034	.041	**.827**	**.813**	**.018**	.008
			30	.035	.034	.043	**.833**	**.820**	**.018**	.009
			40	.031	.033	.036	**.843**	**.825**	.012	.007
		1200/600	20	.046	.044	.050	**.869**	**.825**	**.001**	**.001**
			30	.041	.038	.045	**.872**	**.834**	**.001**	**.001**
			40	.043	.043	.044	**.877**	**.828**	**.000**	**.000**

Table 3: Empirical Type I error rates for detecting DIF when $N = 1800$ and all items are included in the fitted model. Type I error rates greater than .075 and less than .025 are highlighted in bold.

Anchors	Focal Distribution	Sample Sizes	Test Length	*sDIF*	*dDIF*	Wald	$NCDIF_{LF}$	$NCDIF_{SF}$	$P(NCDIF_{LF} > .006)$	$P(NCDIF_{SF} > .006)$
5	$\mathcal{N}(0, 1)$	1350/1350	20	.046	.049	.054	**.824**	**.826**	**.004**	**.004**
			30	.044	.048	.044	**.838**	**.836**	**.002**	**.002**
			40	.044	.043	.044	**.841**	**.843**	**.003**	**.003**
		1800/900	20	.046	.048	.050	**.863**	**.837**	**.000**	**.000**
			30	.040	.052	.046	**.870**	**.846**	**.000**	**.000**
			40	.048	.053	.050	**.881**	**.855**	**.000**	**.000**
	$\mathcal{N}(1/2, 2/3)$	1350/1350	20	.044	.041	.054	**.849**	**.827**	.010	.005
			30	.037	.037	.047	**.853**	**.840**	.008	**.003**
			40	.040	.040	.046	**.848**	**.845**	.006	**.004**
		1800/900	20	.047	.047	.050	**.882**	**.841**	**.000**	**.001**
			30	.050	.053	.047	**.896**	**.848**	**.001**	**.000**
			40	.046	.047	.039	**.900**	**.855**	**.000**	**.000**
10	$\mathcal{N}(0, 1)$	1350/1350	20	.043	.045	.050	**.827**	**.828**	**.001**	**.001**
			30	.041	.047	.043	**.840**	**.839**	**.001**	**.001**
			40	.035	.042	.035	**.841**	**.840**	**.000**	**.000**
		1800/900	20	.047	.052	.049	**.861**	**.836**	**.000**	**.000**
			30	.046	.050	.047	**.872**	**.845**	**.000**	**.000**
			40	.044	.051	.048	**.869**	**.845**	**.000**	**.000**
	$\mathcal{N}(1/2, 2/3)$	1350/1350	20	.042	.036	.041	**.841**	**.834**	**.003**	**.001**
			30	.041	.044	.044	**.842**	**.838**	.005	**.001**
			40	.042	.039	.040	**.854**	**.835**	**.003**	**.001**
		1800/900	20	.048	.044	.057	**.882**	**.837**	**.000**	**.000**
			30	.048	.053	.052	**.891**	**.836**	**.000**	**.000**
			40	.045	.047	.034	**.889**	**.843**	**.000**	**.000**

Table 4: Empirical Type I error rates for detecting DIF when $N = 2700$ and all items are included in the fitted model. Type I error rates greater than .075 and less than .025 are highlighted in bold.

Appendix B Empirical Power Rates for DIF Simulations

			5 Anchors						10 Anchors					
			Groups Equal			Groups Unequal			Groups Equal			Groups Unequal		
DIF Item	Focal Distribution	Test Length	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald
1	$\mathcal{N}(0,1)$	20	.102	.286	.994	.221	.827	1.000	.108	.316	.998	.215	.864	1.000
		30	.106	.304	.996	.216	.846	1.000	.107	.344	1.000	.202	.882	1.000
		40	.104	.285	.995	.218	.893	.999	.081	.336	1.000	.233	.904	1.000
	$\mathcal{N}(1/2,2/3)$	20	.093	.109	.996	.103	.603	1.000	.061	.092	1.000	.134	.629	1.000
		30	.054	.079	1.000	.098	.629	1.000	.059	.084	1.000	.106	.694	1.000
		40	.057	.086	.998	.110	.614	1.000	.049	.072	1.000	.115	.698	1.000
2	$\mathcal{N}(0,1)$	20	.101	.475	.994	.190	.917	1.000	.112	.511	.999	.197	.932	1.000
		30	.106	.461	.993	.183	.902	1.000	.107	.527	1.000	.180	.952	1.000
		40	.081	.418	.992	.161	.943	.999	.096	.507	1.000	.173	.953	1.000
	$\mathcal{N}(1/2,2/3)$	20	.085	.229	.995	.137	.883	1.000	.102	.290	1.000	.152	.906	1.000
		30	.077	.253	.999	.144	.906	1.000	.107	.312	1.000	.151	.926	1.000
		40	.076	.233	.999	.139	.893	1.000	.074	.277	.999	.160	.935	1.000
3	$\mathcal{N}(0,1)$	20	.485	.546	.985	.811	.937	1.000	.527	.658	.990	.855	.950	1.000
		30	.456	.599	.986	.801	.937	.999	.498	.662	.991	.837	.961	1.000
		40	.375	.521	.980	.796	.941	1.000	.520	.655	.996	.850	.965	1.000
	$\mathcal{N}(1/2,2/3)$	20	.482	.567	.992	.814	.949	1.000	.507	.658	.998	.846	.973	1.000
		30	.473	.583	.996	.798	.956	1.000	.539	.682	.998	.829	.969	1.000
		40	.475	.585	.992	.798	.962	1.000	.475	.680	.997	.824	.978	1.000
4	$\mathcal{N}(0,1)$	20	.630	.569	.932	.938	.921	1.000	.692	.681	.951	.944	.948	.998
		30	.606	.582	.951	.898	.896	.995	.688	.671	.955	.950	.949	1.000
		40	.566	.531	.923	.931	.938	.997	.671	.659	.959	.955	.960	.999
	$\mathcal{N}(1/2,2/3)$	20	.727	.661	.950	.948	.936	1.000	.816	.791	.977	.972	.971	1.000
		30	.699	.669	.964	.960	.956	.999	.806	.778	.971	.984	.981	1.000
		40	.700	.673	.971	.945	.941	1.000	.797	.770	.974	.983	.983	.998
5	$\mathcal{N}(0,1)$	20	.617	.527	.671	.928	.880	.918	.708	.621	.714	.969	.943	.930
		30	.640	.536	.706	.906	.868	.932	.707	.639	.704	.969	.962	.937
		40	.542	.467	.669	.929	.896	.952	.682	.592	.736	.958	.942	.946
	$\mathcal{N}(1/2,2/3)$	20	.738	.620	.766	.961	.935	.940	.819	.740	.760	.983	.972	.953
		30	.729	.641	.742	.948	.912	.940	.824	.764	.722	.976	.958	.956
		40	.717	.625	.742	.960	.937	.937	.819	.751	.748	.985	.970	.954

Table 5: DIF Power rates for $N = 900$.

			5 Anchors						10 Anchors					
			Groups Equal			Groups Unequal			Groups Equal			Groups Unequal		
DIF Item	Focal Distribution	Test Length	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald
1	$\mathcal{N}(0,1)$	20	.214	.787	1.000	.425	.993	1.000	.204	.838	1.000	.409	.995	1.000
		30	.191	.817	1.000	.387	.997	1.000	.195	.857	1.000	.429	.999	1.000
		40	.200	.836	1.000	.413	.999	1.000	.196	.874	1.000	.446	.999	1.000
	$\mathcal{N}(1/2,2/3)$	20	.079	.439	1.000	.204	.971	1.000	.082	.478	1.000	.223	.977	1.000
		30	.084	.496	1.000	.213	.984	1.000	.092	.538	1.000	.209	.990	1.000
		40	.099	.460	1.000	.218	.989	1.000	.079	.494	1.000	.268	.995	1.000
2	$\mathcal{N}(0,1)$	20	.186	.894	1.000	.341	.997	1.000	.200	.923	1.000	.366	1.000	1.000
		30	.186	.910	1.000	.331	1.000	1.000	.209	.942	1.000	.353	1.000	1.000
		40	.188	.934	1.000	.349	1.000	1.000	.206	.954	1.000	.371	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.145	.857	1.000	.273	.999	1.000	.152	.880	1.000	.285	.999	1.000
		30	.139	.886	1.000	.258	.998	1.000	.159	.925	1.000	.293	.999	1.000
		40	.128	.867	1.000	.266	.999	1.000	.176	.918	1.000	.282	1.000	1.000
3	$\mathcal{N}(0,1)$	20	.836	.923	1.000	.986	.999	1.000	.859	.960	1.000	.988	1.000	1.000
		30	.816	.939	1.000	.985	1.000	1.000	.856	.971	1.000	.993	1.000	1.000
		40	.786	.948	1.000	.980	.999	1.000	.844	.966	1.000	.991	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.808	.964	1.000	.980	.997	1.000	.818	.986	1.000	.996	1.000	1.000
		30	.829	.978	1.000	.987	1.000	1.000	.808	.985	1.000	.992	1.000	1.000
		40	.796	.970	1.000	.986	1.000	1.000	.836	.984	1.000	.987	1.000	1.000
4	$\mathcal{N}(0,1)$	20	.940	.912	1.000	.997	.995	1.000	.962	.952	.999	.999	.999	1.000
		30	.936	.924	1.000	1.000	1.000	1.000	.967	.971	1.000	.999	1.000	1.000
		40	.950	.935	1.000	.999	.999	1.000	.961	.969	.999	.998	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.971	.967	1.000	.998	.998	1.000	.985	.983	1.000	1.000	1.000	1.000
		30	.980	.989	1.000	.999	1.000	1.000	.984	.987	1.000	1.000	1.000	1.000
		40	.963	.971	1.000	1.000	1.000	1.000	.984	.989	1.000	1.000	1.000	1.000
5	$\mathcal{N}(0,1)$	20	.938	.898	.963	.997	.995	1.000	.971	.952	.967	1.000	.998	.999
		30	.941	.905	.956	.997	.995	1.000	.970	.959	.968	1.000	.999	.999
		40	.943	.906	.966	.999	.999	1.000	.975	.962	.975	1.000	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.977	.958	.973	1.000	.999	.999	.994	.987	.974	1.000	1.000	1.000
		30	.981	.959	.979	1.000	1.000	.998	.992	.986	.981	1.000	1.000	1.000
		40	.978	.958	.972	1.000	1.000	1.000	.991	.989	.973	1.000	1.000	1.000

Table 6: DIF Power rates for $N = 1800$.

			5 Anchors						10 Anchors					
			Groups Equal			Groups Unequal			Groups Equal			Groups Unequal		
DIF Item	Focal Distribution	Test Length	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald	*sDIF*	*dDIF*	Wald
1	$\mathcal{N}(0,1)$	20	.279	.964	1.000	.578	1.000	1.000	.293	.978	1.000	.621	1.000	1.000
		30	.315	.983	1.000	.581	1.000	1.000	.321	.983	1.000	.612	1.000	1.000
		40	.293	.982	1.000	.573	1.000	1.000	.305	.983	1.000	.624	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.122	.815	1.000	.299	.999	1.000	.137	.862	1.000	.308	1.000	1.000
		30	.146	.862	1.000	.331	1.000	1.000	.142	.880	1.000	.342	1.000	1.000
		40	.136	.864	1.000	.332	.999	1.000	.133	.894	1.000	.337	1.000	1.000
2	$\mathcal{N}(0,1)$	20	.264	.993	1.000	.482	1.000	1.000	.303	.993	1.000	.506	1.000	1.000
		30	.276	.993	1.000	.457	1.000	1.000	.260	.993	1.000	.488	1.000	1.000
		40	.267	.996	1.000	.482	1.000	1.000	.260	.999	1.000	.489	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.219	.988	1.000	.371	1.000	1.000	.234	.989	1.000	.399	1.000	1.000
		30	.191	.993	1.000	.379	1.000	1.000	.213	.989	1.000	.383	1.000	1.000
		40	.189	.987	1.000	.358	1.000	1.000	.217	.997	1.000	.397	1.000	1.000
3	$\mathcal{N}(0,1)$	20	.940	.992	1.000	1.000	1.000	1.000	.960	.994	1.000	1.000	1.000	1.000
		30	.939	.993	1.000	1.000	1.000	1.000	.965	1.000	1.000	1.000	1.000	1.000
		40	.963	.994	1.000	1.000	1.000	1.000	.972	.998	1.000	.999	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.936	.998	1.000	1.000	1.000	1.000	.941	1.000	1.000	.999	1.000	1.000
		30	.932	.994	1.000	1.000	1.000	1.000	.948	.997	1.000	.999	1.000	1.000
		40	.927	.996	1.000	1.000	1.000	1.000	.940	1.000	1.000	.999	1.000	1.000
4	$\mathcal{N}(0,1)$	20	.992	.987	1.000	1.000	1.000	1.000	.995	.996	1.000	1.000	1.000	1.000
		30	.993	.995	1.000	1.000	1.000	1.000	.994	.994	1.000	1.000	1.000	1.000
		40	.994	.994	1.000	1.000	1.000	1.000	.998	.999	1.000	1.000	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.998	1.000	1.000	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000	1.000
		30	.996	.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		40	.995	.999	1.000	1.000	1.000	1.000	.997	1.000	1.000	1.000	1.000	1.000
5	$\mathcal{N}(0,1)$	20	.994	.984	.995	1.000	1.000	1.000	.999	.995	.995	1.000	1.000	1.000
		30	.991	.983	.999	1.000	1.000	1.000	.999	.998	.996	1.000	1.000	1.000
		40	.995	.987	.997	1.000	1.000	1.000	.998	.997	.996	1.000	1.000	1.000
	$\mathcal{N}(1/2,2/3)$	20	.999	.998	.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		30	.999	.994	.999	1.000	1.000	1.000	.999	.999	.998	1.000	1.000	1.000
		40	.996	.993	.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 7: DIF Power rates for $N = 2700$.

Appendix C Type I Error Rates for DBF and DTF Simulations

				Three Focal Items				Five Focal Items			
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DBF_{LF}	DBF_{SF}	*sDBF*	*dDBF*	DBF_{LF}	DBF_{SF}
5	$\mathcal{N}(0,1)$	450/450	20	.039	.041	**.762**	**.765**	.033	.048	**.757**	**.764**
			30	.030	.041	**.776**	**.767**	.031	.045	**.770**	**.768**
			40	.025	.025	**.765**	**.772**	**.023**	.044	**.760**	**.759**
		600/300	20	.034	.041	**.788**	**.818**	.027	.053	**.814**	**.787**
			30	.025	.041	**.792**	**.825**	.027	.045	**.840**	**.795**
			40	.032	.038	**.803**	**.829**	.030	.043	**.834**	**.801**
	$\mathcal{N}(1/2,2/3)$	450/450	20	.035	.039	**.765**	**.756**	.032	.038	**.785**	**.754**
			30	.022	.029	**.790**	**.788**	**.023**	.034	**.782**	**.762**
			40	**.018**	**.023**	**.796**	**.789**	**.014**	**.020**	**.802**	**.784**
		600/300	20	.040	.059	**.793**	**.854**	.039	.052	**.848**	**.789**
			30	.039	.047	**.799**	**.831**	.036	.045	**.836**	**.779**
			40	.034	.044	**.798**	**.854**	.043	.041	**.831**	**.790**
10	$\mathcal{N}(0,1)$	450/450	20	.030	.048	**.750**	**.757**	.030	.046	**.721**	**.728**
			30	.032	.036	**.755**	**.754**	.027	.049	**.770**	**.770**
			40	**.024**	.027	**.763**	**.761**	**.022**	.041	**.742**	**.759**
		600/300	20	.036	.055	**.763**	**.799**	.033	.058	**.814**	**.771**
			30	.036	.045	**.759**	**.803**	.031	.045	**.800**	**.753**
			40	.032	.039	**.779**	**.817**	.035	.054	**.812**	**.768**
	$\mathcal{N}(1/2,2/3)$	450/450	20	.029	.033	**.755**	**.783**	.032	.035	**.782**	**.764**
			30	**.020**	**.018**	**.764**	**.786**	**.017**	.027	**.753**	**.745**
			40	**.016**	**.024**	**.753**	**.779**	**.016**	**.024**	**.781**	**.756**
		600/300	20	.049	.063	**.777**	**.828**	.049	.061	**.817**	**.762**
			30	.045	.052	**.798**	**.853**	.043	.041	**.857**	**.794**
			40	.037	.046	**.774**	**.845**	.034	.048	**.836**	**.778**

Table 8: Empirical Type I error rates for detecting DBF when $N = 900$. Type I error rates greater than .075 and less than .025 are highlighted in bold.

				Three Focal Items				Five Focal Items			
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DBF_{LF}	DBF_{SF}	*sDBF*	*dDBF*	DBF_{LF}	DBF_{SF}
5	$\mathcal{N}(0,1)$	900/900	20	.053	.058	**.826**	**.827**	.060	.066	**.816**	**.813**
			30	.032	.045	**.810**	**.806**	.031	.052	**.790**	**.794**
			40	.039	.042	**.798**	**.807**	.040	.048	**.821**	**.817**
		1200/600	20	.037	.053	**.822**	**.842**	.048	.059	**.849**	**.820**
			30	.043	.050	**.830**	**.862**	.038	.058	**.852**	**.823**
			40	.040	.061	**.823**	**.851**	.044	.066	**.847**	**.815**
	$\mathcal{N}(1/2,2/3)$	900/900	20	.035	.044	**.791**	**.817**	.032	.050	**.801**	**.820**
			30	**.023**	.035	**.817**	**.846**	.026	.041	**.824**	**.806**
			40	.038	.041	**.807**	**.831**	.035	.051	**.845**	**.822**
		1200/600	20	.040	.049	**.815**	**.872**	.043	.043	**.882**	**.788**
			30	.045	.046	**.816**	**.875**	.038	.044	**.870**	**.808**
			40	.035	.042	**.818**	**.888**	.038	.056	**.884**	**.808**
10	$\mathcal{N}(0,1)$	900/900	20	.041	.055	**.793**	**.790**	.040	.052	**.792**	**.794**
			30	.044	.051	**.807**	**.814**	.043	.057	**.805**	**.808**
			40	.029	.046	**.797**	**.802**	.049	.067	**.794**	**.801**
		1200/600	20	.049	.061	**.823**	**.853**	.051	.070	**.848**	**.807**
			30	.032	.063	**.813**	**.838**	.040	.066	**.818**	**.795**
			40	.038	.052	**.809**	**.836**	.047	.062	**.858**	**.831**
	$\mathcal{N}(1/2,2/3)$	900/900	20	.041	.047	**.817**	**.828**	.049	.058	**.811**	**.805**
			30	.029	.049	**.803**	**.828**	.035	.051	**.826**	**.789**
			40	.035	.038	**.809**	**.795**	.034	.040	**.811**	**.777**
		1200/600	20	.042	.055	**.798**	**.880**	.049	.070	**.855**	**.805**
			30	.035	.051	**.808**	**.857**	.030	.059	**.866**	**.798**
			40	.044	.055	**.834**	**.869**	.040	.056	**.856**	**.808**

Table 9: Empirical Type I error rates for detecting DBF when $N = 1800$. Type I error rates greater than .075 and less than .025 are highlighted in bold.

				Three Focal Items				Five Focal Items			
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DBF_{LF}	DBF_{SF}	*sDBF*	*dDBF*	DBF_{LF}	DBF_{SF}
5	$\mathcal{N}(0,1)$	1350/1350	20	.036	.047	**.802**	**.801**	.036	.061	**.781**	**.776**
			30	.036	.056	**.837**	**.831**	.044	.059	**.820**	**.827**
			40	.040	.050	**.827**	**.828**	.046	.043	**.831**	**.838**
		1800/900	20	.048	.057	**.844**	**.874**	.042	.058	**.853**	**.817**
			30	.045	.066	**.836**	**.861**	.041	.061	**.871**	**.846**
			40	.049	.053	**.826**	**.851**	.056	.064	**.870**	**.833**
	$\mathcal{N}(1/2,2/3)$	1350/1350	20	.049	.065	**.815**	**.830**	.047	.059	**.838**	**.810**
			30	.040	.059	**.840**	**.837**	.030	.048	**.850**	**.842**
			40	.030	.042	**.822**	**.846**	.033	.043	**.853**	**.834**
		1800/900	20	.043	.055	**.828**	**.883**	.054	.050	**.869**	**.838**
			30	.038	.062	**.841**	**.901**	.048	.062	**.884**	**.839**
			40	.052	.055	**.852**	**.875**	.047	.053	**.895**	**.842**
10	$\mathcal{N}(0,1)$	1350/1350	20	.046	.043	**.804**	**.803**	.037	.057	**.803**	**.810**
			30	.058	.052	**.805**	**.816**	.044	.055	**.811**	**.811**
			40	.046	.063	**.839**	**.835**	.045	.060	**.853**	**.842**
		1800/900	20	.052	.057	**.829**	**.860**	.054	.072	**.847**	**.817**
			30	.038	.059	**.834**	**.852**	.038	.054	**.850**	**.822**
			40	.045	.072	**.830**	**.853**	.044	.068	**.857**	**.834**
	$\mathcal{N}(1/2,2/3)$	1350/1350	20	.051	.042	**.790**	**.829**	.047	.056	**.805**	**.800**
			30	.043	.052	**.807**	**.810**	.038	.050	**.812**	**.809**
			40	.043	.054	**.821**	**.842**	.048	.054	**.844**	**.824**
		1800/900	20	.051	.055	**.807**	**.877**	.040	.055	**.862**	**.815**
			30	.051	.066	**.826**	**.890**	.047	.053	**.868**	**.822**
			40	.025	.046	**.825**	**.852**	.025	.051	**.875**	**.841**

Table 10: Empirical Type I error rates for detecting DBF when $N = 2700$. Type I error rates greater than .075 and less than .025 are highlighted in bold.

				5 Anchor Items						10 Anchor Items					
Sample Size	Focal Distribution	Sample Sizes	Test Length	$sDTF$	$dDTF$	SIBTEST	SIBTEST$_{UC}$	DTF_{LF}	DTF_{SF}	$sDTF$	$dDTF$	SIBTEST	SIBTEST$_{UC}$	DTF_{LF}	DTF_{SF}
900	$\mathcal{N}(0,1)$	450/450	20	.030	.045	**.092**	.052	**.747**	**.754**	.034	.057	.050	.047	**.675**	**.667**
			30	.025	.045	**.102**	.059	**.759**	**.764**	.029	.051	.074	.055	**.731**	**.729**
			40	**.022**	.032	**.111**	.058	**.779**	**.775**	.025	.031	**.076**	.043	**.710**	**.715**
		600/300	20	.041	.055	**.091**	.047	**.811**	**.775**	.042	.063	.061	.053	**.759**	**.725**
			30	.034	.049	**.123**	.050	**.820**	**.786**	.039	.066	**.082**	.054	**.757**	**.714**
			40	.038	.045	**.106**	.055	**.854**	**.827**	.031	.047	.075	.046	**.774**	**.737**
	$\mathcal{N}(1/2,2/3)$	450/450	20	.027	.040	**.132**	–	**.760**	**.746**	.038	.037	**.100**	–	**.726**	**.687**
			30	.030	.039	**.143**	–	**.800**	**.753**	**.022**	.030	**.092**	–	**.723**	**.701**
			40	**.022**	.031	**.142**	–	**.773**	**.780**	.025	.030	**.109**	–	**.748**	**.737**
		600/300	20	.039	.051	**.132**	–	**.840**	**.764**	.051	.056	**.080**	–	**.770**	**.704**
			30	.037	.051	**.148**	–	**.840**	**.780**	.048	.050	**.101**	–	**.816**	**.748**
			40	.038	.048	**.164**	–	**.850**	**.778**	.035	.040	**.104**	–	**.800**	**.749**
1800	$\mathcal{N}(0,1)$	450/450	20	.061	.066	**.111**	.053	**.784**	**.784**	.040	.069	.067	.049	**.738**	**.736**
			30	.037	.053	**.107**	.048	**.780**	**.781**	.047	.053	**.083**	.055	**.780**	**.766**
			40	.039	.050	**.110**	.069	**.813**	**.811**	.041	.067	**.086**	.061	**.794**	**.783**
		600/300	20	.047	.062	**.094**	.042	**.847**	**.811**	.049	.062	.073	.059	**.809**	**.777**
			30	.049	.065	**.106**	.049	**.855**	**.816**	.043	.069	**.083**	.052	**.809**	**.770**
			40	.029	.054	**.098**	.048	**.847**	**.816**	.042	.052	**.082**	.058	**.840**	**.811**
	$\mathcal{N}(1/2,2/3)$	450/450	20	.041	.048	**.116**	–	**.774**	**.789**	.051	.055	**.092**	–	**.748**	**.745**
			30	.033	.040	**.132**	–	**.816**	**.791**	.036	.049	**.098**	–	**.781**	**.784**
			40	.044	.037	**.165**	–	**.843**	**.811**	.033	.043	**.085**	–	**.771**	**.770**
		600/300	20	.043	.038	**.133**	–	**.846**	**.772**	.055	.072	**.077**	–	**.816**	**.755**
			30	.037	.045	**.141**	–	**.874**	**.831**	.047	.064	**.077**	–	**.832**	**.798**
			40	.041	.056	**.147**	–	**.856**	**.807**	.049	.063	**.123**	–	**.845**	**.804**
2700	$\mathcal{N}(0,1)$	450/450	20	.043	.062	**.107**	.071	**.801**	**.801**	.047	.064	.062	.048	**.778**	**.776**
			30	.049	.065	**.095**	.052	**.806**	**.798**	.050	.067	.074	.049	**.793**	**.800**
			40	.038	.047	**.092**	.045	**.830**	**.817**	.034	.054	.074	.045	**.812**	**.815**
		600/300	20	.046	.067	**.091**	.051	**.841**	**.824**	.049	.066	.071	.053	**.802**	**.771**
			30	.049	.067	**.101**	.046	**.873**	**.841**	.042	.067	.064	.055	**.836**	**.806**
			40	.038	.045	**.111**	.041	**.870**	**.851**	.043	.067	**.081**	.057	**.833**	**.811**
	$\mathcal{N}(1/2,2/3)$	450/450	20	.040	.058	**.126**	–	**.799**	**.807**	.049	.058	**.088**	–	**.763**	**.768**
			30	**.024**	.029	**.145**	–	**.828**	**.812**	.040	.056	**.102**	–	**.804**	**.787**
			40	.039	.042	**.156**	–	**.827**	**.838**	.037	.038	**.084**	–	**.810**	**.788**
		600/300	20	.049	.051	**.130**	–	**.885**	**.831**	.042	.056	**.082**	–	**.835**	**.771**
			30	.040	.040	**.159**	–	**.879**	**.816**	.047	.055	**.105**	–	**.837**	**.792**
			40	.046	.053	**.140**	–	**.862**	**.843**	.040	.052	**.091**	–	**.868**	**.830**

Table 11: Empirical Type I error rates for detecting DTF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

Appendix D Empirical Power Rates for DTF and DBF Simulations

				5 Anchors					10 Anchors				
DIF Items	Focal Distribution	Sample Sizes Equal	Test Length	*sDTF*	SIBTEST	*dDTF*	*sDBF*	*dDBF*	*sDTF*	SIBTEST	*dDTF*	*sDBF*	*dDBF*
1,3,5	$\mathcal{N}(0,1)$	Yes	20	.052	.173	.155	.159	.915	.062	.280	.455	.166	.967
			30	.040	.150	.074	.161	.912	.045	.148	.139	.146	.975
			40	.031	.119	.038	.130	.918	.034	.106	.064	.158	.978
		No	20	.063	.243	.303	.313	1.000	.109	.445	.811	.319	1.000
			30	.059	.160	.155	.308	1.000	.077	.215	.355	.338	1.000
			40	.054	.141	.089	.282	.999	.062	.156	.179	.306	1.000
	$\mathcal{N}(1/2,2/3)$	Yes	20	.041	.303	.131	.132	.842	.043	.449	.387	.134	.923
			30	.038	.201	.051	.125	.867	.031	.251	.166	.138	.946
			40	.022	.218	.034	.111	.840	.022	.172	.059	.129	.928
		No	20	.059	.367	.315	.217	.998	.081	.645	.775	.221	1.000
			30	.045	.265	.119	.216	.998	.057	.346	.343	.245	1.000
			40	.031	.214	.068	.190	.999	.038	.244	.166	.220	1.000
1,2,3,4,5	$\mathcal{N}(0,1)$	Yes	20	.111	.321	.338	.399	.981	.227	.638	.864	.478	.996
			30	.058	.187	.127	.374	.983	.090	.299	.362	.453	.997
			40	.028	.150	.061	.324	.978	.063	.195	.187	.454	.997
		No	20	.176	.447	.677	.716	1.000	.429	.848	.996	.804	1.000
			30	.090	.248	.304	.681	1.000	.192	.448	.738	.794	1.000
			40	.058	.176	.155	.681	1.000	.102	.294	.443	.777	1.000
	$\mathcal{N}(1/2,2/3)$	Yes	20	.073	.464	.322	.295	.972	.166	.834	.843	.314	.998
			30	.039	.319	.141	.244	.979	.065	.491	.386	.319	.999
			40	.019	.246	.043	.261	.983	.047	.316	.188	.292	.997
		No	20	.172	.672	.716	.546	1.000	.353	.979	.997	.651	1.000
			30	.077	.436	.314	.536	1.000	.141	.676	.765	.610	1.000
			40	.055	.302	.158	.537	1.000	.100	.449	.447	.633	1.000

Table 12: DBF and DTF Power rates when $N = 900$ when all items are included.

				5 Anchors					10 Anchors				
DIF Items	Focal Distribution	Sample Sizes Equal	Test Length	*sDTF*	SIBTEST	*dDTF*	*sDBF*	*dDBF*	*sDTF*	SIBTEST	*dDTF*	*sDBF*	*dDBF*
1,3,5	$\mathcal{N}(0,1)$	Yes	20	.065	.233	.369	.307	.995	.157	.540	.834	.364	1.000
			30	.052	.164	.133	.298	.999	.067	.224	.407	.320	1.000
			40	.039	.141	.092	.311	1.000	.064	.167	.186	.324	1.000
		No	20	.096	.372	.616	.549	1.000	.228	.734	.986	.606	1.000
			30	.068	.230	.287	.566	1.000	.123	.363	.705	.601	1.000
			40	.054	.169	.164	.575	1.000	.084	.236	.397	.634	1.000
	$\mathcal{N}(1/2,2/3)$	Yes	20	.062	.361	.369	.214	.999	.086	.710	.809	.206	1.000
			30	.040	.256	.138	.199	.997	.060	.369	.389	.261	1.000
			40	.037	.235	.075	.207	1.000	.059	.263	.238	.251	1.000
		No	20	.083	.582	.681	.395	1.000	.166	.930	.989	.400	1.000
			30	.056	.354	.293	.395	1.000	.074	.573	.715	.433	1.000
			40	.048	.290	.166	.417	1.000	.065	.374	.393	.441	1.000
1,2,3,4,5	$\mathcal{N}(0,1)$	Yes	20	.184	.471	.698	.710	1.000	.490	.909	.997	.812	1.000
			30	.096	.257	.357	.708	1.000	.222	.491	.793	.815	1.000
			40	.069	.200	.179	.701	1.000	.121	.296	.453	.782	1.000
		No	20	.361	.715	.959	.956	1.000	.770	.993	1.000	.981	1.000
			30	.183	.414	.659	.952	1.000	.342	.702	.980	.987	1.000
			40	.108	.272	.388	.955	1.000	.200	.458	.790	.983	1.000
	$\mathcal{N}(1/2,2/3)$	Yes	20	.191	.682	.755	.528	1.000	.340	.980	.999	.562	1.000
			30	.071	.443	.385	.523	1.000	.136	.693	.806	.610	1.000
			40	.068	.331	.196	.535	1.000	.084	.477	.520	.586	1.000
		No	20	.287	.886	.963	.852	1.000	.636	.999	1.000	.904	1.000
			30	.139	.617	.688	.866	1.000	.300	.919	.987	.898	1.000
			40	.111	.432	.396	.870	1.000	.196	.727	.827	.907	1.000

Table 13: DBF and DTF Power rates when $N = 1800$ when all items are included.

				5 Anchors					10 Anchors				
DIF Items	Focal Distribution	Sample Sizes Equal	Test Length	*sDTF*	SIBTEST	*dDTF*	*sDBF*	*dDBF*	*sDTF*	SIBTEST	*dDTF*	*sDBF*	*dDBF*
1,3,5	$\mathcal{N}(0,1)$	Yes	20	.089	.326	.527	.452	1.000	.169	.635	.956	.469	1.000
			30	.073	.233	.254	.444	1.000	.098	.311	.584	.486	1.000
			40	.052	.166	.139	.437	1.000	.054	.202	.307	.482	1.000
		No	20	.130	.462	.839	.715	1.000	.287	.881	1.000	.773	1.000
			30	.092	.276	.440	.755	1.000	.142	.468	.892	.787	1.000
			40	.070	.210	.257	.756	1.000	.086	.303	.597	.790	1.000
	$\mathcal{N}(1/2,2/3)$	Yes	20	.070	.456	.523	.277	1.000	.146	.862	.958	.331	1.000
			30	.049	.332	.228	.294	1.000	.097	.498	.615	.332	1.000
			40	.058	.274	.156	.302	1.000	.060	.340	.352	.326	1.000
		No	20	.127	.684	.862	.544	1.000	.278	.978	.999	.568	1.000
			30	.070	.465	.468	.546	1.000	.109	.745	.893	.576	1.000
			40	.048	.265	.217	.552	1.000	.082	.472	.588	.612	1.000
1,2,3,4,5	$\mathcal{N}(0,1)$	Yes	20	.310	.663	.898	.899	1.000	.655	.977	1.000	.919	1.000
			30	.141	.356	.573	.852	1.000	.293	.666	.945	.930	1.000
			40	.093	.277	.325	.869	1.000	.167	.399	.656	.926	1.000
		No	20	.491	.853	1.000	.986	1.000	.904	.999	1.000	.996	1.000
			30	.232	.544	.870	.995	1.000	.509	.866	.998	1.000	1.000
			40	.158	.365	.579	.991	1.000	.291	.596	.935	.998	1.000
	$\mathcal{N}(1/2,2/3)$	Yes	20	.243	.823	.929	.671	1.000	.488	.998	1.000	.737	1.000
			30	.110	.528	.549	.698	1.000	.201	.849	.957	.740	1.000
			40	.069	.403	.328	.708	1.000	.125	.602	.709	.757	1.000
		No	20	.442	.969	.998	.949	1.000	.795	1.000	1.000	.971	1.000
			30	.236	.746	.876	.974	1.000	.415	.988	.999	.981	1.000
			40	.122	.556	.606	.976	1.000	.252	.833	.936	.975	1.000

Table 14: DBF and DTF Power rates when $N = 2700$ when all items are included.

Appendix E Type I Errors for Complete DTF and DBF Cancellation Simulations

				Bundle				Test					
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DTF_{LF}	DTF_{SF}	*sDTF*	*dDTF*	SIBTEST	SIBTEST_{UC}	DTF_{LF}	DTF_{SF}
5	$\mathcal{N}(0,1)$	450/450	20	.041	.025	**.741**	**.754**	.035	.049	.074	.036	**.671**	**.692**
			30	.034	**.010**	**.736**	**.735**	.029	.050	**.095**	.056	**.751**	**.761**
			40	.031	**.011**	**.764**	**.763**	.039	.038	**.120**	.049	**.765**	**.771**
		600/300	20	.045	.030	**.796**	**.748**	.048	.039	**.101**	.046	**.769**	**.687**
			30	.036	**.024**	**.828**	**.781**	.033	.053	**.104**	.048	**.790**	**.731**
			40	**.022**	**.022**	**.825**	**.782**	.033	.048	**.087**	.045	**.796**	**.771**
	$\mathcal{N}(1/2,2/3)$	450/450	20	.027	**.016**	**.793**	**.754**	.026	.044	**.140**	–	**.719**	**.698**
			30	.027	**.013**	**.794**	**.774**	.037	.040	**.159**	–	**.777**	**.744**
			40	**.019**	**.010**	**.789**	**.767**	**.022**	.028	**.166**	–	**.773**	**.771**
		600/300	20	.042	**.024**	**.814**	**.774**	.036	.056	**.128**	–	**.781**	**.714**
			30	.039	.030	**.837**	**.788**	.033	.047	**.139**	–	**.805**	**.738**
			40	.035	**.013**	**.836**	**.767**	.026	.032	**.126**	–	**.820**	**.749**
10	$\mathcal{N}(0,1)$	450/450	20	.035	**.014**	**.725**	**.752**	.033	.053	.068	.053	**.633**	**.607**
			30	.035	**.016**	**.778**	**.769**	.033	.056	**.080**	.054	**.676**	**.704**
			40	.030	**.006**	**.773**	**.779**	.028	.040	**.081**	.053	**.709**	**.722**
		600/300	20	.042	.027	**.793**	**.751**	.036	.054	.064	.057	**.665**	**.619**
			30	.043	**.023**	**.816**	**.765**	.042	.057	**.096**	.061	**.719**	**.686**
			40	.037	**.021**	**.789**	**.758**	.035	.049	**.078**	.048	**.730**	**.691**
	$\mathcal{N}(1/2,2/3)$	450/450	20	**.024**	**.010**	**.787**	**.755**	.026	.046	**.076**	–	**.650**	**.590**
			30	.025	**.013**	**.778**	**.725**	**.024**	.040	**.107**	–	**.712**	**.680**
			40	**.021**	**.011**	**.802**	**.772**	**.016**	.035	**.106**	–	**.744**	**.732**
		600/300	20	.033	**.017**	**.834**	**.734**	.047	.047	.073	–	**.719**	**.628**
			30	.043	.027	**.861**	**.782**	.041	.046	**.092**	–	**.764**	**.688**
			40	.037	**.019**	**.832**	**.778**	.041	.066	**.090**	–	**.798**	**.715**

Table 15: Cancellation Type I error rates when $N = 900$ with two focal items containing balanced DIF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

				Bundle				Test					
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DTF_{LF}	DTF_{SF}	*sDTF*	*dDTF*	SIBTEST	SIBTEST_{UC}	DTF_{LF}	DTF_{SF}
5	$\mathcal{N}(0,1)$	450/450	20	.031	.025	**.699**	**.727**	.029	.044	**.087**	.047	**.674**	**.695**
			30	.028	**.018**	**.757**	**.758**	.032	.050	**.112**	.059	**.719**	**.746**
			40	.024	**.014**	**.780**	**.776**	.030	.029	**.104**	.055	**.752**	**.754**
		600/300	20	.027	.029	**.761**	**.719**	.035	.052	**.083**	.046	**.745**	**.688**
			30	.044	.030	**.766**	**.739**	.032	.044	**.095**	.044	**.770**	**.721**
			40	.037	**.018**	**.804**	**.775**	.031	.047	**.105**	.050	**.814**	**.763**
	$\mathcal{N}(1/2,2/3)$	450/450	20	.030	.025	**.740**	**.750**	**.024**	.034	**.134**	–	**.741**	**.673**
			30	.030	**.016**	**.769**	**.745**	.026	.037	**.147**	–	**.738**	**.763**
			40	**.017**	**.017**	**.815**	**.738**	.029	.033	**.173**	–	**.789**	**.729**
		600/300	20	.035	.030	**.813**	**.756**	.041	.042	**.118**	–	**.783**	**.688**
			30	.030	**.019**	**.840**	**.754**	.031	.042	**.127**	–	**.797**	**.739**
			40	.039	**.017**	**.841**	**.776**	.028	.035	**.154**	–	**.859**	**.768**
10	$\mathcal{N}(0,1)$	450/450	20	.039	**.016**	**.707**	**.727**	.034	.037	.072	.053	**.608**	**.603**
			30	.032	**.017**	**.753**	**.745**	.034	.048	**.076**	.063	**.695**	**.707**
			40	.030	**.012**	**.761**	**.764**	**.021**	.038	.074	.051	**.715**	**.727**
		600/300	20	.044	.032	**.755**	**.714**	.043	.046	.067	.055	**.640**	**.619**
			30	.033	.025	**.811**	**.748**	.044	.066	**.076**	.056	**.742**	**.695**
			40	.032	**.024**	**.809**	**.774**	.025	.046	.074	.048	**.765**	**.726**
	$\mathcal{N}(1/2,2/3)$	450/450	20	.030	**.018**	**.753**	**.732**	.035	.038	**.093**	–	**.644**	**.625**
			30	.028	**.020**	**.772**	**.773**	.035	.030	**.123**	–	**.699**	**.680**
			40	.027	**.017**	**.769**	**.789**	**.024**	.035	**.097**	–	**.705**	**.713**
		600/300	20	.039	.028	**.829**	**.707**	.037	.042	.060	–	**.672**	**.623**
			30	.037	.026	**.835**	**.742**	.029	.040	**.084**	–	**.718**	**.669**
			40	.031	**.017**	**.807**	**.780**	.033	.038	**.088**	–	**.767**	**.717**

Table 16: Cancellation Type I error rates when $N = 900$ with four focal items containing balanced DIF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

				Bundle				Test					
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DTF_{LF}	DTF_{SF}	*sDTF*	*dDTF*	SIBTEST	SIBTEST_{UC}	DTF_{LF}	DTF_{SF}
5	$\mathcal{N}(0,1)$	900/900	20	.049	.034	**.798**	**.788**	.042	.060	**.090**	.052	**.722**	**.706**
			30	.032	.030	**.789**	**.792**	.046	.068	**.104**	.046	**.780**	**.743**
			40	.037	.033	**.793**	**.779**	.047	.061	**.093**	.040	**.784**	**.772**
		1200/600	20	.063	.032	**.811**	**.787**	.052	.065	**.097**	.053	**.771**	**.692**
			30	.047	.034	**.834**	**.808**	.040	.057	**.107**	.060	**.807**	**.770**
			40	.044	.030	**.831**	**.792**	.038	.042	**.105**	.050	**.802**	**.799**
	$\mathcal{N}(1/2,2/3)$	900/900	20	.029	.027	**.800**	**.769**	.035	.046	**.143**	–	**.770**	**.717**
			30	.037	.025	**.849**	**.810**	.029	.042	**.173**	–	**.790**	**.787**
			40	.039	**.024**	**.827**	**.816**	.036	.049	**.157**	–	**.791**	**.803**
		1200/600	20	.053	.050	**.847**	**.761**	.042	.065	**.129**	–	**.771**	**.739**
			30	.049	.036	**.862**	**.789**	.040	.061	**.128**	–	**.819**	**.783**
			40	.051	**.024**	**.856**	**.823**	.039	.044	**.157**	–	**.830**	**.801**
10	$\mathcal{N}(0,1)$	900/900	20	.030	**.023**	**.769**	**.775**	.035	.070	.061	.049	**.633**	**.649**
			30	.033	**.024**	**.771**	**.777**	.046	.065	**.080**	.064	**.692**	**.710**
			40	.025	**.017**	**.801**	**.806**	.036	.047	**.077**	.044	**.740**	**.728**
		1200/600	20	.040	.027	**.823**	**.767**	.040	.067	.057	.039	**.688**	**.592**
			30	.037	.037	**.822**	**.795**	.048	.066	**.087**	.060	**.776**	**.684**
			40	.049	.025	**.853**	**.804**	.037	.058	.073	.054	**.778**	**.730**
	$\mathcal{N}(1/2,2/3)$	900/900	20	.037	**.019**	**.797**	**.780**	.039	.054	**.094**	–	**.663**	**.656**
			30	.038	**.018**	**.832**	**.793**	.031	.059	**.089**	–	**.760**	**.721**
			40	.038	**.018**	**.824**	**.808**	.029	.046	**.088**	–	**.765**	**.728**
		1200/600	20	.051	.037	**.854**	**.774**	.044	.071	.067	–	**.738**	**.624**
			30	.053	.041	**.854**	**.804**	.049	.067	**.102**	–	**.790**	**.704**
			40	.037	.024	**.871**	**.812**	.034	.053	**.084**	–	**.811**	**.763**

Table 17: Cancellation Type I error rates when $N = 1800$ with two focal items containing balanced DIF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

				Bundle				Test					
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DTF_{LF}	DTF_{SF}	*sDTF*	*dDTF*	SIBTEST	SIBTEST_{UC}	DTF_{LF}	DTF_{SF}
5	$\mathcal{N}(0,1)$	900/900	20	.038	.029	**.750**	**.728**	.041	.056	**.079**	.044	**.710**	**.693**
			30	.038	.029	**.778**	**.780**	.039	.037	**.100**	.043	**.756**	**.734**
			40	.034	.026	**.775**	**.785**	.031	.043	**.107**	.046	**.782**	**.782**
		1200/600	20	.046	.036	**.797**	**.746**	.049	.067	**.097**	.052	**.743**	**.707**
			30	.045	.040	**.812**	**.777**	.048	.056	**.093**	.058	**.791**	**.737**
			40	.043	.032	**.812**	**.803**	.036	.046	**.107**	.055	**.838**	**.768**
	$\mathcal{N}(1/2,2/3)$	900/900	20	.034	.035	**.804**	**.761**	.052	.054	**.154**	–	**.759**	**.718**
			30	.036	**.022**	**.782**	**.789**	.033	.039	**.124**	–	**.771**	**.742**
			40	.039	.033	**.796**	**.778**	.025	.029	**.149**	–	**.781**	**.763**
		1200/600	20	.048	.037	**.823**	**.751**	.046	.057	**.127**	–	**.803**	**.728**
			30	.037	.033	**.848**	**.800**	.040	.035	**.142**	–	**.808**	**.751**
			40	.040	.034	**.857**	**.803**	.030	.050	**.136**	–	**.850**	**.787**
10	$\mathcal{N}(0,1)$	900/900	20	.034	.025	**.754**	**.721**	.048	.055	.072	.064	**.627**	**.628**
			30	.049	.029	**.786**	**.765**	.041	.061	.069	.049	**.695**	**.686**
			40	.035	.037	**.799**	**.787**	.038	.054	**.085**	.062	**.739**	**.750**
		1200/600	20	.039	.025	**.779**	**.722**	.046	.058	.055	.048	**.655**	**.623**
			30	.048	.040	**.812**	**.750**	.039	.060	.058	.044	**.735**	**.705**
			40	.041	.043	**.821**	**.788**	.041	.059	.072	.049	**.790**	**.729**
	$\mathcal{N}(1/2,2/3)$	900/900	20	.034	.027	**.781**	**.753**	.032	.043	**.078**	–	**.685**	**.620**
			30	.043	.026	**.800**	**.758**	.032	.048	**.084**	–	**.721**	**.672**
			40	.038	.026	**.793**	**.787**	.047	.047	**.105**	–	**.764**	**.761**
		1200/600	20	.043	.036	**.827**	**.738**	.051	.050	**.079**	–	**.738**	**.632**
			30	.040	.039	**.829**	**.765**	.034	.058	.074	–	**.774**	**.681**
			40	.038	.034	**.835**	**.784**	.038	.046	**.088**	–	**.812**	**.704**

Table 18: Cancellation Type I error rates when $N = 1800$ with four focal items containing balanced DIF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

				Bundle				Test					
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DTF_{LF}	DTF_{SF}	*sDTF*	*dDTF*	SIBTEST	SIBTEST_{UC}	DTF_{LF}	DTF_{SF}
5	$\mathcal{N}(0,1)$	1350/1350	20	.045	**.020**	**.783**	**.793**	.040	.055	**.083**	.045	**.708**	**.725**
			30	.042	**.022**	**.801**	**.808**	.055	.072	**.101**	.063	**.767**	**.761**
			40	.042	.031	**.825**	**.803**	.040	.059	**.099**	.047	**.780**	**.775**
		1800/900	20	.058	.042	**.824**	**.789**	.052	**.083**	**.096**	.062	**.777**	**.709**
			30	.057	.033	**.863**	**.826**	.035	.058	**.097**	.054	**.811**	**.770**
			40	.040	.033	**.858**	**.834**	.035	.057	**.101**	.049	**.844**	**.777**
	$\mathcal{N}(1/2,2/3)$	1350/1350	20	.048	.037	**.822**	**.801**	.043	.052	**.145**	–	**.773**	**.740**
			30	.050	.029	**.840**	**.825**	.039	.043	**.146**	–	**.818**	**.771**
			40	.035	.027	**.838**	**.830**	.031	.043	**.143**	–	**.806**	**.791**
		1800/900	20	.053	.030	**.859**	**.795**	.045	**.077**	**.126**	–	**.791**	**.728**
			30	.045	.033	**.877**	**.806**	.047	.055	**.138**	–	**.843**	**.778**
			40	.044	.032	**.866**	**.838**	.046	.051	**.157**	–	**.850**	**.808**
10	$\mathcal{N}(0,1)$	1350/1350	20	.034	**.024**	**.795**	**.778**	.052	.075	**.076**	.057	**.651**	**.660**
			30	.046	.027	**.793**	**.787**	.032	.060	.062	.041	**.720**	**.730**
			40	.034	.025	**.815**	**.813**	.037	.050	.070	.046	**.753**	**.758**
		1800/900	20	.043	.031	**.815**	**.787**	.055	.072	.064	.047	**.698**	**.650**
			30	.048	.039	**.823**	**.796**	.055	.074	**.080**	.062	**.741**	**.723**
			40	.050	.035	**.866**	**.834**	.056	.068	**.079**	.059	**.807**	**.730**
	$\mathcal{N}(1/2,2/3)$	1350/1350	20	.040	.032	**.816**	**.793**	.041	.061	.071	–	**.693**	**.635**
			30	.033	.032	**.827**	**.830**	.048	.058	**.102**	–	**.762**	**.720**
			40	.044	.036	**.824**	**.818**	.039	.056	**.086**	–	**.781**	**.748**
		1800/900	20	.044	.035	**.845**	**.806**	.048	.066	**.082**	–	**.764**	**.656**
			30	.058	.036	**.886**	**.825**	.050	.073	**.099**	–	**.779**	**.700**
			40	.041	.041	**.872**	**.812**	.034	.064	**.096**	–	**.798**	**.740**

Table 19: Cancellation Type I error rates when $N = 2700$ with two focal items containing balanced DIF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

				Bundle				Test					
Anchors	Focal Distribution	Sample Sizes	Test Length	*sDBF*	*dDBF*	DTF_{LF}	DTF_{SF}	*sDTF*	*dDTF*	SIBTEST	SIBTEST_{UC}	DTF_{LF}	DTF_{SF}
5	$\mathcal{N}(0,1)$	1350/1350	20	.043	.036	**.762**	**.771**	.045	.054	**.095**	.061	**.712**	**.711**
			30	.046	.040	**.783**	**.795**	.042	.064	**.117**	.053	**.778**	**.772**
			40	.039	.043	**.812**	**.802**	.035	.054	**.101**	.056	**.787**	**.792**
		1800/900	20	.060	.027	**.816**	**.757**	.042	.050	**.089**	.041	**.763**	**.689**
			30	.042	.038	**.831**	**.801**	.053	.063	**.094**	.048	**.779**	**.764**
			40	.049	.044	**.829**	**.808**	.033	.049	**.093**	.047	**.840**	**.797**
	$\mathcal{N}(1/2,2/3)$	1350/1350	20	.045	.036	**.813**	**.787**	.043	.050	**.133**	–	**.757**	**.702**
			30	.048	.024	**.819**	**.800**	.038	.044	**.142**	–	**.794**	**.778**
			40	.025	.027	**.797**	**.812**	.035	.041	**.145**	–	**.802**	**.798**
		1800/900	20	.042	.031	**.831**	**.761**	.047	.052	**.136**	–	**.782**	**.730**
			30	.050	.042	**.852**	**.788**	.036	.051	**.127**	–	**.828**	**.793**
			40	.044	.035	**.873**	**.825**	.044	.052	**.144**	–	**.839**	**.792**
10	$\mathcal{N}(0,1)$	1350/1350	20	.044	.038	**.751**	**.766**	.039	.058	.057	.049	**.643**	**.615**
			30	.047	.033	**.786**	**.807**	.060	.058	**.087**	.065	**.727**	**.698**
			40	.044	.032	**.794**	**.812**	.040	.056	.073	.048	**.732**	**.744**
		1800/900	20	.050	.040	**.794**	**.733**	.051	.065	.065	.051	**.691**	**.631**
			30	.042	.045	**.814**	**.771**	.050	**.076**	**.079**	.061	**.772**	**.700**
			40	.049	.044	**.840**	**.813**	.047	.071	**.076**	.045	**.765**	**.734**
	$\mathcal{N}(1/2,2/3)$	1350/1350	20	.052	.030	**.783**	**.765**	.056	.051	**.092**	–	**.720**	**.647**
			30	.040	.040	**.823**	**.787**	.048	.055	**.098**	–	**.737**	**.711**
			40	.030	.031	**.808**	**.824**	.035	.047	**.097**	–	**.788**	**.754**
		1800/900	20	.052	.041	**.838**	**.752**	.046	.059	**.086**	–	**.763**	**.626**
			30	.035	.033	**.866**	**.777**	.045	.057	**.098**	–	**.770**	**.716**
			40	.048	.029	**.871**	**.827**	.048	.052	**.101**	–	**.822**	**.786**

Table 20: Cancellation Type I error rates when $N = 2700$ with four focal items containing balanced DIF. Type I error rates greater than .075 and less than .025 are highlighted in bold.

Appendix F Empirical Coverage Rates for Conditional DRF Measures

θ Sign	Test Length	DRF Measure	10	9	8	7	6	5	4	3	2	1	0
Negative (−)	20	$sDIF_\theta$	.947	.946	.946	.945	.944	.943	.940	.940	.944	.942	.956
		$sDBF_\theta$	.942	.942	.941	.943	.942	.942	.937	.939	.939	.939	.938
		$sDTF_\theta$	.953	.953	.955	.955	.954	.957	.957	.961	.959	.958	.954
	30	$sDIF_\theta$	.959	.960	.961	.962	.958	.955	.952	.950	.950	.948	.954
		$sDBF_\theta$	.950	.951	.950	.950	.951	.952	.950	.951	.952	.963	.953
		$sDTF_\theta$	.960	.962	.960	.958	.955	.955	.956	.957	.955	.961	.955
	40	$sDIF_\theta$	.942	.939	.940	.944	.944	.946	.950	.947	.947	.954	.957
		$sDBF_\theta$	.956	.956	.957	.955	.956	.952	.953	.953	.964	.955	.951
		$sDTF_\theta$	.946	.948	.947	.950	.956	.955	.951	.952	.965	.960	.955
Positive (+)	20	$sDIF_\theta$	.948	.949	.946	.946	.946	.944	.943	.943	.944	.951	–
		$sDBF_\theta$	.941	.941	.940	.942	.943	.944	.939	.934	.941	.948	–
		$sDTF_\theta$	.957	.956	.956	.955	.955	.955	.953	.954	.949	.947	–
	30	$sDIF_\theta$	.953	.955	.954	.949	.947	.948	.946	.950	.951	.953	–
		$sDBF_\theta$	.951	.950	.951	.950	.948	.950	.954	.956	.960	.958	–
		$sDTF_\theta$	.960	.956	.959	.959	.959	.957	.950	.955	.956	.941	–
	40	$sDIF_\theta$	.940	.939	.938	.938	.935	.942	.937	.941	.947	.945	–
		$sDBF_\theta$	.951	.949	.950	.950	.946	.948	.947	.949	.954	.956	–
		$sDTF_\theta$	.946	.950	.951	.950	.950	.952	.954	.955	.953	.953	–

Table 21: Empirical 95% coverage rates for conditional DRF measures at various θ locations.