ROLLINS III, JONATHAN DARRELL, Ph.D. A Comparison of Observed Score Approaches to Detecting Differential Item Functioning among Multiple Groups. (2018) Directed by Dr. Richard M. Luecht and Dr. John T. Willse. 138 pp.

The overall purpose of this dissertation was to compare various observed score approaches in detecting differential item functioning among multiple examinee groups simultaneously. Specifically, this study contributes to the literature base by investigating a lasso-constraint observed score method (i.e., logistic regression lasso; LR lasso) in the context of multiple groups as well as features of test design related to test information targets. Given that a lasso-constraint method has not been extended for multiple groups using observed scores, comparisons are made with other observed score techniques (i.e., generalized Mantel-Haenszel χ^2 and generalized logistic regression) while using item response theory to generate data (thus avoiding model-data congruity complications in the study design).

Multiple variables were manipulated in a simulation study at the test-level (e.g., the location of the test information target relative to the central tendency of the examinee population, and the shape of the test information function), item-level (e.g., the location of DIF items relative to the test information target, and the percentage of DIF items), and for simulees (e.g., the amount of impact and sample size balance). The relative lack of literature which explores DIF as it relates to target test information functions provided the exigency for exploring it within this study, along with its typical absence in literature using IRT generation models. Practitioners may find the results useful in judging the merit of adopting the newer lasso method for detecting DIF within multiple groups as opposed to pre-existing methods. Furthermore, the test design features of this study allow

for the interpretation to be less theoretical in nature and better aligned with standard operational practices, such as building exams to be optimized at test information targets, for example.

The results provide consilience that the LR lasso method has inflated type I error overall with no additional benefit in power. In fact, even when type I error rates are comparable across methods, LR lasso has a lower hit rate in many instances (i.e., higher type II error rate). The sensitivity of LR lasso to detecting DIF items seems to be substantially influenced by having an increased number of DIF items on a form. Recommendations for practitioners, as well as limitations and directions for future research, are provided as well.

Taken collectively, the results of the simulation study can be interpreted to support the claim that LR lasso fails to perform comparably with more established methods for multiple groups DIF detection across numerous instances but could potentially have merit in practical application in situations that have yet to be explored. While some limitations of LR lasso were noted within this study, there are a variety of other conditions which need to be explored before practitioners discard the method altogether (a few such studies are suggested). It may well be the case that the added complexity afforded by the regularization in estimating the group-specific model parameters through lasso constraints may confound the detection of the DIF items.

A COMPARISON OF OBSERVED SCORE APPROACHES TO DETECTING DIFFERENTIAL ITEM FUNCTIONING AMONG MULTIPLE GROUPS

by

Jonathan Darrell Rollins III

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2018

Approved by

____
Committee Co-Chair

____
Committee Co-Chair

APPROVAL PAGE

This dissertation written by JONATHAN DARRELL ROLLINS III has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Co-Chairs ____________________
Dr. Richard M. Luecht

____
Dr. John T. Willse

Committee Members ____________________
Dr. Randall Penfield

____
Dr. Devdass Sunnassee

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

It is with great gratitude and humility that I give thanks in successfully reaching this milestone in my life. Above everything else, God is given all sovereignty and glory. I love both of my parents and brother with the ever-constant support and unconditional love that they have shown me throughout my life. None of this dissertation would have been remotely possible without them. Close friends and family at every step in my education have been such a positive support, and largely have held a role in my progress up to this point.

Without the help of my committee members, a great portion of my knowledge and this dissertation would not exist. My committee co-chairs, Dr. Luecht and Dr. Willse, have had an absolutely pivotal impact on my knowledge and skills that will stay with me for a lifetime. I cannot express enough gratitude for their unwavering support and suggestions during this process, as well as throughout my time in graduate school and as I start my career. They inspire me to strive for greatness. Also, I would like to thank Dr. Sunnassee for his guidance and support. He has had a significant impact on how I approach many challenges and on how I reflect on my progress. I am indebted to Dr. Penfield for his encouragement and suggestions, as well as his influence on many of my current research interests.

Additionally, I would like to underscore the importance of the rest of the ERM faculty with whom I took coursework and collaborated on projects, and Dr. Ackerman, Dr. Chalhoub-Deville, Dr. Downs, and Dr. Henson. All my professors, as well as my experiences with Winston-Salem / Forsyth County Schools, the College Entrance Examination Board, and the Georgia Department of Education molded me as a scholar. It would be remiss of me not to give thanks to the rest of the educators throughout my entire life in giving me knowledge and help in reaching this point.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The overall purpose of this dissertation is to compare various observed score approaches in detecting differential item functioning among multiple examinee groups simultaneously. Specifically, this study contributes to the literature base by investigating a lasso-constraint observed score method in the context of multiple groups as well as features of test design related to test information targets. Given that a lasso-constraint method has not been extended for multiple groups using observed scores, comparisons are made with other observed score techniques while using item response theory to generate data (thus avoiding model-data congruity complications in the study design). To support the overall purpose, the scope of the current chapter includes background information for differential item functioning, a detailed purpose and rationale, research questions, and definitions and notation of key terms used throughout the study.

Background of the Problem

Item-level bias in which the probability of a correct response among equally able persons differs in subgroups is known as differential item functioning (DIF; Tutz & Schauberger, 2015). DIF can also be defined as a violation of item-level invariance across subpopulations (Kamata & Vaughn, 2004). The *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*; AERA, APA, & NCME, 2014) provides a formal definition of DIF:

> *Differential item functioning* occurs when different groups of test takers with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item (p.16).

It is important to clarify the difference between DIF and impact. Plainly stated, a group difference in ability or performance is not DIF. Impact refers to such differences in the overall distributions of the ability or performance of intact groups, and thus is a group-level measure (Dorans & Holland, 1993). DIF, on the other hand, is an item-level phenomenon. DIF is examined by first matching examinees in different groups on a criterion variable, typically ability or performance level. Because of this matching, DIF is unexpected as the groups have been made comparable with respect to the measured construct.

To distinguish between groups (usually demographic groupings), terms are used to describe relative advantage or disadvantage with respect to responding correctly on an item. The reference group is the group which may potentially have an advantage in answering the item correctly, while the focal group is the group of concern because they may potentially have a disadvantage in answering the item correctly. As an example within the context of testing in the United States, and in the case of the usual DIF methods that assume two groups, the reference group tends to be Caucasian American examinees while the focal group would be a combination of students from other racial groups. As another example, males may be considered the reference group and females as the focal group.

As it pertains to this study, when DIF is examined across more than two groups simultaneously there is one reference group and multiple focal groups. Given the example of race described previously, the multiple groups of races would not be combined into one focal group, but each group would remain intact for the analysis and each be treated as a unique focal group. Often, all focal groups are combined into a single focal group to alleviate issues related to balance between groups (i.e., statistical power) and pairwise comparisons (i.e., increased type I family error rates). When multiple focal groups are combined, there is an underpinning assumption that the groups have roughly the same ability distributions and potentially the same level of disadvantage on a given item. However, when the groups are not truly comparable, multiple group methods which allow the groups to remain separated are warranted.

Another example can be found whenever exams are administered in multiple languages (Angoff & Sharon, 1974; Ellis & Kimmel, 1992). Instead of designating English speaking examinees as the reference group and examinees with foreign language proficiency as a single focal group, each group of examinees speaking the same non-English language would be permitted to exist as an independent focal group. Still yet, additional examples could include large classrooms within a school, schools within a country, and longitudinal differences for an admissions test (Magis, Tuerlinckx, & De Boeck, 2015).

While the intention of this study does not specifically address the causes of DIF, it is important to understand some theoretical causes to underscore the importance of testing for DIF. As an earlier reference, Jenson (1980) posited the cultural difference hypothesis, which described that people from different cultural backgrounds could have differing levels of familiarity with test content. In a similar manner, Meredith and Millsap

(1992) made the argument that manifest variables are not sufficient measures for capturing the latent variables which actually cause DIF. To expand this idea, the manifest variables are correlated with the latent variables. For example, there is not an inherent physiological or psychological difference in intelligence between males and females which causes DIF; rather, there is likely a sociological/cultural phenomenon which reflects gender-normed behaviors and knowledge. Such phenomena are appropriately modeled as latent traits. However, latent variable procedures of DIF detection could suffer from model-data fit and estimation issues.

On the other hand, issues of DIF are not always cultural in nature. Examinees may simply use different metacognitive skills in responding to items (Tatsuoka, Linn, Tatsuoka, & Yamamoto, 1988). Affective domains may also influence DIF, with such an example being females perhaps having an advantage with content involving social relationships (Stricker, 1981). Additional research supports this notion in finding that familiarity, interest, and emotional reactions may serve to be factors which impact item responses (Stricker & Emmerich, 1999). In the context of cross-lingual exams, Benítez and Padilla (2014) used cognitive interviews following quantitative DIF analyses to uncover that DIF may be caused by specific jargon which has different interpretations across languages.

DIF is influenced by and related to other phenomena. Once a DIF item has been flagged, it is important to explore what potentially is causing the item to be biased. Differential distractor functioning (or differential alternative functioning), differential speededness, and differential omission are analyses which can be used to better understand why subgroups of examinees have differential performance (Dorans & Holland, 1993). As an example, Ben-Shakur and Sinai (1991) explored how differential guessing tendencies between males and females were influenced by formula scoring as opposed to number correct scores. Formula scoring was found to provide an advantage to males in both samples they examined (i.e., ninth graders and applicants to Israeli universities).

Among some of the more empirical DIF studies are those which attempt to experimentally cause DIF. There are multiple strengths with these studies. First, they contain typical features of real data that simulation studies do not fully capture. Second, they avoid the model-data congruity issue which faces many simulation studies. Third, different causes of DIF, as well as various categorizations of subpopulations of examinees, can be explored across multiple detection methods. However, there are no guarantees that an item intended to show DIF will do so. For example, a study by Scheuneman (1987) evaluated 16 hypotheses related to experimental (non-scored) test items which were manipulated to cause DIF on the Graduate Record Examination (GRE), and found that just 10 of the hypotheses supported DIF. The manipulations of item features were related to item format, vocabulary in antonyms, wording of the item stem, inference, test wiseness, key placement, and abstraction. Not all manipulations were detected as significant DIF.

Other studies have used experimental manipulation to induce DIF. A 50-item vocabulary test was constructed by Subkoviak, Mack, Ironson, and Craig (1984) which contained 10 items that favored black students over white students. Their findings supported using the area between item characteristic curves (ICCs) corresponding with a unidimensional three-parameter logistic model (U3PL) to detect DIF, when compared with the transformed item difficulty index (Angoff & Ford, 1973; Angoff, 1982) and two chi-squared approaches (Scheuneman, 1979; Camilli, 1979). Kim and Cohen (1991) later reanalyzed the same data to elaborate on IRT-based area measures for detecting DIF. Other researchers have experimentally manipulated features of language to construct DIF items, and found that using an iterative logit method was appropriate for detecting DIF when it was supposed to exist (Kok, Mellenbergh, & Van der Flier, 1985). Still yet, others altered item order within content clusters between males and females, and found significant differences in calibrated IRT difficulties (Plake, Patience, & Whitney, 1988).

To combat the issues surrounding biased test items, the *Standards* (AERA, APA, & NCME, 2014) express the need for analyzing and reporting on issues surrounding item bias and, more specifically, DIF. In fact, chapter three in Part I of the *Standards* is devoted specifically to issues pertinent to fairness in testing. As it relates directly to DIF, suggestions surround the need for preventing construct irrelevant variance at all steps of the testing process (3.0), including minimizing all sources of construct-irrelevant variance which could stem from linguistic, communicative, cognitive, cultural, physical, or other characteristics (3.2), as well as including all subgroups when pilot testing items to screen for bias (3.3). *The Standards* also describe the need for documenting procedures used in evaluating item quality, including screening for DIF among major examinee groups (4.10). Although, relatively few suggestions are provided in how to obtain these goals. In consideration, psychometricians are free to use their professional judgement as to what methods are appropriate to use in a given scenario. Unfortunately, not all methodologies are comparable in performance, even when they are designed for similar situations. Therefore, it is crucial to understand how error can be introduced through choice of methodology alone.

Assuming that error associated with all those aforementioned portions has been successfully mitigated, there are errors potentially introduced by choice of methodology (which is one concern of this study). It is entirely possible that using a particular statistical method to detect DIF may not work properly in certain scenarios. For example, some methods (such as logistic regression; LR) are known to contribute to increased type I errors (i.e., false positives) whenever overall group differences exist in the midst of guessing behavior by examinees (DeMars, 2010). Of particular concern is the level of type II error (i.e., false negatives), which occurs when items that truly exhibit DIF are not detected. Practically speaking, type I errors could potentially lead to good items being removed from exam scoring, while type II errors could potentially allow biased items to be included in determining exam scores.

An item which is flagged according to a statistical criterion does not necessarily mean that the item is biased against subgroups, nor does an item which is deemed appropriate guarantee that the item is not biased. Detecting DIF items is further complicated by data requirements and assumptions with more complex DIF detection techniques. Selecting a method which is too complex may accidentally result in modeling noise along with signal found in data because the model contains more parameters than necessary for describing the data. As such, it is possible that placing statistical constraints (e.g., a lasso constraint) on existing methods may improve performance in these instances.

Determining which items possess DIF in a given test can vary depending on the DIF detection method chosen, given that each method has different assumptions. While these differences are practically non-existent under ideal testing conditions (e.g., large numbers of examinees, excellent model-data fit, and unidimensionality), data which are not ideal will exacerbate differences between the methods. Unfortunately, matters are further complicated by the possibility of multiple types of DIF.

Distinguishing between multiple types of DIF is important because DIF detection methods may be better suited for particular types of DIF. As such, different causes of DIF often result in different types of DIF. When considering item response functions, DIF can be viewed as uniform or non-uniform. Uniform DIF occurs whenever the same group is favored at any level of ability or performance. Stated differently, there is no meaningful interaction between group membership and ability level. This type of DIF is typically associated with a between-group difference in the difficulty of a given item. Uniform DIF frequently occurs if the DIF occurs in the correct response option, though it can also appear in the question stem, distractors, or supporting materials when answering an item.

On the other hand, non-uniform DIF generally refers to when the relative advantage at any given level of ability or performance changes with respect to the other group(s). In other words, there is a meaningful interaction between group membership and ability. Non-uniform DIF can be observed as non-uniform non-crossing DIF or non-uniform crossing DIF. In the former case, typically both the difficulty and discrimination of an item change across groups in such a way that the item response functions do not intersect at any point along the ability continuum. In the latter case, the primary between-group difference is the item discrimination, which causes the item response function to intersect at or near the difficulty of the item. This type of DIF is rarer, and is an interesting find because the relative advantage reverses depending upon what point of the ability continuum is observed. For example, focal group examinees with higher ability may be disadvantaged on the item while examinees with lower ability are advantaged when compared with the reference group.

While multidimensional approaches can be used to account for secondary traits (e.g., SEM and MIRT), such traits which cause DIF are undesirable and cannot justifiably be supported as appropriately entering the item writing process as long as a single score is reported for interpretation and use. Nevertheless, one possibility could be to model a secondary dimension, and base the scoring using only parameters of the primary dimension. However, these modeling procedures can be complex and may not lead to stable estimates. Observed score approaches offer a parsimonious way of detecting DIF, and are used by major testing organizations (e.g., ETS, ACT) even in the present time.

The peculiar nature of detecting DIF is that many statistical tests for doing so analyze one item at a time, making an inherent assumption that there is not contamination in the total score introduced by other items. That is, an assumption is made that all other items except the one under examination are non-DIF items. However, this assumption is not necessarily the case, and is a very strict requirement to meet. Most DIF methods are performed at the item level, but approaches which fit a global model and can estimate item-level parameters is a potential strategy that is less restrictive in the assumptions made on items.

As a way to control for ability level, many DIF models use the observed total score (i.e., a proxy for ability) as a matching criterion to match examinees between groups. This matching helps to lessen the chances that differences observed at the item-level are influenced by differences in group ability (i.e., impact). However, if there are one or more DIF items on an exam, the matching criterion will be contaminated by construct-irrelevant variance. One strategy to potentially improve the matching criterion is item purification (Candell & Drasgow, 1988; Holland & Thayer, 1988; Lautenschlager & Park, 1988; Clauser, Mazor, & Hambleton, 1993; Fidalgo, Mellenbergh, & Muñiz, 2000; Wang & Yeh, 2003; Wang & Su, 2004; as cited in Magis, Beland, Tuerlinckx, & De Boeck, 2010). Item purification is an iterative procedure which removes DIF items from the calculation of a total score or estimation of ability. A DIF method which is calculated for each item individually is first used. Any items with DIF are removed, and the calculations are performed again using the remaining items which were determined to be free of DIF. These steps are continued until none of the remaining items are flagged as having DIF, and the remaining items are used to determine the total score or ability estimate for matching. While purification minimizes issues related to DIF influencing the matching criterion, it introduces an additional confound if many items are removed because there are less data being used to determine the matching variable.

The matching variable, even with improvements or estimation with a latent trait model, is not a perfect representation of ability. Absolute truth cannot be known with a latent trait, nor can it be known in regards to how subgroups of examinees will interact with items. Consequently, it benefits greatly to speculate situations where the truth is assumed be to known, and deviation from truth can be quantitatively measured. Simulation studies accomplish this feat, which allow for absolute manipulation of a constructed reality (Baudrillard, 1994). That is, a study can be conducted which purposefully creates simulated data that contain DIF items, and various methods can be directly compared in how well they correctly identify DIF, as well as fail to recognize it.

Purpose and Rationale

Demographic information is often collected for variables which have more than two groups (e.g., race/ethnicity and language), and being able to explore DIF with these variables provides the exigency of this inquiry. Multiple researchers have asserted that a limitation of most existing DIF methods is that only two groups can be tested (Penfield, 2001; Tutz & Schauberger, 2015; Oshima, Wright, & White, 2015). The purpose of this proposed study is to compare and contrast more traditional multiple group observed score (i.e., non-IRT) DIF detection methods (e.g., generalized Mantel-Haenszel χ^2 and generalized logistic regression) with the more recently developed logistic regression lasso DIF technique (LR lasso DIF; Magis, Tuerlinckx, & De Boeck, 2015). In fact, this purpose was suggested by the authors:

> The method can easily be extended to more than two groups of respondents. It is straightforward to extend the definition of the DIF to any number of groups and to perform lasso penalization onto all DIF parameters for all groups simultaneously. One can then determine on the basis of the lasso approach which items function differently between which groups of respondents. The LR method has been extended to multiple groups' framework before (Magis, Raîche, Béland & Gérard,

> 2011; Magis & De Boeck, 2011), so that it can be used as a basis of comparison (p. 131).

Additionally, this study adds to the literature base through exploring how features of test design, specifically those surrounding information targets, may affect the extent to which DIF items can be correctly identified. A simulation study will be used to demonstrate and summarize scenarios which distinguish between the performances of the methods in detecting true DIF items. The proposed study aims to inform practitioners and researchers of situations where they may find the results useful in judging the merit of adopting the newer lasso method for detecting DIF within multiple groups as opposed to using the pre-existing methods. While several studies have explored detecting DIF in multiple groups, there are no studies to date which explore to use of lasso constraints while detecting DIF among multiple groups using observed score approaches. It is worth mentioning, however, that a recently developed DIF procedure for the Rasch model by Tutz & Schauberger (2015) examined its performance when there are multiple simulated groups. Applying the lasso constraint in the context of multiple groups has not been done for an observed score approach, and this study aims to fill the gap in the literature.

However, observed score approaches have several limitations (Spray, 1989). First, an observed test score is not a perfect representation of an examinee's latent ability, given that the measured scale is not perfectly reliable and is influenced by various sources of measurement error. Second, the observed scores reflect sampling errors. There is no guarantee that the samples for each subgroup are reflective of their respective populations, especially when sample sizes of particular subgroups decreases. Third, because observed scores are summed across individual item scores, items with DIF directly influence the matching variable in an observed score DIF method. Thus, creating a study which manipulates variables related to these limitations advances the understanding of DIF detection in the presence of multiple groups.

Research Questions and Study Variables

This study was guided by two main research questions, each of which is composed of several subquestions. The aim of the first research question was to determine the comparability of the observed score DIF methods with respect to classification accuracy of DIF items. The evaluation of the DIF methods based upon absolute criterion are examined in subquestions *1a* through *1d*, because the ultimate goal of DIF methods is to correctly detect items which are biased against subgroups. These subquestions consider correct classification, type I error, type II error (specified in terms of hit rates), and consistency with truth. Subquestion *1e* concerns relative comparisons among the methods by determining the extent to which they classify DIF items in a similar manner.

The second research question was posited to determine how the methods are directly influenced by changes in types of independent variables which are commonly considered in DIF studies. These subquestions are directly related to the simulation conditions that are manipulated in this study. Specifically, visual inspection of conditional plots can answer what proportion of error can be directly attributed to characteristics at the test-level (e.g., the location of the information target relative to the examinee population and the shape of the test information function; subquestions *2a*, *2b*, respectively), the item-level (e.g., the location of DIF items relative to the information target and the percentage of DIF items; subquestions *2c*, *2d*, respectively), and of simulees (e.g., the amount of impact and sample size; subquestions *2e*, *2f*, respectively). Specifying the research questions in this manner allowed for the interaction between simulation conditions to be examined, as opposed to examining each condition only in isolation. The research questions are explicitly stated as the following:

1. How does the penalized LR DIF detection method (i.e., LR lasso) compare to more traditional non-IRT multiple-group methods (i.e., generalized Mantel-Haenszel χ^2 and generalized logistic regression) as it relates to:
    a. correct classification rate of DIF items?
    b. type I error rate in the classification of DIF items?
    c. hit rates (defined as one minus the type II error rate) in the classification of DIF items?
    d. phi correlations of true and detected DIF items?
    e. agreement statistics among methods?
2. When detecting items that truly exhibit DIF, to what degree is classification error for each analysis model influenced by changes in:
    a. the location of the information target relative to the examinee population?
    b. the shape of the information function?
    c. the location of DIF items relative to the information target?
    d. the percentage of DIF items?

e. the amount of impact?

f. sample size?

Definition of Key Terms and Notation

A list of select terminology and abbreviations is provided in Table 1. The hope is that this table serves as a quick and accessible reference for readers as they encounter unfamiliar abbreviations and to clarify terminology that may be ambiguous due to multiple existing definitions. More detailed descriptions are provided throughout the text of this document where relevant.

Study Organization

A total of five chapters are used to describe this study in-depth. The current chapter was an introduction to DIF and described the importance of this study to the measurement field. In order to frame the study purpose and research questions, a review of relevant DIF literature is summarized in Chapter Two. Chapter Three is used to specify the simulation study design along with the methodologies that will be used to evaluate the results for each research question and subquestion. Chapter Four contains a presentation of the results, accompanied by summary tables and figures. Finally, Chapter Five comprises a discussion of the results, with consideration given to comparisons alongside the DIF literature more generally.

Table 1. List of Selected Terms and Notation Along with Brief Definitions

Term	Description
a_i	item discrimination
b_i	item difficulty
c_i	item lower asymptote for probability of correct response
χ^2	chi-squared
D	scaling constant (i.e., 1.000 or 1.702) used in logistic IRT models
DIF	differential item functioning
ETS	Educational Testing Service
ICC	item characteristic curve
IRF	item response function
IRT	item response theory
GMH	generalized Mantel-Haenszel
GLM	generalized linear model
GLR	generalized Logistic Regression
GRE	Graduate Record Examination
k	number of items
LR	logistic regression
MH	Mantel-Haenszel
N	number of examinees/simulees
$Q1$	first quartile
$Q3$	third quartile
R	right
TCC	test characteristic curve
TIF	test information function
θ_n	examinee/simulee ability level
U2PL	unidimensional two-parameter logistic model
U3PL	unidimensional three-parameter logistic model
W	wrong
X	total score obtained for the entirety of Form X
x_i	item score obtained for item i on Form X

CHAPTER II

LITERATURE REVIEW

The current chapter is organized by first providing a synopsis of observed score approaches to detecting DIF, with descriptions flowing from simpler to more complex models for each of the three analysis models used in this study. Subsequent sections are devoted to providing background research related to conditions which are manipulated later in the simulation study that have been considered in previous studies. Finally, an overview of item response theory (IRT) is provided to inform later discussions surrounding the data generation model. While the emphasis of this study is observed score approaches, some literature from IRT approaches may appear because LR and IRT share similarities under the generalized linear model (GLM).

Overview of Observed Score Multiple Groups DIF Methods

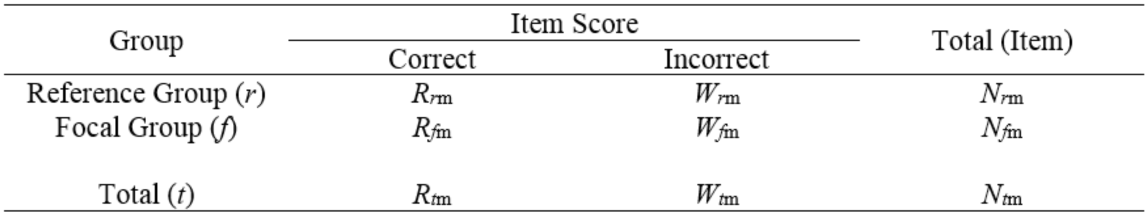
The following section provides a brief overview of the statistical techniques using observed scores to detect DIF items. Readers interested in more detailed coverage are referred to the foundational articles for each of the methods (as provided in each section). The models discussed hereafter are used as the analysis models later in this study.

Mantel-Haenszel and Generalized MH

The development of DIF indices historically has included non-parametric approaches. In educational measurement, two similar approaches based upon χ^2 (chi-squared) were suggested in the late 1970s (Subkoviak, Mack, Ironson, & Craig, 1984). Scheuneman (1979) proposed a procedure similar to χ^2 which used only correct item responses to determine DIF. By conditioning on total score as a proxy for ability, the procedure calculated the probability of a correct response for each possible observed score category. Items with unequal probabilities across score categories were identified as DIF items. In the same year, Camilli (1979) described a χ^2 statistic which used both correct and incorrect responses to reach a very similar statistical test. The strength of these conditioning procedures is that they do not make assumptions regarding the score distributions for each group. However, this type of non-parametric conditioning procedure was actually described decades before.

Mantel and Haenszel (1959), outside of the context of educational statistics and psychometrics, introduced a χ^2 procedure which allowed for a comparison of matched groups. The resulting statistical test is traditionally referred to as Mantel-Haenszel χ^2 (MH). It was later adapted by Holland and Thayer (1988; Holland, 1985) for detecting DIF as a hypothesis test on the constant odds ratio of getting an item correct for two groups across all ability levels.

An item is considered to possess DIF when the MH test statistic exceeds a critical value that is established *a priori*. The calculation of MH is based upon a three-way contingency table (see Figure 1), with dimensions for the matching criterion (typically integer values spanning the range of observed values of the total score), frequencies of correct and incorrect item responses (or score categories in a polytomous case when using a generalized model), and categories (typically group membership for two groups, or more than two groups in a generalized model). The resulting MH statistic follows an asymptotic χ^2 distribution with one degree of freedom (see Equation 1). The MH formula differs from the usual χ^2 formula because the denominator term is not the expectation, and the summation is over the matching criterion as opposed to all observations within a single contingency table. This difference is because MH conceptually (and not algebraically) is summing across individual χ^2 tests conditional on the matching criterion. MH also incorporates Yate's correction for continuity by subtracting 0.5 from the absolute difference in the numerator. To calculate Equation 1, the expectation (see Equation 2) and the variance (see Equation 3) terms are needed.



Group	Item Score		Total (Item)
	Correct	Incorrect	
Reference Group (*r*)	R_{rm}	W_{rm}	N_{rm}
Focal Group (*f*)	R_{fm}	W_{fm}	N_{fm}
Total (*t*)	R_{tm}	W_{tm}	N_{tm}

Figure 1. Contingency Table for Any Given Score Level, *m*. This Figure Has Been Adapted From the One Provided by Dorans & Holland (1993).

The null hypothesis (see Equation 4) states that there is no conditional association (i.e., across all permissible total scores, or bins/strata of score levels in scenarios with smaller sample sizes) between group membership and responding to an item correctly. The alternative hypothesis states that there is a conditional association between group membership and item response. Furthermore, the conditional association is a consistent, unidirectional difference (i.e., uniform DIF).

$$\chi^2 = \frac{\left[\left|\sum_m R_{rm} - \sum_m E(R_{rm})\right| - .5\right]^2}{\sum_m Var(R_{rm})} \tag{1}$$

$$E(R_{rm}) = \frac{N_{rm} R_{tm}}{N_{tm}} \tag{2}$$

$$Var(R_{rm}) = \frac{N_{rm} R_{tm} N_{fm} W_{tm}}{N_{tm}^{2} (N_{tm} - 1)} \tag{3}$$

$$H_0 : \frac{R_{rm} / W_{rm}}{R_{fm} / W_{fm}} = 1 \tag{4}$$

An effect size measure of MH (α_{MH}; see Equation 5) was provided by Mantel and Haenszel (1959). In view of the fact that the interpretation of odds ratios are bounded between zero and positive infinity, the log-odds of α_{MH} (see Equation 6) are typically calculated so that the values are theoretically bounded between negative infinity and positive infinity, and are asymptotically normally distributed (Agresti, 2002). Sometimes, this calculation is linearly translated to the delta metric to ease interpretation of the log-odds. ETS, as well as some other companies, use a transformation of the common-odds ratio for interpretation, a statistic known as MH D-DIF (see Equation 7; Holland & Thayer, 1988; Dorans & Holland, 1993).

$$\alpha_{MH} = \frac{\sum_{m} (R_{rm} * W_{fm}) \Big/ N_{tm}}{\sum_{m} (R_{fm} * W_{rm}) \Big/ N_{tm}} \tag{5}$$

$$\lambda_{MH} = \ln(\alpha_{MH}) \tag{6}$$

$$MH \;\; D - DIF = -2.35 * \lambda_{MH} \tag{7}$$

Significantly positive values of MH D-DIF suggest that an item is biased against the focal group, while negative values of MH D-DIF suggest that the bias is against the reference group. A three-category classification system was developed to describe the magnitude of DIF detected in an item (Dorans & Holland, 1993). The three levels are "A" (negligible DIF), "B" (intermediate DIF), and "C" (large DIF). The absolute values of MH D-DIF, or the MH-LOR, are used along with significance testing to determine the level of DIF that an item exhibits. An item is designated as Level A if |MH D-DIF| is less than 1.0 delta unit (or |MH-LOR| < .426) or |MH D-DIF| is not significantly different from 0. An item is designated as Level C when both |MH D-DIF| is greater than 1.5 delta units (or |MH LOR| ≥ 0.638) and is significantly greater than 1.0. By default, any items which do not belong to Levels A or C are classified into Level B.

The MH approach to detecting DIF makes a few assumptions. First, it inherently assumes unidimensionality of the scale score, given that a single total score is used to perform the matching for the hypothesis test. Second, the direction of bias between the two groups is assumed to be unidirectional across all levels of the matching variable, which means that MH is truly appropriate for detecting uniform DIF only. Third, a hypergeometric assumption is made with regards to the marginal totals. In calculating the expected values for the χ^2 statistic, the marginal totals are assumed to be fixed at a given total score (or stratum). While the χ^2 statistic is non-parametric, the resulting value is sample-dependent. Fourth, it is assumed that the two groups in the test are independent of each other. However, it is often the case that there are shared characteristics or dependencies between the two groups (e.g., males and females may be similar with regards to school, community, culture, and socio-economic status).

In the context of multiple groups, it is not appropriate to conduct a test of DIF for each focal group separately. As Penfield (2001) notes, there are three limitations to doing so: (1) inflated type I error rates; (2) decreased statistical power to detect DIF; and (3) increased run-time and computing resources. Alternatively, it is better to test for DIF among all groups simultaneously. The generalized Mantel-Hanszel procedure (GMH; Mantel & Haenszel, 1959; Somes, 1986) can be used to detect uniform DIF among multiple groups. In addition to better controlling for issues with statistical power, this technique also controls the type I error rate without requiring *post hoc* adjustments to familywise-error rates such as the Bonferonni correction or adjustments to the false discovery rate such as the Benjamini–Hochberg procedure (Kim & Oshima, 2013). Additionally, it also makes no assumptions concerning the cause of the item responses (unlike IRT-based approaches).

GMH is essentially a measure of average partial association in a three-way contingency table. It differs from MH in that it potentially allows for more than two groups and/or polytomous item scores. Furthermore, it is potentially advantageous for use on polytomous data because it does not assume that the data are ordinal, and looks across the distribution of item scores without assuming a particular distributional form. The calculation for GMH χ^2 is given by Equation 8. The GMH χ^2 statistic is chi-squared distributed, with the degrees of freedom under the null hypothesis being equal to one less than the number of demographic groups (assuming dichotomous data, which simplifies the second degree of freedom in the set). Like MH, the null hypothesis under GMH is no conditional association between group membership and response category.

The formula contains bolded letters to indicate vectors of values. The vectors $\mathbf{A}_k$ (see Equation 9) and $\mathbf{E}(\mathbf{A}_k)$ (see Equation 10) have a length one less than the number of groups (i.e., G-1), and $\mathbf{V}(\mathbf{A}_k)$ is a covariance matrix of the same rank (see Equation 11). The vector $\mathbf{A}_k$ is analogous to the R_{rm} term from MH, and it represents the pivotal cells for each level of the matching variable. The expectation of this vector is given by Equation 10, and its variance in Equation 11. The plus sign that is included as a subscript indicates summation over that dimension. A general form of GMH also was given by Landis, Heyman, and Koch (1978) which allows it to more directly simplify to the MH procedure (as cited by Zwick, Donoghue, & Grima, 1993).

$$\boldsymbol{GMH}\,\chi^2 = \left[\sum \boldsymbol{A_k} - \sum \boldsymbol{E}(\boldsymbol{A_k})\right]' \left[\sum \boldsymbol{Var}(\boldsymbol{A_k})\right]^{-1} \left[\sum \boldsymbol{A_k} - \sum \boldsymbol{E}(\boldsymbol{A_k})\right] \qquad (8)$$

$$\boldsymbol{A_k} = \left(n_{11k}, n_{12k}, \cdots, n_{1(G-1)k}\right) \qquad (9)$$

$$\boldsymbol{E}(\boldsymbol{A_k}) = \frac{n_{1+k}\boldsymbol{n}'_k}{n_{++k}} \qquad (10)$$

$$\boldsymbol{V}(\boldsymbol{A_k}) = n_{1+k}n_{0+k}\left(\frac{n_{++k}diag(\boldsymbol{n}_k) - \boldsymbol{n}_k\boldsymbol{n}'_k}{n^2_{++k}(n_{++k}-1)}\right) \qquad (11)$$

Additional advantages of using GMH include increased power under balanced designs, as well as not collapsing focal groups in a manner such that truly differential performance is subsequently undetected. While not considered in this study, GMH has a stronger literature base with applying the procedure to polytomous data, as opposed to multiple groups (Welch & Hoover, 1993; Zwick, Donoghue, & Grima, 1993; Zwick & Thayer, 1996; Chang, Mazzeo, & Roussos, 1996; Zwick, Thayer, & Mazzeo, 1997; Ankenmann, Witt, & Dunbar, 1999; Camilli & Congdon, 1999; Penfield, 2001; Penfield & Algina, 2003; Meyer, Huyah, & Seaman, 2004; Wang & Su, 2004; Su & Wang, 2005; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). The limitations of GMH are that it cannot discern between uniform and nonuniform DIF (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005), the power of the procedure is decreased by smaller sample sizes and impact (Welch & Hoover, 1993, as cited in Penfield & Lam, 2000), and type I errors may possibly be inflated when impact is present (Welch & Hoover, 1993).

Logistic Regression and Generalized LR

As described by Agresti (2002), the generalized linear model can be described as composed of three components: the random component (i.e., the dependent variable and its probability distribution), the systematic component (i.e., the independent variables), and the link function (i.e., the relationship between the independent variables and the dependent variable). When predicting a dichotomous outcome, the GLM can be expressed as a logistic regression. More specifically, a logistic regression is expressed by having mixed effects independent variables (i.e., categorical and/or continuous predictors), a binomial dependent variable, and a logit link. Whenever the GLM is constrained using the logit link it is often referred to as a logit model.

Unlike linear regression, LR makes no assumptions regarding normality, linearity, homogeneity, and normally distributed error terms (Howell, 2010). However, the independent variables are assumed not to have multicollinearity. As usual, the estimation method used makes additional assumptions in addition to those of the model itself. The model parameters of LR cannot be estimated using least squares methods due to the logit link, so maximum likelihood estimation (MLE) methods are typically used to perform the model estimation. Additionally, estimates are solved under MLE using iterations because no closed form solution exists. MLE assumes that data are independently drawn from a multivariate normal distribution (Myung, 2003).

The premise behind LR as a DIF detection method is to predict item responses when using total scores and group membership as predictors. In short, an item is determined to be DIF based upon testing regression coefficients for statistical significance. The first substantial mention of an LR-like approach as a possible DIF detection method (in a non-IRT context) was made in the early-to-mid 1980s (Van der Flier, 1980; Mellenbergh, 1982; Van der Flier, Mellenbergh, Ader, & Wijn, 1984). An iterative logit model was used to correct for the influence that DIF items have on the total score, which is typically a limitation of observed score methods (others have used purification techniques to accomplish the same feat). This method built upon the contingency table approaches by using loglinear models to analyze the data, which is comparable to using the odds ratio estimator in the MH technique. Much like the later LR method for DIF detection (Swaminathan & Rogers, 1990), this method modeled the item difficulty as the intercept, and included parameters for the observed score category and group membership, in addition to an interaction effect of score category with group membership. However, it differed from the later LR method because it treated the observed scores as discrete unordered categories.

More in line with the usual LR framework for dichotomous data with two groups, an LR framework was being studied prior to the more formal conception of the LR method (Spray & Carlson, 1986; Bennett, Rock, & Kaplan, 1987). Swaminathan and Rogers (1990) were the first to provide a detailed model which improved upon IRT methods by reducing issues related to sample size and model-data fit, and improved upon the previous logit models by better accounting for the continuous nature of the ability scale.

Swaminathan and Rogers (1990) further provided a conceptual relationship between MH and LR which involves constraining the LR. Two assumptions must be made. First, the ability variable must be discrete (e.g., observed total scores). Second, there must be no interaction term between the ability variable and group membership, which excludes testing for non-uniform DIF. While this relationship is not exact (given that MH is non-parametric and LR is parametric, and both are based on different assumptions), the hypothesis for uniform DIF is being tested in both.

However, Swaminathan and Rogers (1990) demonstrated that LR was more effective than MH in detecting non-uniform crossing DIF in their foundational paper, particularly across varying test lengths and sample sizes. Using the notation of Magis, Tuerlinckx, and De Boeck (2015) to keep consistency with their model described later, the LR model specified by Swaminathan and Rogers is given by Equation 12.

$$Logit[\Pr(Y_{ijg} = 1)] = \alpha_{0j} + \alpha_{1j} S_i + \alpha_{2j} G_{ig} \tag{12}$$

Using a logit link, the model specifies the probability of a correct response (Y=1) to a single dichotomously scored item, *j*, for examinee *i* belonging to group *g*. The intercept term, $\alpha_{0j,}$ is related to the difficulty of the item. The first logistic regression coefficient, α_{1j}, gives the change in log-odd units of the item-level scores for a single unit increase in the total test score, S_i, of examinee *i*. The second logistic regression coefficient, α_{2j}, describes the change in log-odd units of the item-level scores for a change in group membership from the reference group (0) to the focal group (1). This latter coefficient is of primary importance, because there should be no discernable difference with respect to item performance between the reference and focal groups.

While there are multiple approaches to testing the null hypotheses, primarily two methods have been used in prior studies: the Wald test (Wald, 1939) and the likelihood ratio test. Both approaches are similar in that they can be conceived as being nested model comparisons, and they both share the same asymptotic chi-squared distribution. The Wald test is a significance test of a vector of parameters used to see if each parameter is significantly different from zero. Non-significant parameters can subsequently be omitted from the model. The Wald test was used by Swaminathan and Rogers (1990), which is a good reference for interested readers. As an alternative, the likelihood ratio test, not to be confused with the IRT-based DIF detection technique having the same name (Thissen, Steinberg, & Wainer, 1988), compares null and alternative hypotheses within the nested model comparison. The formula for the likelihood ratio test is provided in Equation 13. In words, Wilks' lambda (Λ; Wilks,

1938) is equal to double the opposite of the natural log ratio of the maximized likelihoods of the nested models, where L_0 is the null model and L_1 is the alternative model.

$$\Lambda = -2 * \ln\left(\frac{L_0}{L_1}\right) \tag{13}$$

Table 2 highlights the comparisons made with respect to the model parameters. The basic model is written in abbreviated form, S + G + S*G, to indicate the predictors for observed score (S), group membership (G), and the interaction term (S*G).

Table 2. Comparison of LR Model Parameters for the Three Null Hypotheses Tested in the Likelihood Ratio Test.

DIF Type	Null	Alternative	Difference in Nested Models
Uniform	S + G	S	G
Non-Uniform	S + G + S*G	S + G	S*G
Both	S + G + S*G	S	G + S*G

Magis, Raîche, Béland, and Gérard (2011) were the first to build upon the LR technique to create a generalized model, namely the GLR method of DIF detection. As noted in their paper, Millsap and Everson (1993) suggested that LR could be expanded into the GLR. Moreover, Van den Noortgate and De Boeck (2005) presented a logistic mixed model capable of considering multiple groups, which was essentially a reformulation of an IRT model. Although flexible, their model involved the estimation of ability, which is circumvented in the GLR because it is an observed score approach. And compared with the GMH, the GLR potentially allows for a more direct detection of non-uniform DIF through a significance test on its interaction term. However, if a test only possesses one or more items with uniform DIF, including the interaction term in the model could potentially lessen the chance that uniform DIF is properly detected.

Equation 14 is the model equation for the GLR, where π_{ig} is the probability that respondent *i* from group *g* responds correctly to a dichotomous item. The reference group is *g* = 0, and non-zero values represent the focal groups. The common intercept and slope are given by α and β, respectively. The specific intercepts and slopes are given by α_g and β_g, where the specific coefficients for the reference group (i.e., α_0 and β_0) are constrained to be equal to zero. This constraint allows the interpretation of the focal group coefficients to be relative to the reference group. Three null hypotheses are included in the model. The null hypothesis for uniform DIF (see Equation 15) is characterized by uniform DIF being present if at least one intercept is significantly different from zero while having all slope parameters equal to zero. On the other hand, the null hypothesis for non-uniform DIF (see Equation 16) states that non-uniform DIF is characterized by at least one slope being significantly different from zero, irrespective of the value of the intercept parameters. Taken together, the null hypothesis for both types of DIF (see Equation 17) requires that all parameters be equal to zero across groups.

$$\boldsymbol{logit}(\pi_{ig}) = \begin{cases} \alpha + \beta S_i & \boldsymbol{if\ g} = 0 \\ (\alpha + \alpha_g) + (\beta + \beta_g) S_i & \boldsymbol{if\ g} \neq 0 \end{cases} \qquad (14)$$

$$H_0 : \alpha_1 = \ldots = \alpha_F = 0 \mid \beta_1 = \ldots = \beta_F = 0 \quad \boldsymbol{(UDIF)} \qquad (15)$$

$$H_0 : \beta_1 = \ldots = \beta_F = 0 \quad \boldsymbol{(NUDIF)} \qquad (16)$$

$$H_0 : \alpha_1 = ... = \alpha_F = \beta_1 = ... = \beta_F = 0 \quad \textbf{\textit{(DIF)}} \tag{17}$$

Seeing as the GLR has three null hypotheses and explicitly tests for non-uniform DIF, it theoretically has a benefit over the GMH. However, a potential limitation of GLR is that increasing the number of groups potentially also increases the type I error rate. Furthermore, the use of maximum likelihood in GLR could potentially be a limitation whenever item scores are subject to variance restriction because of extreme difficulty values.

Logistic Regression Lasso Approach

In linear algebra, vector norms are used to regularize estimation of prediction models. Two common examples of vector norms are the L_1-norm (i.e., lasso) and L_2-norm (i.e., ridge regression), which serve as penalties in regularized estimation of a generalized linear model. Both are used to place constraints on model parameters. A major difference is that the lasso performs the shrinkage of parameters towards zero using absolute values, while ridge regression uses sum of squares to perform the penalization. In doing so, the lasso translates coefficients by a constant factor, while ridge regression scales coefficients by a constant factor. The translation allows the former to successfully obtain values of zero, and permits variable selection through the remaining non-zero coefficients. Ridge regression, on the other hand, does not perform variable selection and is not appropriate for DIF analyses because it cannot discern between predictors at the item-level.

Recall that when the GLR is estimated, it performs the model parameter estimation through an item-by-item basis. The LR lasso DIF method provides a theoretical improvement over the GLR by fitting a generalized linear model to an entire data set. The global nature of LR lasso allows for the relationships between items to be better captured. However, it does not escape the ipsative nature of DIF (i.e., the total score is a property of the test and not an external criterion). LR lasso places the lasso constraints on the variables for each item that describe differences in item performance given group membership. That is, not all items have a meaningful difference in group performance that should be explicitly modeled. Described within the context of LR lasso, logistic regression is a special case of the generalized linear model that can be estimated using lasso regularization. In fewer words, the LR lasso is a lasso penalized version of a generalized logistic regression. The lasso is performed using penalty terms (λ), which cause it to be a shrinkage estimator. Estimated coefficients for covariate terms (e.g., group membership) are multiplied by λ. Other LR terms (e.g., item difficulty and test score) are not influenced by λ.

The LR lasso model (see Equation 18) bears resemblance to the original LR DIF method by having coefficients for test score and group membership. However, the coefficient for test score, S_i, is constrained to be the same for all items for two reasons. First, it circumvents problems with model inconsistency because allowing items to be weighted differently is akin to a weighted sum, and defeats the purpose of using an observed score method where the total score is a sufficient estimate of ability. In this respect, the LR lasso model is more akin to the U1PL model than the U2PL model. Second, allowing different item weights potentially increases the type II error rate (DeMars, 2010, as cited in Magis, Tuerlinckx, & De Boeck, 2015).

$$Logit[\Pr(Y_{ijg}=1)]=\alpha_{0j}+\alpha_1 S_i+\alpha_{2j}G_{ig} \qquad (18)$$

Given that MLE fits the entirety of the data, the penalized log likelihood (see Equation 19) must be maximized with respect to a vector of all parameters (see Equation 20) simultaneously. For model identifiability, a constraint is added so that α_{21} is equal to zero. Given the summation of α_{2j} in the penalized log likelihood, the estimated difference across all groups is multiplied by the penalty parameter, λ. The product of those two terms is referred to as the penalty term.

$$\overset{\wedge}{\boldsymbol{\tau}}(\lambda)=\arg\max l(\boldsymbol{\tau})-\lambda\sum_{j=1}^{J}|\alpha_{2j}| \qquad (19)$$

$$\boldsymbol{\tau}=(\alpha_{01},\ldots,\alpha_{0J},\alpha_1,\alpha_{21},\ldots,\alpha_{2J}) \qquad (20)$$

The optimal λ value can be determined based upon two primary methods. The first is using relative fit statistics describing information criteria (not to be confused with information in IRT), such as Akaike information criterion (AIC; Akaike, 1974), AIC correction for finite samples (AICc; Hurvich & Tsai, 1989; Burnham & Anderson, 2002), Bayesian information criterion (BIC; Schwarz, 1978), Corrected AIC (CAIC; Bozdogan, 1987), Hannan–Quinn information criterion (HQIC; Hannan & Quinn, 1979), or weighted information criterion (WIC; Magis, Tuerlinckx, & De Boeck, 2015). Another method is cross-validation (CV; Hastie et al., 2009). CV splits data into a number of subsets (*k*), and the prediction error is accumulated through *k*-1 iterations in which each subset is removed and the model is refit during each iteration. To provide a comparison, CV is used to select a λ value which minimizes prediction error, while BIC is used to provide the most parsimonious/conservative solution.

The WIC is advocated by Magis, Tuerlinckx, and De Boeck (2015) because it provides an intermediate solution which differs from CV but also falls between the AIC (which is liberal) and the BIC (which is conservative) criteria. They reported that it outperformed the AIC, BIC, and CV criteria under most conditions for percentage of DIF items, DIF magnitude, and sample size and balance. WIC is a weighted average of AIC and BIC that allows the weighting for each to be influenced by characteristics of a given data set, such as sample size and number of items, because of how deviance terms and degrees of freedom are used in the calculation. Equation 21 provides the formula for WIC. The optimal penalty value is found by minimizing the WIC criterion conditional on ω_i, where *i* refers to the weights on an interval inclusive of zero and one.

$$WIC(\lambda \mid \omega_i) = \omega_i * AIC(\lambda) + (1 - \omega_i) * BIC(\lambda) \qquad (21)$$

Item Response Theory and Dichotomous Data

Models

IRT models exist for both unidimensional and multidimensional data, though the former is explicated herein to provide necessary background and justification for the simulation conditions described later. In educational testing, IRT models describe the probability of a correct response for an examinee with a given ability (θ_n) to a particular item. Item response functions (IRFs), plotted as item characteristic curves (ICCs), are used to characterize how the probability of a correct response changes as ability-level changes.

The most widely used models have three or fewer parameters in describing the item properties. These three properties are difficulty (b_i), discrimination (a_i), and lower asymptote (c_i; also known as pseudo-guessing). Defined more specifically, item difficulty is the location on the ability scale where the probability of a correct response equals .5 plus half of the lower-asymptote parameter, and is also the location on the θ scale where the inflection of the ICC occurs. Item discrimination is related to the slope of the item characteristic curve at the point of inflection, and is intended to model the extent to which an item can be used to distinguish between examinees of lower and higher abilities than the item difficulty. The lower asymptote sets a lower bound to the probability space, and is typically used to partially buffer for the impact of guessing in scored responses and improve model-data fit. It represents the probability of a correct response for an examinee with infinitely low ability.

The unidimensional three-parameter logistic model (U3PL) uses all three of the aforementioned item parameters along with θ_n to calculate the probability of a correct response (see Equation 22). A scaling constant, D, of 1.702 has been used historically to allow the cumulative distribution function of the logit model to approximate that of a probit model. However, this practice has largely fallen out of favor, and the scaling constant usually equals one to retain the logit scale. Whenever c_i is constrained to be equal to zero across all items, the model reduces to the unidimensional two-parameter logistic model (U2PL). Additionally, both a_i and c_i can be constrained to one and zero, respectively, to yield the unidimensional one-parameter logistic model (U1PL). If the scaling constant (D) is set to unity also, it reduces mathematically to the Rasch model. The advantages of these unidimensional models are that they have simpler mathematical forms. However, they generally fail to capture the complex interaction among persons and items, which may actually involve a set of traits instead of a single ability (Reckase, 2009). This recapitulates the philosophy held by Box and Draper (1987) that, "Essentially, all models are wrong, but some are useful" (p. 424).

$$P_{ni}(x_{ni}=1|\theta_n,\alpha_i,b_i,c_i)=c_i+(1-c_i)\frac{\exp(D\,a_i(\theta_n-b_i))}{1+\exp(D\,a_i(\theta_n-b_i))} \tag{22}$$

Conceptually, IRT models are a special case of the GLM, as their notation can be expressed through transformations in slope-intercept form. In fact, changing the form of the model to other equivalent expressions can change the interpretation of the lower-asymptote parameter (von Davier, 2009). Through the current parameterization of the U3PL, it is assumed that examinees with higher ability levels do not "slip," which is incorrectly responding to an item when truly possessing the knowledge to answer it correctly (i.e., an accidental or careless mistake). However, the absence of this phenomenon being explicitly modeled in the U3PL does not mean that slipping does not occur. IRT models are employed to represent the probabilistic responses of examinees, and are descriptive instead of prescriptive.

In that respect, IRT models are a smoothing function of the empirical ICC (Petersen, Cook, & Stocking, 1983), where there is better model-data fit as the two approach the same shape. Better fit results from having less residual variance which has not been explained by the IRT model. For all the models explained above, there are additional assumptions for an upper asymptote of one and for perfect symmetry of the IRF. While having more parameters usually lends itself to describing data better, consideration must ultimately be given to sample size as it influences the likelihood of obtaining stability through convergence of item parameter estimates.

Assumptions

Assumptions of unidimensional IRT models include local independence, monotonicity, model-data fit, invariance of item parameters over comparable examinee samples, non-speededness, and a causal relationship between θ_n and item responses (De Ayala, 2009; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Under the Rasch model, additional assumptions of sufficiency and specific objectivity (sometimes called objective measurement) are acknowledged. Whenever IRT models are applied to datasets in operational settings, certain assumptions are made regarding additional concepts, which include treating the standard error of estimate (SEE) around item parameter estimates as negligible when estimating θ_n so long as the SEE values are reasonably low given the sample size.

These assumptions are important to note, because the appropriateness of a model is determined by the extent to which data corresponds with the assumptions that are made. Substantial violations of these assumptions will likely result in degradation of how the model performs with respect to accurately describing the data. For instance, violations of invariance can result in differential item functioning (DIF) across subgroups of examinees.

Specifying Test Information Targets

Test specification is the process of selecting items according to various constraints to build one or more test forms. The purpose of test specification is to provide guidelines to better ensure that various properties of the constructed exam forms are met, as well as provide the best possible estimates of ability for examinees (especially near cut points). Oftentimes, testing organizations use a content blueprint to guide item selection to ensure that the test is representative of the construct being measured. Ideally, a content blueprint is developed in such a way that every domain of a given set of content standards (or set of knowledge, skills, and abilities) are sampled adequately.

Linear programming techniques are generally used as a means of building exam forms through optimizing objective functions (van der Linden, 2005). Statistical specifications are considered in this process and can be specified at various levels, such as the item level, stimulus level, item-set level, subtest level, test level, or multiple tests level. More specifically, constraints related to psychometric properties of the items are chosen to fall within specified ranges, with some examples being p-values and biserial (or point-biserial) correlations, word counts, response time, and the exclusion of enemy items on the same form. Exam-level features are sought after as well. Some examples may include test length, minimum reliability, and test information targets.

Test information targets are particularly important because they directly correspond with cut points used in test use and interpretation. It is customary practice to build exams so that there is a peak in the test information function at the location on the ability scale where decisions are being made about examinees (Hambleton &

Swaminathan, 1985, p. 104-115). A convenient feature of a TIF is that the information contributed by individual items is additive. Stated simply, the sum of the item information functions (IIFs) is equal to the TIF for each possible value of θ_n. Individual items cannot be studied in this manner under CTT because it is strictly a property which arises with IRT modeling. The discrimination parameters of items have more of an influence on the shape of a TIF when compared with item difficulty parameters. Hence, it is often preferred that items with better discrimination are selected during form building, because information has an inversely proportional relationship with the standard errors of estimate for θ_n.

Though, the values associated with a TIF cannot be readily interpreted as meaningful. The θ metric itself inherently lacks meaning and is statistically indeterminate. Subsequently, the choice of scaling for the θ metric directly influences the shape of a TIF. Scaling changes in the discrimination parameters also change the shape of a TIF. Despite an immediate intuitive understanding of TIF values being rather difficult to obtain, it is important to underscore the importance of the relative location of maximal information when compared with the intended cut points. If a TIF is not aligned with the intended cut point location(s), then test development has been misspecified with respect to its intended use and interpretation.

Person Characteristics

Sample Size

Historically, DIF studies have explored the influence of small and/or unbalanced sample sizes on the effectiveness of detecting DIF items. In general, performance degradation occurs as sample size decreases. Unbalanced designs often contain smaller samples for specific groups, which impacts the overall analysis. Studies have varied in suggesting a minimum sample size required for adequate power to detect DIF. One study reported the lack of statistical power to adequately detect biased items using a sample size of 300 within two-group dichotomous data (Candell & Drasgow, 1988). Others have suggested that 200 is a stable lower bound (Narayanan & Swaminathan, 1996). Still yet, sample sizes as low as 147 have been examined before using Mantel-Haenszel, but with conclusions related to instability of the estimates (Ryan, 1991). In the context of multiple groups DIF, Welch and Schauberger (2015) used 250 as a minimum sample size when calibrating with their Rasch Lasso approach.

As such, observed score methods do not have the additional burden of estimating person ability, which allows the detection of DIF to be less demanding with respect to sample size requirements. This benefit was evidenced by Magis, Tuerlinckx, and De Boeck (2015) when they showed that 100 examinees in a two-group case had adequate DIF identification for both the LR and LR lasso methods. However, certain factors have a more detrimental influence when they occur in the midst of small sample sizes. The presence of impact may require that the minimum sample size be increased to 500 examinees to avoid issues with increased type I error (Welch & Hoover, 1993). Moreover, items within the same exam having different item-total correlations (or discriminations) can inflate type I error, especially for MH (Roussos & Stout, 1996).

Additional strategies can be used in cases when smaller sample sizes exist. McLaughlin and Drasgow (1987) advocated adjusting the significance level of hypothesis tests to obtain a more rigorous nominal level rejection rate. As mentioned previously, thick matching can be used to circumvent having too few observations at any given score level.

Impact

Impact (i.e., group differences in underlying ability or group differences in observed score means) has received considerable attention in DIF literature because of its potential influence on obfuscating the detection of DIF items. Methods which use matching variables are hampered because the increased distance between the distributions of values for groups effectively leads to floor and ceiling values that potentially do not have matches. Furthermore, methods using prediction models would seemingly be less influenced by impact because the estimated parameters are robust and arguably sample-independent. However, the estimation methods used in those methods can make distributional assumptions and may have convergence issues as data become more extreme. Examinee test-taking strategies not only impose construct-irrelevant variance, but are perhaps different across different levels of ability. In the presence of impact, students who are part of groups with lower performance may have higher guessing tendencies (Uttaro & Millsap, 1994). In general, DIF analyses are performed using discrete outcome measures (which are typically scored dichotomously), meaning that there are less numeric information available, leading to additional complexity in distinguishing between the influence of impact and DIF on scored data.

Impact has been found to increase type I error rates for both Mantel-Haenszel (Holland & Thayer, 1988; Clauser, Mazor, & Hambleton, 1993; Welch & Hoover, 1993;

Penfield, 2000; Fidalgo, Ferreres, & Muniz, 2004; Li, Brooks, & Johanson, 2012) and logistic regression methods (Li & Stout, 1996; Narayanan and Swaminathan, 1996; Whitmore & Schumacker, 1999; DeMars, 2009; Güler & Penfield, 2009; Li, Brooks, & Johanson, 2012). In fact, simulation studies can amplify the influence of impact on observed score techniques by simultaneously using non-Rasch data generation models (Roussos & Stout, 1996). Though, having at least 40 items may not have as large of an influence compared with shorter tests (Uttaro & Millsap, 1994).

Impact has a slightly stronger influence on GMH when compared with MH when there are only two groups, with the distribution of χ^2 values obtained for the former increasing more than the latter (Zwick, Donoghue, & Grima, 1993). That is, given the same level of impact between two groups when analyzing dichotomous data, the GMH test statistic is slightly more prone to type I error than MH. On the other hand, if there are more than two groups being analyzed, GMH better controls for type I error than MH does (Penfield, 2001). In short, collapsing multiple focal groups may lead to increased error. Tian (1999; as cited in Kristjansson, Aylesworth, McDowell, & Zumbo, 2005) also found that GMH had increased type I error in the presence of impact. Finch (2016) found increased type I error rates for both GMH and GLR across conditions for sample size, level of DIF, and the presence of impact. The results also suggested that GMH may have superior performance in detecting DIF when compared to GLR when there are more than three unbalanced groups being tested simultaneously. Moreover, LR lasso was found to be promising for DIF detection in two groups when compared with LR in the presence of impact (Magis, Tuerlinckx, & De Boeck, 2015).

CHAPTER III

DATA AND METHODOLOGY

Simulation Design and Conditions

Given that the overall purpose of this study is to compare observed score DIF approaches for multiple groups, a simulation study is employed to determine the extent to which “truth” can be recovered across three different methods. The strength of a simulation approach is that the performance of each method can be evaluated against a known criterion (i.e., detected DIF items compared to generated DIF items), treating the generated parameters as absolute truth. The limitation of simulation studies is that they cannot perfectly capture the reality of the testing process and the nuances of real data, regardless of how many precautions are taken with writing code and stipulating conditions. Thus, simulation studies are not intended to prove any concepts, but rather to build an argument for situations where degradation of methodological performance can or cannot be readily observed. Furthermore, the robustness of models and estimators can be tested without any adverse impact on examinees. The term simulees is specified hereafter to refer to simulated examinees. The entire simulation study was conducted in the R programming environment (Revolution R Enterprise version 8.0 – 64 bit; R Core Team, 2015; Microsoft Corporation, 2015), with some additional analyses performed in SAS 9.4 (SAS Institute, 2012) to further analyze the results.

Error is defined as any discrepancy between predicted DIF status and actual DIF status in this case, and this study offers insight into which observed score approach for detecting DIF among multiple groups has the best recovery given data that measure different features related to test information targets, as well as item-level characteristics and simulee population characteristics. Given the extent to which the simulated data mimic real data, the results potentially can be generalized to situations where non-IRT methodology is used because there are many conditions explored in the study. As a caveat, McLaughlin and Drasgow (1987) noted that generalizing the results of studies are often limited to samples containing normally distributed abilities, which is the case in this study. Plus, strict normality is difficult to find in practice (Micceri, 1989).

Specified in Table 3 is a summary of the conditions which were manipulated by the researcher. The summary table of simulation conditions allows for study variables to be manipulated at the test-level (e.g., the location of the information target relative to the central tendency of the examinee population, and the shape of the test information function), item-level (e.g., the location of DIF items relative to the information target, and the percentage of DIF items), and for simulees (e.g., the amount of impact and sample size). These conditions are likely to have an influence on predicting which items possess DIF above and beyond the error that results in detecting those items even in ideal situations. The relative lack of literature which explores DIF as it relates to target test information functions provided the exigency for exploring it within this study, along with its typical absence in literature using IRT generation models. Practitioners may find the results useful in judging the merit of adopting the newer lasso method for detecting DIF within multiple groups as opposed to pre-existing methods. Furthermore, the test design features of this study allow for the interpretation to be less theoretical in nature and better aligned with standard operational practices, such as building exams to be optimized at test information targets, for example.

All conditions were fully crossed for each analysis model. Constants held across all conditions were a total form length of 40 items (Donoghue & Allen, 1993; Raju, van der Linden, & Fleer, 1995; Fidalgo, Ferreres, & Muñiz, 2004), no missing data, and 250 replications for each crossing of conditions. The rationale for selecting 40 items is that the reliability of the scale scores should not present an additional confound into the study, while still allowing for integer values to be obtained for the percentage of DIF items. Also, an exam consisting of 40 items presents itself as an acceptable lower bound for the number of scored items that may be included on an achievement test (e.g., an end-of-grade exam for third grade students). Four groups have been used previously in studying DIF across multiple groups (Stark, Chernyshenko, & Drasgow, 2004; Magis, Raîche, Béland & Gérard, 2011). However, the simulation does not consider issues related to test speededness being observed, though it is acknowledged that some proportion of DIF could arise from speededness of a test. A detailed description of the conditions in the table and their corresponding levels is provided throughout the next section.

Data Generation

The simulated data sets (already assumed to be scored) were generated using a unidimensional three parameter logistic IRT (U3PL) model (see Equation 22). In the U3PL, the lower asymptote can model guessing behavior to some extent, which is a type of construct irrelevant variance (Wright, 1991). The incorporation of the lower asymptote parameter more closely resembles the complex reality of assessment data, wherein

Table 3. Summary of Simulation Conditions Including the List and Count of Levels for Each Condition.

Condition	Levels	Totals	Literature/Rationale
Test Type	• Commensurate $[TIF_{max} - \bar{\theta} \approx 0]$ • Disparate $[TIF_{max} - \bar{\theta} \approx$ -1.15]	2	Roughly 87.5% pass rate for disparate condition, which is comparable to first-time test-taker results (e.g., ABIM, 2015).

TIF Shape	• Spread
	rtruncnorm(k, min=-2, max=2, mean = *x*, sd = 1.5)
	• Narrow
	rtruncnorm(k, min=-2, max=2, mean = *x*, sd = .5)
		2	Truncation range same as:
			DeMars (2009).
			SD of *b*-parameters is manipulated to reflect the amount of precision near the information target.

DIF Location	• Near Information Target [Select items closest to target to be DIF items] • Offset above Information Target [Select items closest to +1 logit above the target to be DIF items]	2	As a practical example, DIF may be induced by new item types or revised content standards that subsequently causes the DIF item to be more difficult than the other items measured on a given form.
Percentage of DIF Items	• 0 % (0 items) • 5 % (2 items) • 10 % (4 items)	3	Exactly as: Magis, Tuerlinckx, & De Boeck (2015)
Impact ($\bar{\theta} + \varepsilon$, in logit units to group means)	• {0, 0, 0, 0} • {.00, -.17, -.33, -.50}	2	Same total range as: Jodoin & Gierl (2001). First group is the reference group.
Sample Size	• Balanced: {500, 500, 500, 500} • Unbalanced: {800, 600, 400, 200}	2	Total sample size used in: Magis, Tuerlinckx, & De Boeck (2015); French & Miller (1996). Minimum for unbalanced taken from: Narayanan & Swaminathan (1996).
SUBTOTAL	-	96	-
Analysis Models	• Generalized Mantel-Haenzsel χ^2 • Generalized Logistic Regression • LR lasso DIF	3	-
TOTAL	-	288	-

*Constants: Total test length of 40 items (Donoghue & Allen, 1993; Raju, van der Linden, & Fleer, 1995; Fidalgo, Ferreres, & Muñiz, 2004). 250 replications for each crossing of conditions.

response variability is more difficult to predict with lower performing examinees when item types and scoring processes allow for guessing behavior to be advantageous potentially. However, DIF detection is more likely to result in false positives when guessing behavior influences responses, especially in the presence of large group impact (DeMars, 2010), which ultimately hinders the interpretation of variability in errors attributed to impact. For this reason, study conditions include crossings where there are no DIF items but impact in examinees in order to have a baseline for comparison. Though not addressed in this study, it may be entirely possible that guessing behavior is positively correlated with examinee ability, with higher ability examinees being more likely to correctly guess through eliminating incorrect responses.

While generating the data sets, all the relevant changes/shifts to item parameters were considered in calculating the probability of a correct response for each item and examinee per the U3PL. This matrix of model-implied probabilities was compared against a matrix of the same size containing random values from the uniform distribution, $U(0,1)$. If the probability value in each cell was lower than the random uniform value in the same cell location in the comparison matrix, the scored item response was a zero. Otherwise, if the probability was higher than the corresponding random value, the scored item response was a one. This procedure is commonly used to prevent a deterministic model of data generation. Data were generated for each of four groups separately within a given replication, with the data being concatenated into a single data set for subsequent analysis. Ability parameters were sampled from a standard normal distribution for each group, with any subsequent level of impact added to the ability parameters for each group after having been sampled.

Test-Level Conditions

Careful consideration must be given in selecting the U3PL generating parameters, as the resulting data must be at least adequate for use with the non-IRT analysis models. The true b_i parameters were drawn from a truncated random normal distribution bounded between -2 and +2 logits (DeMars, 2009) of the intended test information target. This truncation was done largely with three reasons related to the observed score analysis models. First, it helped in preventing completely homogenous response patterns (e.g., all zeros or all ones). Second, it minimized the chance that a "difficulty" dimension would impact the results, because items with extreme difficulty parameters often have smaller variance when compared to other items on an exam. Extreme difficulty values would be more likely to impact procedures involving estimation (i.e., GLR and LR lasso) as opposed to calculations (GMH), thus presenting a bias which would not be purposefully manipulated through the study conditions. Third, the truncation better mimics an operational item selection process which typically avoids items with extreme difficulties (i.e., items with extreme p-values) and favors items closer to the intended cut score locations.

The standard deviation of the difficulty parameters was directly manipulated to influence the shape of the target information function. Given that the moments chosen for the theoretical truncated distribution will not directly match that of a sampling distribution (which is not truncated), standard deviation values of 1.5 and 0.5 where chosen in the generating distribution to subsequently yield empirical standard deviation values of 1.0 and 0.5, respectively, in the difficulty parameters on average.

The intended TIF target was also changed as a condition to mimic tests which have different purposes. Commensurate targets correspond with the central tendency of the examinee population, while disparate targets correspond with a non-central portion of the examinee population. Oversimplified examples would be achievement versus certification/licensure tests, respectively. However, it should be noted that certification/licensure tests are very rarely 40 items long, so cautions must be taken in such an overly simplistic interpretation, though the examples are convenient for describing instances in which TIF targets may change depending upon the type of exam. In this study, the TIF targets were set with respect to the average ability value of a standard normally distributed population of simulees. To clarify, two levels of this condition were used. First, the theoretical distribution of difficulty values was centered around zero to create a commensurate target. Second, the theoretical distribution of difficulty values was centered at -1.15 to create a disparate target. Under the standard normal distribution, roughly 87.5% of the area underneath the curve can be found at or above -1.15. This simulated level of pass/fail rate is comparable to first-time test-taker results (e.g., ABIM, 2015).

The discrimination parameters were simulated to follow a lognormal distribution with a resulting mean of 1.00 and standard deviation of .1225 (Donoghue & Allen, 1993; Penfield, 2001). To control for "item quality" within each replication of the simulation, the same discrimination parameters were used for both the commensurate and disparate testing scenarios. This constraint prevented the item discrimination from directly influencing the DIF detection not only across testing scenarios, but did not allow for the GLR and LR Lasso DIF procedures to have an inherent advantage over GMH, given the former better captures discrimination through estimation.

The lower asymptote parameters were not simulated based upon a distributional assumption, but rather were set constant at 0.20 for all items based upon a consistent practice established in literature (Lautenschlager & Park, 1988; Park & Lautenschlager, 1990; Mazor, Clauser, & Hambleton, 1992; Donoghue & Allen, 1993; Rogers & Swaminathan, 1993; Uttaro & Millsap, 1994; Allen & Donoghue, 1996; Narayanan & Swaminathan, 1996; Marañón, Garcia, & Costas, 1997; Penny & Johnson, 1999; Fidalgo, Mellenbergh, & Munoz, 2000; Penfield, 2001; Jodoin & Gierl, 2001; Hidalgo & Perez-Pina, 2004; Wang & Su, 2004; Herrera & Gómez, 2008; DeMars, 2009; Güler & Penfield, 2009; DeMars, 2010).

Item-Level Conditions

Whenever DIF was introduced into the simulated data sets (that is, when the percentage of DIF items was a non-zero value), the magnitude of DIF was held constant at a total range of 0.8 logit units across the four groups (Rogers & Swaminathan, 1993; Penfield, 2001; Magis, Tuerlinckx, & De Boeck, 2015; Finch, 2016). To elaborate, the following values were added to the difficulty parameters for the four groups on the DIF items (once again, assuming the first group is the reference group): .00, .27, .53, and .80. The non-zero values were added to the three focal groups randomly across replications to more closely reflect the reality that certain subgroups are not always more disadvantaged than others in the presence of DIF. However, the reference group remained the same across all replications, given that the reference group often does not change in standard operational practice in many settings. The approach of adding DIF to item parameters is like that done in other DIF studies (Lim & Drasgow, 1990; Kim & Cohen, 1991; Miller & Oshima, 1992; Gomez-Benito & Navas-Ara, 2000). The primary concern of this study is uniform DIF that is unidirectional/asymmetric with respect to groups (these constraints could be relaxed in future studies to explore non-uniform DIF and/or symmetric DIF). Considering the impact condition explained above, the focal groups which would potentially have lower observed performance have an increasing disadvantage on the DIF items.

The percentage of DIF items was manipulated as an item-level condition. Three levels were chosen for this study: 0% (i.e., no items), 5% (i.e., two items), and 10% (i.e., four items). These levels were also used by Magis, Tuerlinckx, and De Boeck (2015). Historically, changing the percentage of DIF items has been studied in the DIF literature. Having no DIF items informed a baseline type I error rate under various crossings of conditions, while having up to 10% of an exam better informs type II error rates. In contrast, increasing the number of DIF items contaminates the matching variable (i.e., the total score) and may affect both type I error and power.

The location of the DIF items relative to the TIF target was explored to see if there would be a difference between the methods. A calculated statistic like GMH may perform differently amid variance restriction of item scores when compared with a globally fitted model like the LR Lasso. As a practical example of how DIF location may be impacted, DIF may be induced by new item types or revised content standards that subsequently causes the DIF item to be more difficult than the other items measured on a given form. As such, a new technology-enhanced item could introduce construct-irrelevant variance that causes an item simultaneously to be more difficult and exhibit DIF. To introduce DIF location into simulated data within this study, two levels were considered: items near the TIF target, and items offset from the TIF target. In conditions with non-zero DIF items, the relevant number of items closest to the observed information target (and not the theoretical distribution, to account for sampling fluctuations) were chosen to exhibit DIF. Likewise, for the items offset from the TIF target, the number of simulated DIF items closest to +1 logit unit above the TIF target exhibited DIF. It was not apparent that any previous DIF studies have explored DIF location relative to the TIF target, so this condition contributes novel information to the literature base.

Simulee Conditions

For simulees, a total sample size of 2000 was used as it represents a stable lower bound (French & Miller, 1996; Magis, Tuerlinckx, & De Boeck, 2015). The total sample size was distributed among four groups (Stark, Chernyshenko, & Drasgow, 2004). A standard normal distribution, $N(0,1)$, was used to generate ability parameters for both the commensurate and disparate testing scenarios, though different samples were generated for them within each replication because the samples of simulees are assumed to belong to inherently different populations. In the presence of group-level impact, the four groups were assumed to have different means for the normal distribution. More precisely, the distributions for the groups were as follows (with the first group being the reference group): *N*(0,1), *N*(-.17,1), *N*(-.33,1), and *N*(-.50,1). The values in the reporting herein are truncated at two decimal places, but precise fractions were used in the actual code. This range of a half logit is the same as that used by Jodoin and Gierl (2001) and others (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Finch & French, 2007; Paek, 2010; Finch, 2016). It is common for DIF studies to use a one logit range for impact, but doing so within the scope of this study would frequently lead to perfect response patterns in the conditions where shifts in items are already occurring due to the presence of DIF, test type, and DIF location, in addition to the shifts in ability distributions in the presence of impact.

While the total sample size was not manipulated, the balance of the sample size was changed as a condition. A balanced design was specified so that each of the four groups had 500 simulees, in contrast to an unbalanced design where the four groups had 800, 600, 400, and 200 simulees, respectively. The former could represent a case where an exam is administered in multiple languages, while the latter could represent different racial/ethnic groups. Moreover, the balanced design represented an ideal case statistically (with regards to statistical power, as derived for dichotomous LR DIF detection by Li, 2015), though the unbalanced design was considered because of practical limitations which occur frequently. The minimum sample size of 200 in the unbalanced case was designated due to a finding by Narayanan and Swaminathan (1996) that supported using MH only if that sample size requirement was met in a focal group. Other studies support various minimum sample sizes (e.g., Güler & Penfield, 2009, suggest 200 to 250;

Swaminathan & Rogers, 1990, suggest 250), which is directly attributable to different bin sizes used in the matching variable in MH. In the case of thin matching, the number of bins equals the number of possible scores on an exam (i.e., the number of dichotomous items plus one). Only thin matching was explored in this study; thus 41 bins were used in the GMH procedure.

Analysis Models

Item purification was not performed with the analysis models. Lasso constraints inherently are a selection procedure, so allowing item purification would add an additional confound to the study because the results would be impacted by different selection procedures. Moreover, the intention of this study is not to study selection procedures, but to examine the benefit of using the lasso against a baseline (which is not having a selection/purification process in a multiple group setting). However, such a comparison would likely provide a fruitful investigation as a future study.

The R package *difR* (Magis, Beland, Tuerlinckx, & De Boeck, 2010) was used to calculate the GMH statistics within each replication. Anchor items were not provided to the function (i.e., the total score was used as the matching criterion variable). Given a nominal α level of .05, the one-tailed hypothesis test was calculated using a threshold on a chi-squared distribution with the degrees of freedom equal to the total number of groups minus one (i.e., the number of focal groups in the analysis). At three degrees of freedom, the critical value is roughly 7.815.

The R package *difR* was also used to estimate the GLR statistics within each replication. As it was with the other analyses, anchor items were not provided to the function. Significance testing was based upon the likelihood ratio test, given that it is less impacted by estimation issues from smaller sample sizes when compared with the Wald test of significance (Agresti, 2002; as cited in Magis, Raîche, Béland, & Gérard, 2011). Additionally, *post hoc* group-level comparisons for significant findings were not required in this study, because DIF identification is of primary concern. Both uniform and non-uniform DIF were tested (unlike the comparison made in Magis, Tuerlinckx, & De Boeck, 2015) because mostly identical results would be obtained between GMH and GLR if uniform DIF were only being tested, and because the benefit of GLR is that practitioners can simultaneously test for both types of DIF.

The R code for LR Lasso DIF used by Magis, Tuerlinckx, and De Boeck (2015) was obtained through personal communication with Magis. Detailed description of the functions appears in the unpublished appendix of their paper. In summary, the code makes a call to the R package *glmnet* (Friedman, Hastie, & Tibshirani, 2010) to fit the lasso penalized logistic regression to a given data set in long-format. Additional functions are provided to obtain the optimal λ value given an information or CV criterion. For this study, the WIC was used and determined across 1000 weight values ranging from zero to one (i.e., analyzed at increments of .001). The code was adapted by the author of this study to accommodate for multiple groups.

The overall shell of the R code for the study was written in a way to ensure that the full set of 250 replications was met across all conditions and analysis models. In keeping with many programming paradigms, a *main* function was written to run all the data generation and analyses of a given replication. To decrease the run time of the entire simulation study, the main function was optimized with byte code compilation using the *compiler* package that is included in base R. The compiled main function was then placed inside of a *try* statement to prevent the study from crashing in the event of errors. The single line of code (i.e., the compiled main function surrounded by a try statement) was iterated through a *repeat* loop with logic evaluation of the replication count with a break from the loop occurring only after a successful run of a given replication. Writing the code in this manner ensured that all 250 replications contained only successful analyses. Subsequently, *for* loops were used to iterate through all possible combinations of condition levels. Global scoping assignment (i.e., the double arrow assignment operator) was used inside of the compiled *main* function to keep counts of replications to avoid variable confounding.

Evaluation of Results

The first research question was analyzed across all 250 replications for each crossing of conditions. The correct classification rate (CCR) of DIF items was determined by the instances in which an item that did not exhibit DIF in the generating parameters was deemed to be a non-DIF item by a given analysis model, as well as the instances in which an item that exhibited DIF in the generating parameters was deemed to be a DIF item by the analysis model. Given that each replication was composed of 40 items for a simulated exam, the overall proportion reflects the proportion of the 10,000 (i.e., 250 times 40) items within a given crossing of conditions which were correctly classified.

Type I error rates were determined to be the proportion of items which were not generated to have DIF, but were indicated as having DIF by a given analysis model. In other words, it is the false positive rate for DIF identification. On the other hand, hit rates (also referred to as power) were defined as one minus the proportion of type II errors (i.e., false negatives) made by a given analysis model. That is, the type II errors are the proportion of items where DIF was induced in the generating difficulty parameters, but the analysis model failed to identify those items as having DIF. In general, type II errors in DIF detection can be argued as possessing a greater threat to validity than type I errors, because DIF items which are not detected can influence total scores (or ability estimates) by introducing construct irrelevant variance. For this reason, more emphasis is placed upon hit rates than type I errors in the interpretation of results. Phi correlations were computed within each replication as an additional measure of the success of each method in correctly predicting DIF items, given that they consider CCRs, type I errors, and type II errors simultaneously in one measure of association that's appropriate for data which truly are dichotomous (which corresponds with the assumption made by the generation model as well as the typically binary nature of DIF flagging).

Agreement statistics were also computed on the correct classification rates (CCRs) across replications as a relative comparison between methods to further answer research question one. Namely, these statistics were unweighted κ (kappa), weighted κ, percent exact agreement, percent adjacent agreement, and combined agreement. To further clarify, a quadratic weight was used in weighted κ so that more flagrant disagreements between the analysis models were more heavily penalized. Also, the percent exact agreement of CCRs was calculated as the percentage of times in which the number of correctly classified items perfectly matched between analysis models. In other words, it is the percentage of replications where two methods had the same CCR.

To clarify, this statistic does not consider if any two analysis models are correctly identifying the same items, just only the same number of items. Such an inquiry is not fully warranted because the data are simulated and the cause cannot be readily identified. Additionally, the percentage of adjacent agreement for CCRs captures scenarios in which the number of correctly classified items differ only by one item. Subsequently, the combined percentage agreement is the sum of exact and adjacent agreement. The combined agreement provides a general idea of the extent to which different analysis models are reaching similar conclusions with only minor differences. In summary, research subquestions 1*a* through 1*d* were calculated within replications, while research subquestion 1*e* was calculated across replications.

To answer research question two, comparisons of the three DIF detection methods were analyzed marginally with respect to the six manipulated conditions. Such comparisons included comparative boxplots to inspect error variability with respect to CCRs, as well as line graphs to analyze type II error rates while conditioning on type I error rates. The boxplots included information across all replications (by levels within condition), but the line graphs only included replications where any percentage of DIF items were present. (i.e., replications containing data without true DIF items were excluded to better represent type II errors). Direct comparisons of type II error rates are more trustworthy when the amount of type I error is comparable between procedures, given that the type II error rate of a given procedure is decreased unduly by inflated false positive rates. By examining fidelity of the procedures with respect to levels of the manipulated conditions, it is possible to generalize the findings at a larger grain-size (though, not broad-sweeping generalizations that would be aptly applicable to all the many nuanced scenarios practitioners encounter). In doing so, it is possible to obtain a general sense of the relative prevalence of the conditions in causing error in detecting DIF items. Operational practices would benefit from this knowledge because it provides guidance as to which testing conditions may potentially inhibit the detection of DIF items.

Descriptive statistics and visual inspection were also used to note relationships among the true classification of DIF items and the magnitude of the DIF statistics, in a manner similar to what was done by Penny and Johnson (1999). Furthermore, as inspired by Dorans and Kulick (1986), conditional plots were generated to better explain the relationship between classification accuracy of DIF items and differences in test information targets and the examinee location. Analyzing conditional relationships are crucial to informing the extent to which even seemingly smaller differences in the TIF and examinees could impact classification accuracy.

CHAPTER IV
RESULTS

The simulation study results are presented by research question. The findings for research question one (RQ1) are explained by the crossing of study conditions for correct classification rates, type I error rates, hit rates, and correlations between generated and detected DIF items. These explanations are accompanied by an overall comparison of the DIF detection methods without respect to specific crossing of conditions. Furthermore, the findings for research question two (RQ2) are organized by simulation condition (i.e., marginally). Particular interest is given to comparisons of CCRs, as well as type II error rates conditioned on type I error rates, for each method.

Research Question One

To restate RQ1, how does the penalized LR DIF detection method (i.e., LR lasso) compare to more traditional non-IRT multiple-group methods (i.e., generalized Mantel-Haenszel χ^2 and generalized logistic regression) as it relates to:

a. correct classification rate of DIF items?
b. type I error rate in the classification of DIF items?
c. hit rates (defined as one minus the type II error rate) in the classification of DIF items?
d. phi correlations of true and detected DIF items?
e. agreement statistics among methods?

Subsections within the following section are organized to address the various subquestions listed under RQ1.

Correct Classification Rates

Table 4 contains the CCRs for the conditions where the test targets were commensurate with the location of the simulees and there was no simulee impact. To clarify, each cell in the table summarizes the 250 replications for that given crossing of conditions and pertinent analysis model. Overall, the LR lasso method had comparable performance in correctly identifying non-DIF items in the null cases where DIF was not introduced. That is, the rows in Table 4 that have zero DIF items all provided a similar result, and contain values that were very similar across all three methods. However, as the number of DIF items increased, degradation in LR lasso performance was noted. Even within the LR lasso columns, the difference between scenarios with two DIF items and four DIF items was at least a few percentage points.

Table 5 contains the CCRs for the conditions where the test targets were commensurate with the location of the simulees, but there was a half-logit span total in impact across the four simulee groups. Compared with the CCRs for LR lasso observed in the absence of impact, the presence of impact led to lower CCRs, particularly when the TIF shape was more spread. Also, unlike the results in the absence of impact, the CCRs for LR lasso were not comparable to GMH and GLR in the null conditions when impact was present. To explain the table tersely, the CCRs decreased more precipitously for LR lasso than GMH and GLR as the number of DIF items increased.

Table 4. Correct Classification Rates (as Percentages) for Commensurate Test Targets and No Simulee Impact.

			Balanced Sample Size			Unbalanced Sample Size		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	0	94.96	95.24	94.99	94.99	95.17	95.50
		2	95.10	95.12	93.17	94.77	94.94	93.04
		4	93.40	93.53	89.53	93.02	93.32	89.49
	+1	0	94.85	95.68	94.85	95.20	95.44	95.82
		2	94.59	95.18	93.87	94.38	94.66	93.18
		4	93.55	93.53	89.01	93.60	93.70	88.61
1.5	+0	0	94.94	95.19	95.22	94.70	94.88	95.31
		2	94.29	94.24	92.57	95.04	94.74	92.87
		4	92.94	92.98	90.00	92.61	92.86	89.07
	+1	0	95.01	94.95	95.24	95.05	94.96	95.07
		2	94.44	94.36	93.35	94.38	94.10	92.59
		4	92.75	93.14	89.20	93.19	93.12	89.18

Table 5. Correct Classification Rates (as Percentages) for Commensurate Test Targets and a Half Logit Total of Simulee Impact.

			Balanced Sample Size			Unbalanced Sample Size		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	0	95.14	94.98	93.68	95.23	95.33	95.17
		2	94.96	94.43	92.05	94.40	94.14	91.67
		4	93.54	93.77	88.77	93.60	93.79	88.85
	+1	0	94.68	94.85	94.01	94.68	95.16	94.75
		2	95.04	94.84	91.78	94.57	94.39	92.46
		4	93.25	93.48	88.53	92.93	93.29	88.51
1.5	+0	0	95.04	94.54	93.15	94.85	94.69	93.00
		2	94.46	93.51	90.54	94.23	93.85	90.99
		4	92.59	92.09	86.80	92.75	92.81	87.11
	+1	0	94.99	94.26	92.09	94.92	94.28	93.79
		2	93.97	93.67	89.33	94.41	93.63	90.50
		4	92.69	92.85	86.54	92.72	92.52	86.82

CCRs are contained in Table 6 for the conditions where the test targets were disparate with the location of the simulees (specifically, 1.15 logit units below the mean of the simulee samples that were generated) and there was no simulee impact. When compared with the commensurate scenarios in Table 4, the CCRs in Table 6 did not have many appreciable differences across methods and conditions. Though, a few noteworthy cases should be described. When the TIF shape was narrower and the sample size was unbalanced, LR lasso performance suffered in the null condition by roughly 1.5% percentage points (when compared to Table 4). On the other hand, LR lasso performance improved when the TIF shape was narrower and the sample size was unbalanced by roughly half a percentage point for scenarios when there were four DIF items located a logit above the maximum TIF point (also when compared to Table 4).

Table 6. Correct Classification Rates (as Percentages) for Disparate Test Targets and No Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	0	94.97	95.04	95.12	94.42	94.68	94.04
		2	94.62	95.13	92.71	94.97	94.84	93.21
		4	93.42	93.86	89.13	94.00	93.73	89.43
	+1	0	94.93	95.25	95.21	94.88	94.82	94.37
		2	94.71	94.91	92.78	94.83	95.05	92.62
		4	93.71	93.76	89.69	93.01	93.06	88.96
1.5	+0	0	94.92	95.20	94.92	94.81	94.98	95.01
		2	94.29	94.33	92.47	94.56	94.84	93.23
		4	93.32	93.51	88.61	93.60	93.33	89.69
	+1	0	95.09	95.03	95.22	95.03	95.20	95.35
		2	94.48	94.51	92.89	94.91	94.55	92.95
		4	93.50	93.34	89.16	93.53	92.93	89.57

CCRs are contained in Table 7 for the conditions where the test targets were disparate with the location of the simulees (i.e., 1.15 logit units below the mean of the simulee sample), but there was a half-logit span total in impact across the four simulee groups. When compared with the commensurate scenarios in Table 5, the disparate scenarios in Table 7 reflected a degradation in performance for the GLR. LR lasso had a substantial boost in performance across many of the conditions, particularly when the sample size was balanced. In fact, many of the deficits noted in Table 5 were improved in Table 7. GMH had some improvements between the commensurate and disparate scenarios, though not as drastic as LR lasso. Generally, GMH tended to have the highest CCRs across Tables 4 through 7, so the gains in performance would likely have been lesser because it appeared to be more robust to the variety in conditions.

Table 7. Correct Classification Rates (as Percentages) for Disparate Test Targets and a Half Logit Total of Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	0	95.14	93.73	94.35	95.02	93.49	94.42
		2	94.55	93.65	92.29	95.04	94.26	91.86
		4	93.55	92.73	89.55	93.65	92.95	88.46
	+1	0	95.29	93.55	94.67	95.23	94.15	94.49
		2	94.83	93.40	92.73	94.71	93.98	91.79
		4	93.33	92.76	89.38	93.69	93.24	88.37
1.5	+0	0	94.83	93.57	94.45	95.24	94.17	94.82
		2	94.68	93.21	91.14	95.02	93.77	92.34
		4	92.88	91.87	88.37	93.19	92.09	87.99
	+1	0	95.24	93.41	95.26	94.72	94.01	94.94
		2	94.45	93.32	91.62	94.47	93.76	91.81
		4	93.45	92.28	88.08	93.54	92.53	88.02

Type I Error Rates

Table 8 is comprised of the type I error rates for the conditions where the test targets were commensurate with the location of the simulees and there was no simulee impact. As such, it is a counterpart to Table 4, and provides additional data that help with interpretation of fluctuations in CCRs. As it stands, the type I error rates in Table 8 support the notion that LR lasso had an increased false positive rate more so with a narrow TIF when compared with a more spread TIF. On the contrary, an interesting and seemingly counterintuitive finding emerged between the balanced and unbalanced sample size conditions for LR lasso.

Table 8. Type I Error Rates (as Percentages) for Commensurate Test Targets and No Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	0	5.04	4.76	5.01	5.01	4.83	4.50
		2	4.68	4.46	5.18	4.93	4.62	4.59
		4	5.68	5.07	6.21	6.17	5.33	5.57
	+1	0	5.15	4.32	5.15	4.80	4.56	4.18
		2	5.18	4.42	4.47	5.33	4.94	4.22
		4	5.64	5.10	7.02	5.54	5.01	6.65
1.5	+0	0	5.06	4.81	4.78	5.30	5.12	4.69
		2	5.18	5.01	5.74	4.44	4.61	4.41
		4	5.52	5.02	6.10	5.91	5.42	5.88
	+1	0	4.99	5.05	4.76	4.95	5.04	4.93
		2	4.95	4.90	4.87	5.04	5.14	4.50
		4	5.59	4.86	6.68	5.40	5.15	5.30

Upon further investigation and comparison with Table 12 (*vide infra*), it appeared that LR lasso was flagging fewer items more generally (for both true and false positives) when the sample size was unbalanced. And in the case of Table 8, without the consideration of additional data in other tables, could lead to an erroneous conclusion that false positives were being reduced as an artifact of the method having improved performance in correctly discounting non-DIF items.

Table 9 contains the type I error rates for the conditions where the test targets were commensurate with the location of the simulees, but there was a half logit total of simulee impact. It is a counterpart to Table 5 and provides additional data that help with interpretation of fluctuations in CCRs. When the TIF shape was spread and multiple DIF items existed, there was a notable difference in the performance of LR lasso between the balanced and unbalanced sample size scenarios.

Table 9. Type I Error Rates (as Percentages) for Commensurate Test Targets and a Half Logit Total of Simulee Impact.

TIF Shape	**DIF Loc.**	**DIF Items**	**Balanced Sample Size**			**Unbalanced Sample Size**		
			GMH	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	0	4.86	5.02	6.32	4.77	4.67	4.83
		2	4.69	5.11	6.04	5.15	5.26	5.43
		4	5.29	4.75	6.87	5.54	5.02	5.68
	+1	0	5.32	5.15	5.99	5.32	4.84	5.25
		2	4.58	4.61	6.20	5.01	5.10	4.71
		4	5.67	5.09	6.87	5.91	5.20	5.64
1.5	+0	0	4.96	5.46	6.85	5.15	5.31	7.00
		2	4.74	5.52	7.32	5.01	5.29	6.19
		4	5.24	5.71	8.84	5.34	5.25	7.30
	+1	0	5.01	5.74	7.91	5.08	5.72	6.21
		2	5.09	5.28	8.35	4.85	5.48	6.54
		4	5.41	5.18	8.71	5.48	5.63	7.51

At first, it appeared that LR lasso performance improved with an unbalanced sample size, but it was most likely due to fewer items being flagged altogether (in tandem with Table 13; *vide infra*). This finding was not surprising given the prevalence of impact and is interesting even when test information was spread (which would further have increased the variability in item scores for three groups affected by impact).

Table 10 contains the type I error rates for the conditions where the test targets were disparate with the location of the simulees and there was no simulee impact. Similar to Table 9, if Table 10 is interpreted out-of-context, there was a false sense that LR lasso performance improved with unbalanced samples. Interestingly, LR lasso had a lower type I error rate than GMH and GLR in cases when the TIF shape was spread and there were multiple DIF items with unbalanced sample sizes. However, given the lower hit rates that were observed as well (see Table 14; *vide infra*), it seemed most likely that the disparate test target (without the presence of impact) reduced variability in item scores in such a way that the lasso regularization was not as effective.

Table 11 is comprised of the type I error rates for conditions where the test targets were disparate with the location of the simulees and there was a half logit total of impact. It is a counterpart to Table 7. In general, there was a degradation in the performance of GLR and LR lasso when the information target was offset and simulee groups differed considerably. Like other tables with type I error rates, LR lasso falsely appeared to improve for unbalanced sample sizes (when compared with balanced sample sizes) with a spread TIF shape and multiple DIF items. Interestingly, with GLR in the null cases. Non-DIF variability in the data increased the type I error rates for GLR more than LR lasso.

Table 10. Type I Error Rates (as Percentages) for Disparate Test Targets and No Simulee Impact.

			Balanced Sample Size			Unbalanced Sample Size		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	0	5.03	4.96	4.88	5.58	5.32	5.96
		2	5.21	4.58	5.68	4.86	4.78	4.78
		4	5.87	5.03	7.02	5.46	5.28	5.99
	+1	0	5.07	4.75	4.79	5.12	5.18	5.63
		2	5.06	4.72	5.70	5.06	4.68	5.05
		4	5.75	5.24	6.65	6.46	5.96	6.76
1.5	+0	0	5.08	4.80	5.08	5.19	5.02	4.99
		2	5.30	5.04	5.60	5.17	4.67	4.14
		4	5.84	5.10	7.42	5.37	5.15	5.39
	+1	0	4.91	4.97	4.78	4.97	4.80	4.65
		2	5.20	4.93	5.19	4.81	4.93	4.34
		4	5.47	5.08	6.76	5.52	5.71	5.15

Table 11. Type I Error Rates (as Percentages) for Disparate Test Targets and a Half Logit Total of Simulee Impact.

			Balanced Sample Size			Unbalanced Sample Size		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	0	4.86	6.27	5.65	4.98	6.51	5.58
		2	5.20	6.08	6.19	4.74	5.46	5.90
		4	5.89	6.33	7.19	5.75	6.19	7.36
	+1	0	4.71	6.45	5.33	4.77	5.85	5.51
		2	5.00	6.35	5.56	5.09	5.73	5.98
		4	6.03	6.46	6.79	5.77	6.06	6.82
1.5	+0	0	5.17	6.43	5.55	4.76	5.83	5.18
		2	5.00	6.34	7.12	4.70	5.84	5.22
		4	5.97	6.74	7.96	5.86	6.62	6.84
	+1	0	4.76	6.59	4.74	5.28	5.99	5.06
		2	5.22	6.20	6.48	5.10	5.72	5.40
		4	5.48	6.39	7.88	5.54	6.35	6.91

Hit Rates

The tables in the following subsection (Tables 12 through 15) present the hit rates for all crossings of conditions. The tables omit rows for the null conditions because type II error was not possible in the absence of DIF items. To guide interpretation, the hit rates were calculated as one minus the ratio of missed DIF items over the number of generated DIF items.

Table 12 contains the hit rates for conditions when test targets were commensurate with the simulee population and there was no simulee impact. All three methods were adversely affected by having an increasing number of DIF items, particularly LR lasso. Additionally, an unbalanced sample size had an adverse effect on LR lasso that was not observed with GMH and GLR.

Table 12. Hit Rates (as Percentages) for Commensurate Test Targets and No Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	2	95.60	91.60	67.00	94.00	91.20	52.60
		4	90.80	86.00	57.40	91.90	86.50	50.60
	+1	2	95.40	92.00	66.80	94.20	92.00	48.00
		4	91.90	86.30	60.30	91.40	87.10	52.60
1.5	+0	2	89.40	85.00	66.20	89.60	87.00	45.60
		4	84.60	80.00	61.00	85.20	82.80	49.50
	+1	2	87.80	85.20	64.40	88.40	84.80	41.80
		4	83.40	80.00	58.80	85.90	82.70	44.80

Table 13, when compared with Table 12, appeared to have rather small differences with the introduction of impact. The largest differences were observed for LR lasso when there was a narrow TIF shape. When the TIF shape was spread, the values were relatively close to the values found in Table 12. This finding has an implication for practitioners in that a TIF shape that is very targeted may cause as LR lasso to miss roughly half of the DIF items that may exist on a test form.

Table 13. Hit Rates (as Percentages) for Commensurate Test Targets and a Half Logit Total of Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	2	93.00	90.80	61.80	91.00	88.00	42.00
		4	88.30	85.20	56.40	91.40	88.10	45.30
	+1	2	92.40	89.00	59.60	91.60	89.80	43.40
		4	89.20	85.70	54.00	88.40	84.90	41.50
1.5	+0	2	84.00	80.60	57.20	84.80	82.80	43.60
		4	78.30	78.00	56.40	80.90	80.60	44.10
	+1	2	81.20	79.00	53.60	85.20	82.20	40.80
		4	81.00	80.30	52.50	82.00	81.50	43.30

Table 14 contains hit rates for when test targets are disparate with respect to the simulee population location and there was no simulee impact. In conjunction with Table 10, there appeared to be both increased type I and type II errors for LR lasso. When compared with Table 12, the values for GMH and GLR were ever so slightly better. While a disparate test target would seem like a methodological hurdle because of restricted variance, it actually proves to be somewhat advantageous because all non-DIF items begin to look more similar under a disparate target (i.e., the majority of simulees are responding correctly), which causes any DIF items to be more distinctive.

Table 14. Hit Rates (as Percentages) for Disparate Test Targets and No Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	2	96.60	94.20	67.80	96.60	92.40	59.80
		4	92.90	88.90	61.50	94.60	90.10	54.20
	+1	2	95.40	92.60	69.60	97.80	94.60	53.40
		4	94.60	90.00	63.40	94.70	90.20	57.20
1.5	+0	2	91.80	87.40	61.40	94.60	90.20	47.40
		4	91.60	86.10	60.30	89.70	84.80	50.80
	+1	2	93.60	88.80	61.60	94.40	89.60	45.80
		4	89.70	84.20	59.20	90.50	86.40	47.20

Table 15 contains the hit rates for disparate test targets when there was simulee impact present. The TIF target was more aligned with impacted groups in these cases, so the hit rates would generally be expected to be better than those found in Table 13. However, the hit rates were fairly similar to Table 14.

Table 15. Hit Rates (as Percentages) for Disparate Test Targets and a Half Logit Total of Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	2	95.00	94.60	69.60	95.60	94.40	55.20
		4	94.40	90.60	67.40	94.00	91.40	58.20
	+1	2	96.60	95.00	65.80	96.00	94.20	55.40
		4	93.60	92.20	61.70	94.60	93.00	51.90
1.5	+0	2	93.60	91.00	65.20	94.40	92.20	51.20
		4	88.50	86.10	63.30	90.50	87.10	48.30
	+1	2	93.40	90.40	62.00	91.40	89.60	44.20
		4	89.30	86.70	59.60	90.80	88.80	49.30

Correlations Between Truth and Predicted

Phi coefficients were calculated to determine the strength of the association between the generated and detected DIF items. This statistic captures succinctly what the CCR on its own does not, which was the conglomeration of CCR, type I error, and type II error into a single value. As such, there was not a one-to-one relationship between the φ coefficient and the other values presented herein because the φ coefficient contributes novel information. Guidelines for interpreting Pearson-product moment correlation coefficients generally can be employed for ease of interpretation. However, given the slight data-distribution dependency of the statistic, smaller differences between φ coefficients should not be interpreted as particularly meaningful.

Tables 16 through 19 contain the association values across all crossings of conditions within the study. Rows for null conditions are not included because the φ coefficient cannot be calculated when one variable is a constant (i.e., no DIF items). Principally, LR lasso had weaker relationships between generated and detected DIF items when compared with GMH and GLR. Similar performance was noted between GMH and GLR in general. Paradoxically, Table 19 contains larger values than Table 17, which can be attributed to the previously described scenarios when the lasso regularization has a reduced type I error rate, given that the presence of impact caused the TIF target to be less disparate for the groups experiencing impact.

Table 16. Phi Correlations of Predicted and Truth for Commensurate Test Targets and No Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	2	0.709	0.691	0.551	0.690	0.682	0.489
		4	0.728	0.717	0.499	0.720	0.710	0.493
	+1	2	0.690	0.696	0.566	0.680	0.679	0.465
		4	0.736	0.717	0.493	0.735	0.724	0.467
1.5	+0	2	0.650	0.638	0.513	0.677	0.649	0.420
		4	0.693	0.674	0.519	0.681	0.683	0.445
	+1	2	0.652	0.633	0.534	0.646	0.627	0.368
		4	0.683	0.682	0.498	0.704	0.689	0.444

Table 17. Phi Correlations of Predicted and Truth for Commensurate Test Targets and a Half Logit Total of Simulee Impact.

			Balanced Sample Size			**Unbalanced Sample Size**		
TIF Shape	**DIF Loc.**	**DIF Items**	**GMH**	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	2	0.700	0.660	0.492	0.665	0.636	0.365
		4	0.721	0.718	0.470	0.736	0.732	0.440
	+1	2	0.693	0.672	0.450	0.672	0.660	0.426
		4	0.717	0.713	0.456	0.707	0.701	0.403
1.5	+0	2	0.635	0.587	0.403	0.629	0.603	0.356
		4	0.656	0.644	0.436	0.671	0.671	0.384
	+1	2	0.602	0.572	0.350	0.634	0.588	0.313
		4	0.672	0.675	0.393	0.673	0.665	0.347

Table 18. Phi Correlations of Predicted and Truth for Disparate Test Targets and No Simulee Impact.

TIF Shape	DIF Loc.	DIF Items	Balanced Sample Size			Unbalanced Sample Size		
			GMH	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	2	0.693	0.705	0.516	0.712	0.690	0.503
		4	0.735	0.737	0.508	0.760	0.737	0.467
	+1	2	0.690	0.688	0.532	0.709	0.703	0.437
		4	0.753	0.736	0.526	0.733	0.718	0.472
1.5	+0	2	0.661	0.644	0.489	0.688	0.680	0.447
		4	0.730	0.716	0.478	0.730	0.702	0.484
	+1	2	0.671	0.656	0.505	0.697	0.665	0.434
		4	0.725	0.700	0.503	0.733	0.698	0.451

Table 19. Phi Correlations of Predicted and Truth for Disparate Test Targets and a Half Logit Total of Simulee Impact.

TIF Shape	DIF Loc.	DIF Items	Balanced Sample Size			Unbalanced Sample Size		
			GMH	**GLR**	**LR lasso**	**GMH**	**GLR**	**LR lasso**
0.5	+0	2	0.683	0.653	0.531	0.705	0.671	0.447
		4	0.744	0.712	0.541	0.746	0.720	0.485
	+1	2	0.701	0.639	0.519	0.697	0.663	0.456
		4	0.734	0.716	0.505	0.751	0.732	0.443
1.5	+0	2	0.687	0.617	0.475	0.698	0.646	0.425
		4	0.706	0.669	0.498	0.720	0.676	0.409
	+1	2	0.673	0.621	0.467	0.667	0.629	0.379
		4	0.725	0.679	0.468	0.729	0.697	0.413

Agreement among Methods

The values in Table 20 should be interpreted truly as a reflection of consistency, and not as a reflection of accuracy. The purpose of this table is to demonstrate how frequently the various methods gave identical results. Generally, the use of LR lasso had a tendency to flag a different set of items than GMH and GLR in roughly four out of every five applications of the methods. While this finding was not substantially different than the consistency between GMH and GLR, it was notably lower. The values are relatively small in the column for percent exact agreement because the exact agreement means that the classification of 40 items within a given crossing of conditions and replication was completely identical across the two methods being compared (i.e., the 40 items were flagged identically). CCRs are parsed by condition in the next section.

Table 20. Agreement Measures for Correct Classification Rates among Methods across All Conditions and Replications.

Methods being Compared	**Unweighted Kappa**	**Weighted Kappa**	**Percent Exact Agreement**	**Percent Adjacent Agreement**	**Percent Combined Agreement**
GMH & GLR	.104	.470	27.61	42.06	69.68
GMH & LR Lasso	.041	.197	18.87	31.95	50.82
GLR & LR Lasso	.025	.159	17.38	30.61	48.00

Research Question Two

To restate RQ2, when detecting items that truly exhibit DIF, to what degree is classification error for each analysis model influenced by changes in:

a. the location of the information target relative to the examinee population?

b. the shape of the information function?

c. the location of DIF items relative to the information target?

d. the percentage of DIF items?

e. the amount of impact?

f. sample size?

Subsections within the following section are organized to address the various subquestions listed under RQ2. This presentation of the results allows for conditions to be collapsed marginally by simulation condition, with 8,000 or 12,000 replications contained for every box or line in each figure (figures for percentage of DIF items contained 8,000 replications, and all other conditions contained 12,000 replications). While collapsing the data in such a manner can obfuscate the complexities and nuances of the individual scenarios (which were already described in preceding tables), it also permits a wider range of variability in results for each condition and allows for results to be generalized to a larger subset of scenarios. It is important to note that the presentation of CCRs in this section is different than those in Tables 4 through 7, given that the boxplots below show distributional characteristics of CCRs and present findings at a higher grain-size. As a general description, the findings in this section consistently signal a less accurate performance of the LR lasso method when compared with GMH and GLR.

Location of Information Target Relative to Simulee Population

The boxplots in Figure 2 are a comparison of the three methods when considering the location of the TIFs relative to the simulee population. The degradation of LR lasso performance was readily observed. In fact, the median CCR of LR lasso corresponded with the first quartile (Q1) of GMH and GLR for both commensurate and disparate TIF locations. A difference that was truly negligible can be seen in the plot between GMH and GLR on the lower tails (that is, only a single replication for the commensurate target conditions and just three replications for the disparate target conditions appeared at a

CCR of 30 items for GLR). As noted previously, the improved CCR for LR lasso between the commensurate and disparate target conditions was an artifact of fewer items being flagged in total and was primarily driven by a decrease in type I error rates. In other words, the LR lasso method did not truly perform better when the TIF was offset from the simulees. Instead, it made fewer mistakes because of reduced variability in the data (particularly in the absence of impact).

Figure 3 portrays a similar outcome, with the type II error rate having very small gains in performance with a disparate target only when the type I error rate was zero. To further clarify interpretation of the figure, the relative spikes in performance should not be overinterpreted, as there were fewer replications that corresponded with the increasing type I error rates. It should be noted that most of the conditional plots throughout this section contain much longer lines for LR lasso, which elucidates that the longer tails in the CCR boxplots are primarily due to increased type I error rates.

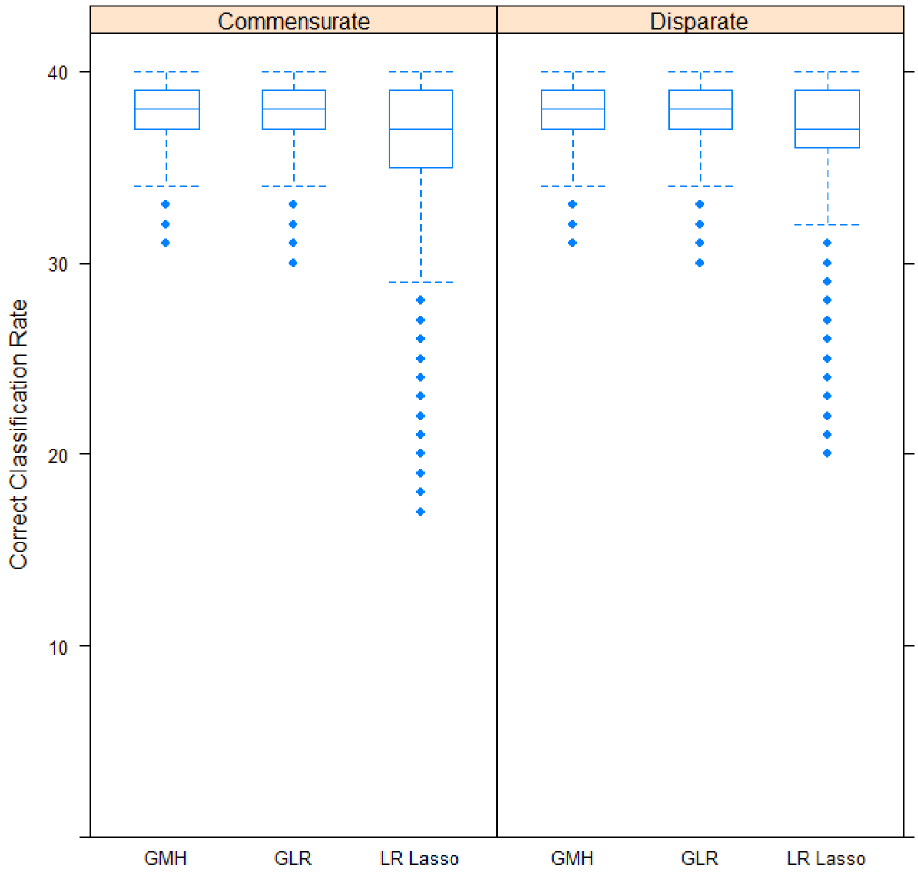


Figure 2. Correct Classification Rates (in Number of Items) across All Replications For Conditions Parsed by Commensurate and Disparate Locations of Test Information Targets Relative to Simulee Populations.

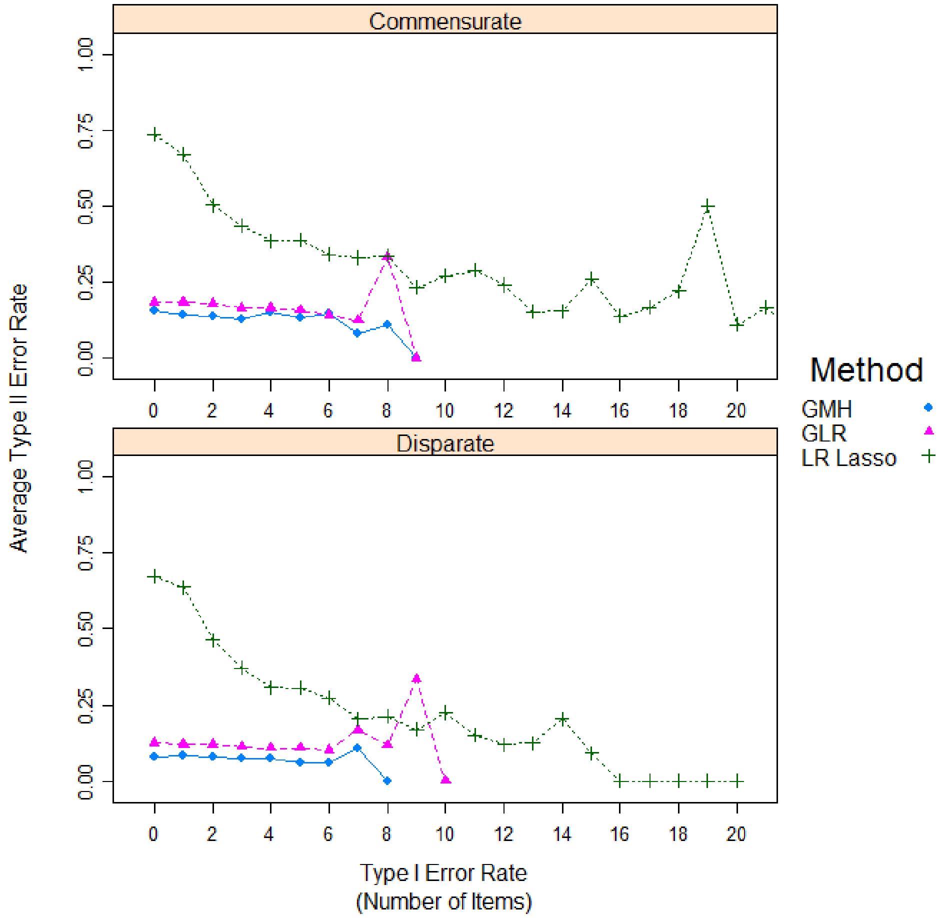


Figure 3. Average Type II Error Rate (by Number of Items) for Various Levels of Type I Error Rates Parsed by Commensurate and Disparate Locations of Test Information Targets Relative to Simulee Populations.

Shape of Information Function

The effects of the relative spread of the test information function can be found in Figures 4 and 5. As seen specifically in Figure 4, the performance of LR lasso was hampered as the TIF was spread further (via the variance of the item difficulties), though the median CCR was the same between the narrow and spread TIF conditions. GMH and GLR seemed to have similar performance in both cases and were not particularly sensitive to changes in the spread of the items. Figure 5 confirms many of these same findings and shows that the type II error rates seemed to be comparable to Figure 3 when conditioned on type I error rates across cases for all methods.

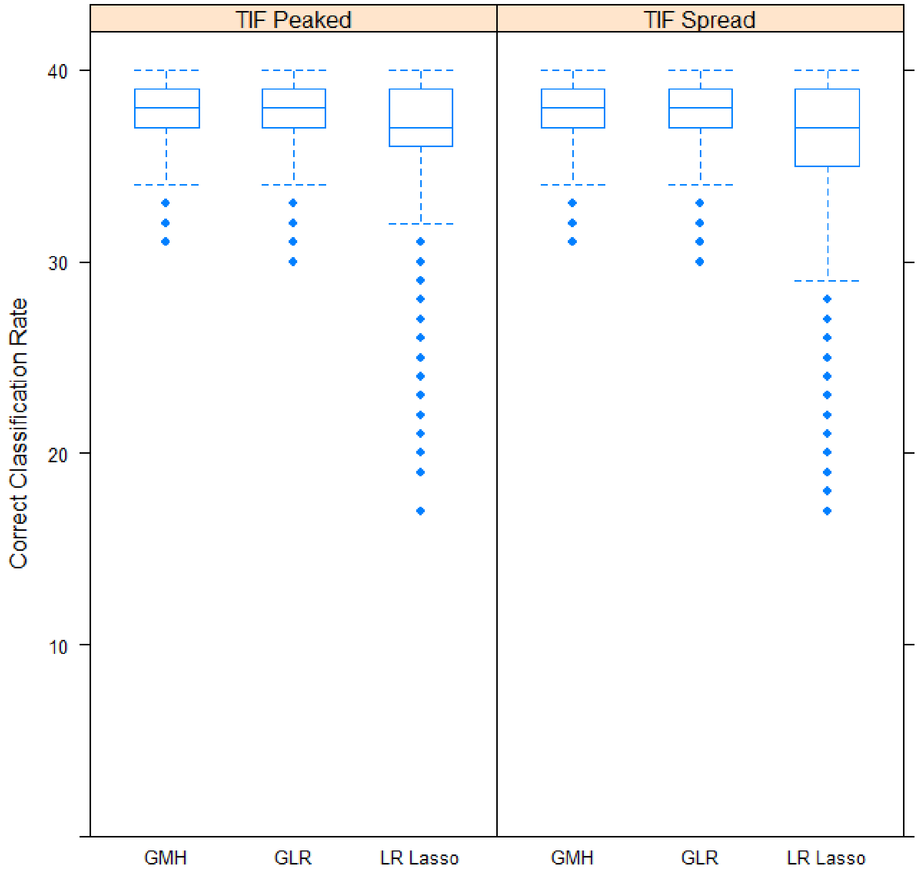


Figure 4. Correct Classification Rates (in Number of Items) across All Replications for Conditions Parsed by the Relative Spread of the Test Information Function.

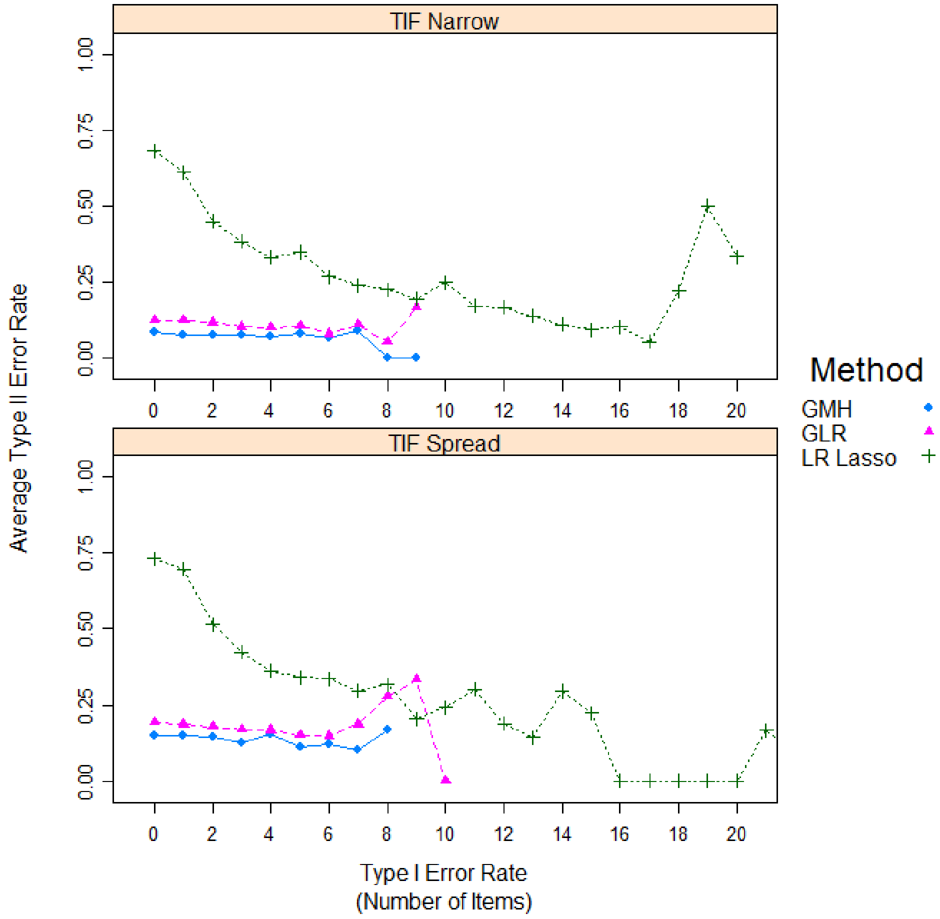


Figure 5. Average Type II Error Rate (by Number of Items) for Various Levels of Type I Error Rates Parsed by the Relative Spread of the Test Information Function.

Location of DIF Items Relative to Information Target

Figures 6 and 7 display the results for when the location of the DIF items was nearest the information target versus offset above the target. The findings closely mirror those found in Figures 4 and 5, with the tails of the LR lasso plots seemingly having minor differences when compared with those before. Stated differently, it appeared that the location of the DIF items with respect to the test information target had a similar effect on the methods as changing the spread of item difficulties. This finding perhaps serves as a caution for test developers considering LR lasso when there are multiple cut points on a test scale, for example.

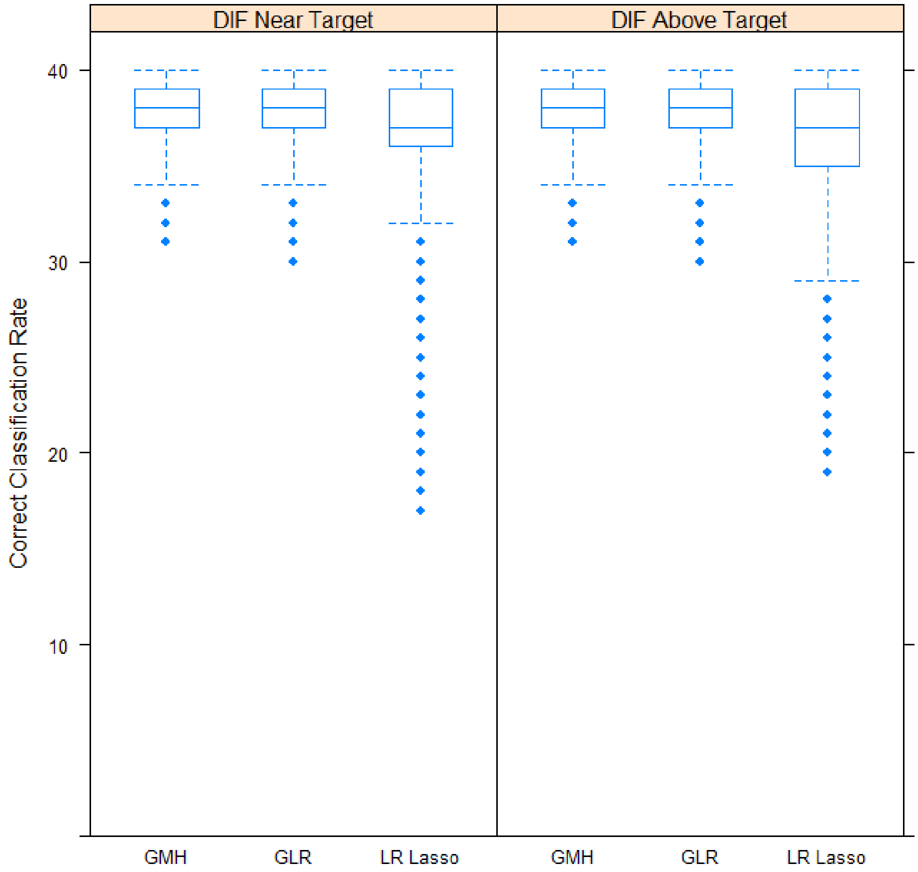


Figure 6. Correct Classification Rates (in Number of Items) across All Replications for Conditions Parsed by Location of DIF Items Relative to the Test Information Target.

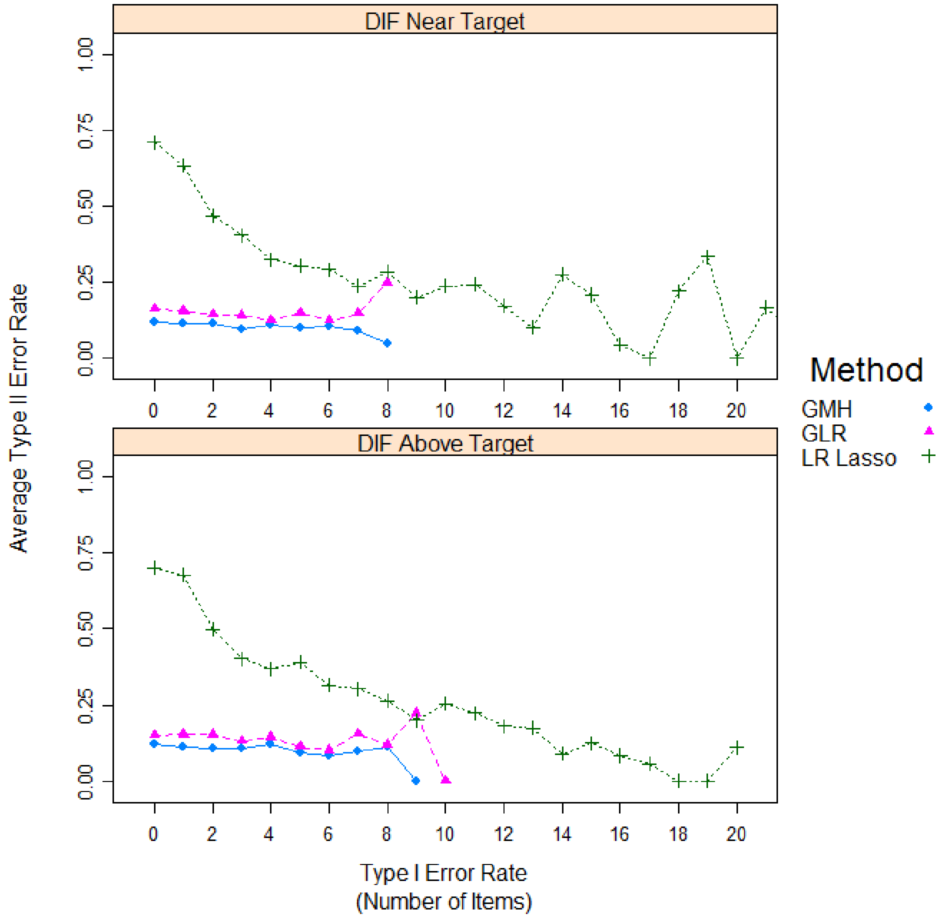


Figure 7. Average Type II Error Rate (by Number of Items) for Various Levels of Type I Error Rates Parsed by Location of DIF Items Relative to the Test Information Target.

Percentage of DIF Items

The effects of the presence of DIF items can be found in Figures 8 and 9. In Figure 8, LR lasso performance decreased as the number of DIF items increased. GMH and GLR only suffered once the test was comprised of 10% DIF items. Figure 9 portrays a lower than usual type II error rate for LR lasso when a test was composed of 5% DIF items. For all three methods the total score was used either as a conditioning variable or predictor variable. As the total score became increasingly contaminated by the presence of DIF items, the less trustworthy it was for conditioning or predicting. Purification procedures are used normally to mitigate issues related to contamination of the total score. However, as mentioned previously, purification was not used in this study because it would have served as a confound in the comparison with the lasso regularization. Even without purification, GMH and GLR outperform LR lasso. One difference between GLR and LR lasso is that the coefficient for test score, S_i, is constrained to be the same for all items with LR lasso. This constraint unintentionally ensures that any DIF items are always included in the prediction.

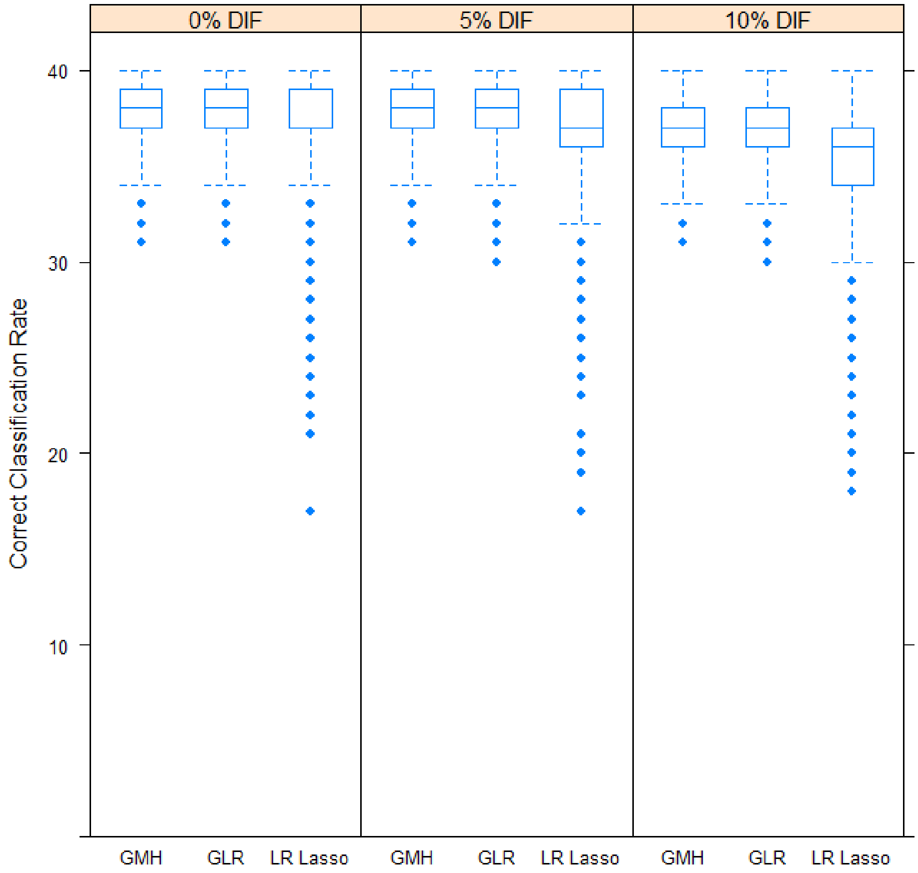


Figure 8. Correct Classification Rates (in Number of Items) across All Replications for Conditions Parsed by Percentage of DIF Items.

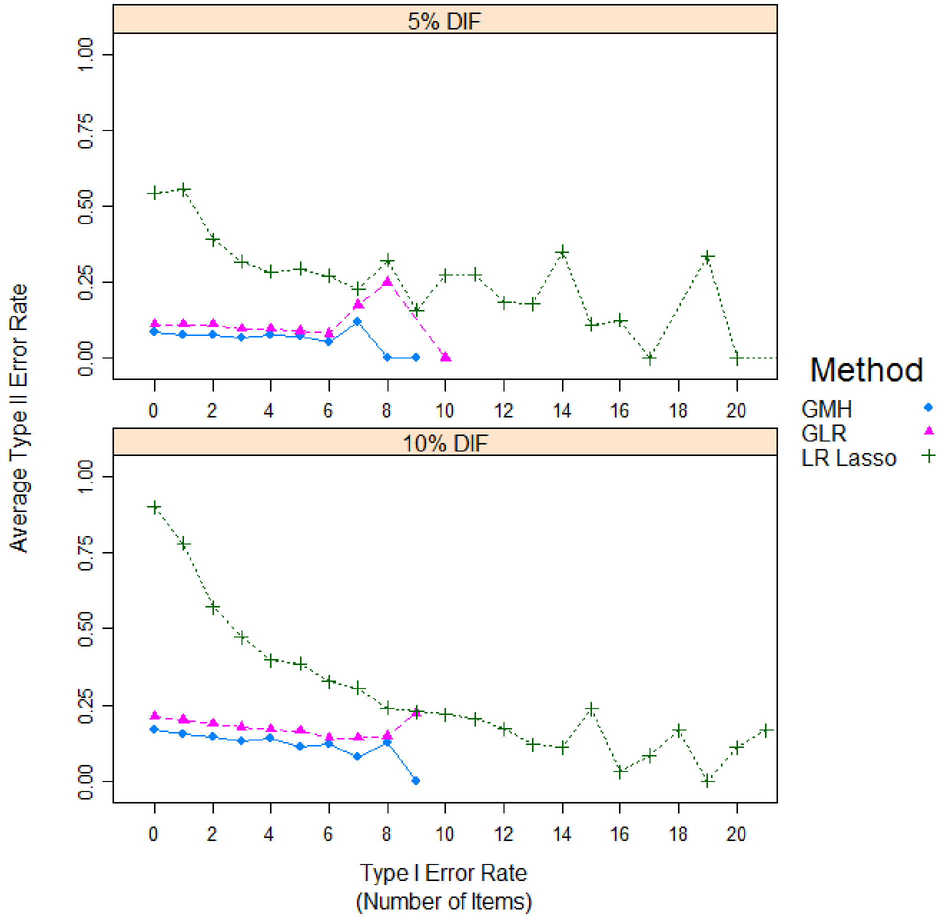


Figure 9. Average Type II Error Rate (by Number of Items) for Various Levels of Type I Error Rates Parsed by Percentage of DIF Items.

Amount of Impact

Figures 10 and 11 contrast the methods in the absence of impact versus a half logit total span of impact on simulees. Relatively speaking, GMH and GLR suffered only minor differences, with a slight increase in type I error rates (see Tables 8 through 11) that was not perceptible in Figure 10. LR lasso suffered noticeably more. When conditioned on type I error rates in Figure 11, the effect on type II error rates with LR lasso was not readily observed, however.

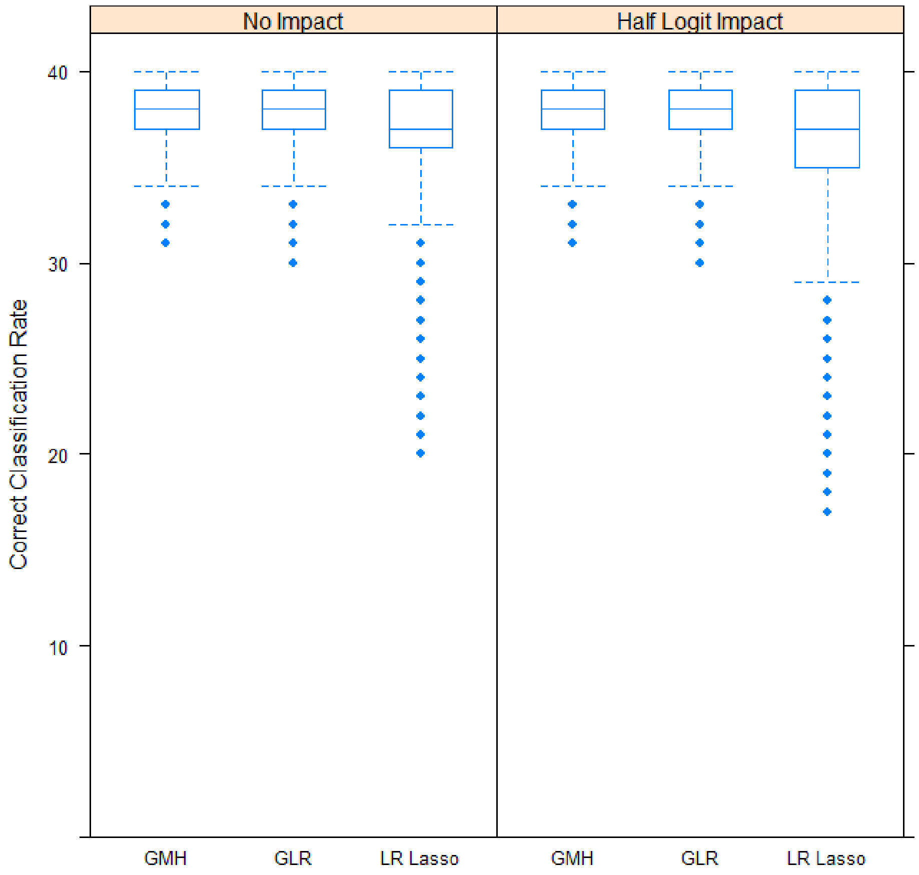


Figure 10. Correct Classification Rates (in Number of Items) across All Replications for Conditions Parsed by Presence of Impact.

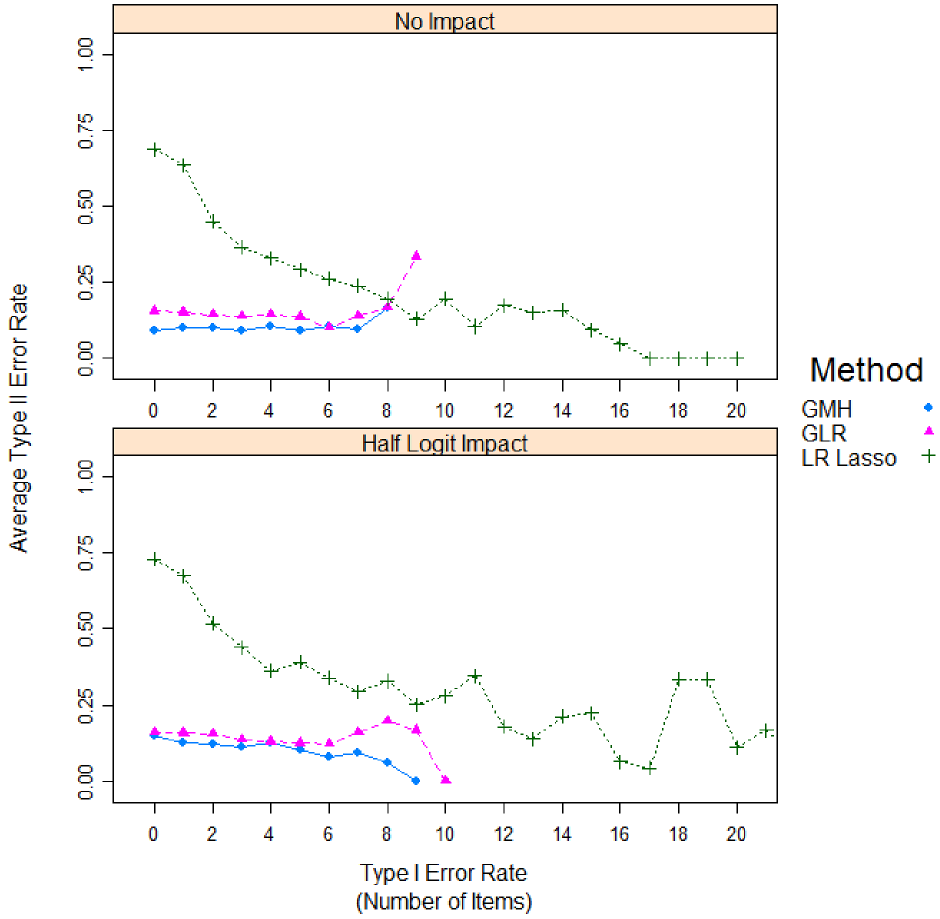


Figure 11. Average Type II Error Rate (by Number of Items) for Various Levels of Type I Error Rates Parsed by Presence of Impact.

Sample Size

Figures 12 and 13 compare conditions with balanced and unbalanced sample sizes. At first glance, Figure 12 seemingly portrays that LR lasso performed better with unbalanced samples. As described previously, it simply made fewer type I error mistakes as a result of reduced variability in the data. Figure 13 is helpful in better understanding that type II error rates, after accounting for type I error rates, in fact were lower for the balanced sample sizes.

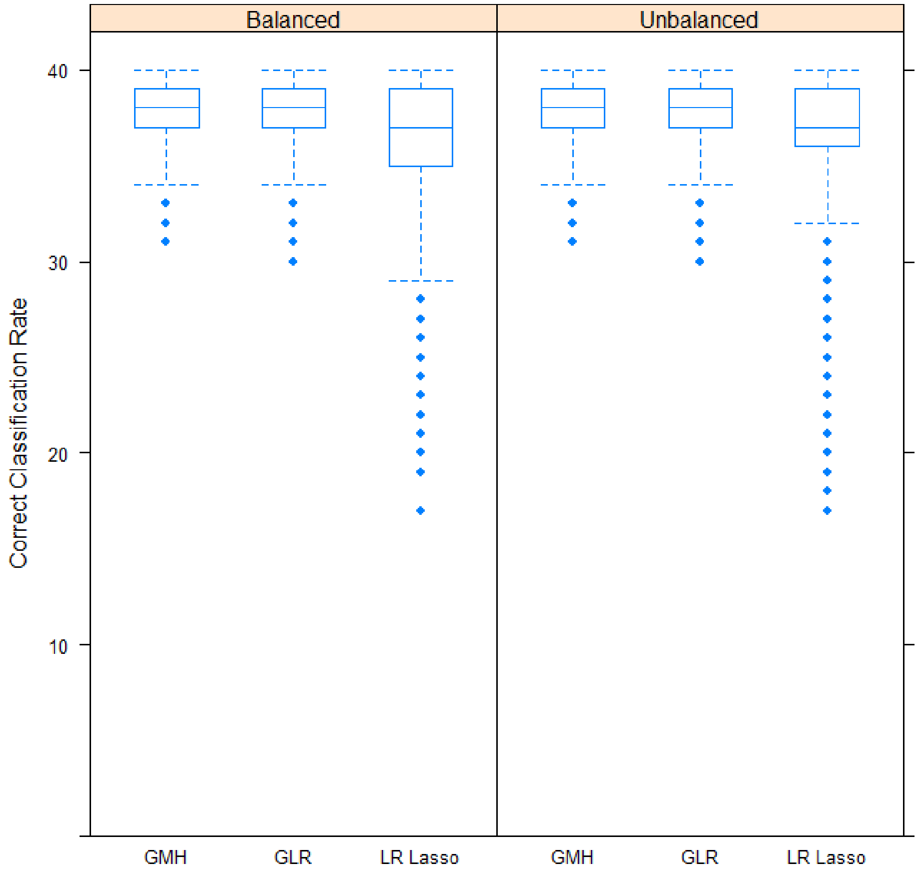


Figure 12. Correct Classification Rates (in Number of Items) across All Replications for Conditions Parsed by Sample Size Balance.

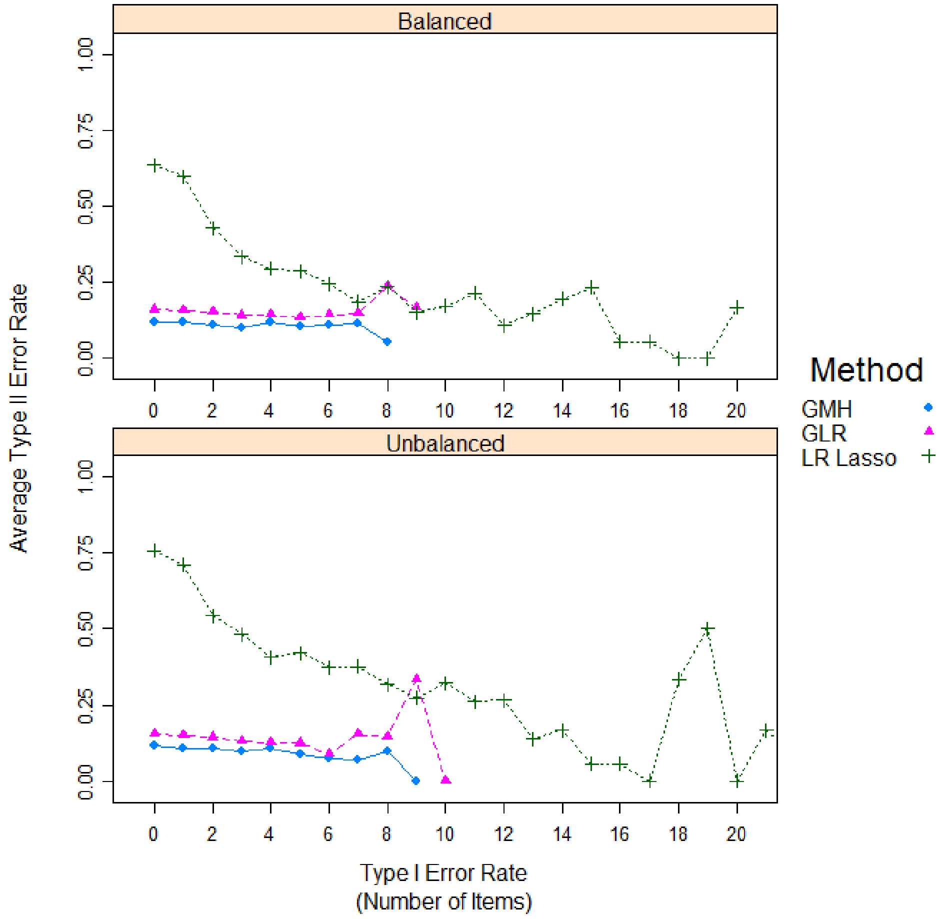


Figure 13. Average Type II Error Rate (by Number of Items) for Various Levels of Type I Error Rates Parsed by Sample Size Balance.

CHAPTER V

DISCUSSION

This chapter provides commentary on trends for variables that were manipulated through the simulation study related to test-level characteristics, item-level characteristics, and simulee characteristics. Additionally, five practical scenarios are described using specific crossings of conditions found within the simulation study. Recommendations are made based upon all the preceding information, as well as extensions are suggested regarding possible effect size measure conceptualizations, limitations, and future research.

Test-Level Characteristics

Figure 14 displays how changing the location of the TIF with respect to the simulee population directly influences the power to detect DIF. As seen in the figure, the ability to detect DIF is best when the location of all the simulee groups is commensurate with the maximum point of the TIF. Any deviations from that ideal scenario, such as a disparate TIF location, begins to limit the power in being able to detect differential item performance.

Given the conditions of this study, however, there are exceptions. For example, the presence of impact can cause some simulee groups to drift closer to a disparate TIF target. Another example is when the TIF is spread. A comparison of the CCRs in Tables

4 and 6 showed that a spread TIF leads to improved performance with a disparate TIF location for both GMH and LR lasso. Particularly for LR lasso, there appeared to be a notable gain in performance due to the reduction in type I error rates observed between Tables 9 and 11. Contrary to expectation, this finding did not hold true for GLR in the presence of impact as was observed in Tables 5 and 7. This provides evidence that GLR seemed to be more strongly influenced by the presence of impact, which was not specifically studied by Magis, Raîche, Béland & Gérard (2011) in their initial formulation of the method.

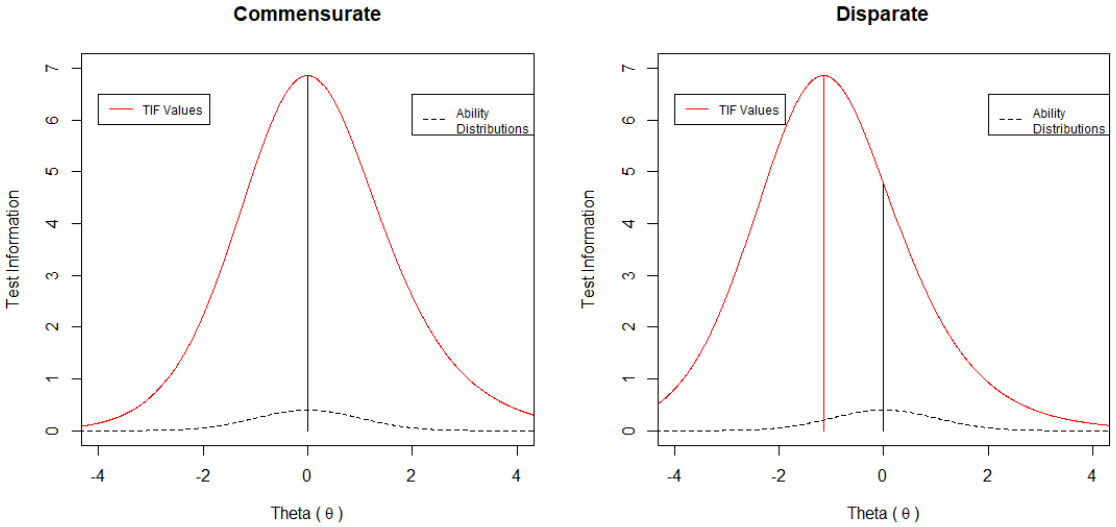


Figure 14. Correspondence of the Generating Ability Distributions and the TIFs by Test Type.

Item-Level Characteristics

The location of the DIF items (i.e., near the information target or offset above the information target) had a different influence dependent upon the location of the TIF target (i.e., commensurate or disparate). Figure 15 shows the correspondence between the generating ability distributions, a prototypical narrow TIF shape, as well as the locations of the DIF items. When the maximum TIF location is commensurate with the location of the simulee population distribution, the offset location of DIF items occurs where there is limited person information to accurately discriminate between the performance of the four groups. When the TIF location is disparate with respect to the simulee central tendency, having DIF items near the TIF target leads to a similar scenario where there is limited person information to adequately discern relative performance between groups. In fact, assuming balanced sample sizes and four DIF items, this similarity was confirmed when comparing hit rate percentages in Table 12 (GMH = 91.90 %; GLR = 86.30 %; LR lasso = 60.30 %) and Table 14 (GMH = 94.60 %; GLR = 90.00 %; LR lasso = 63.40 %). Furthermore, the similarity was seen in the type I error rate percentages in Table 8 (GMH = 5.64 %; GLR = 5.10 %; LR lasso = 7.02 %) and Table 10 (GMH = 5.87 %; GLR = 5.03 %; LR lasso = 7.02 %).

The percentage of DIF items had a profound impact on LR lasso performance. GMH and GLR only seemed to suffer in performance once the test contained at least 10% DIF items. However, when a test contains 10% DIF items in practice, the effectiveness of DIF detection methods becomes less trustworthy in general. In that sense, GMH and GLR seemed to be equally efficient under typically observed scenarios in practice.

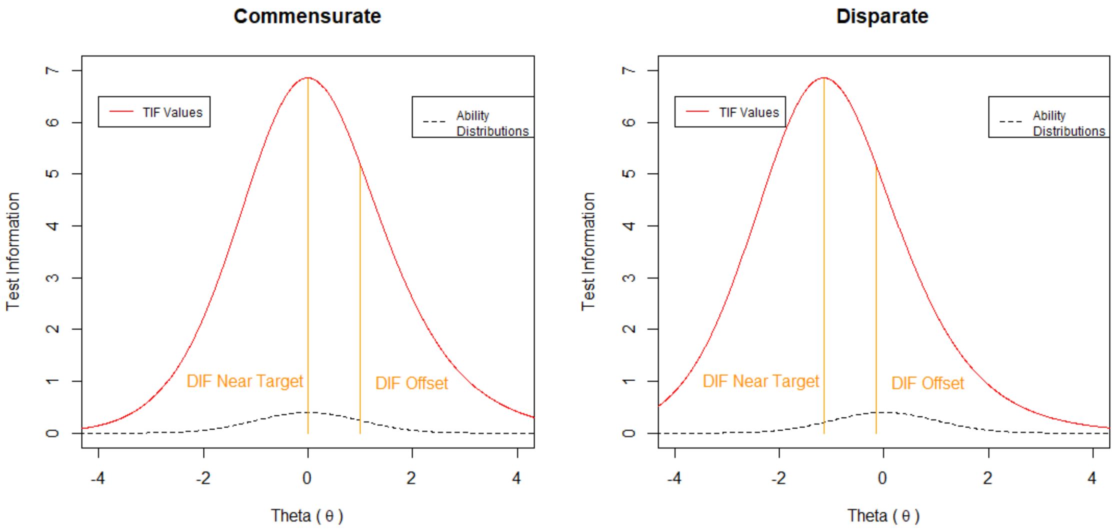


Figure 15. Correspondence of the Generating Ability Distributions and the TIFs by Location of DIF Items.

Regarding LR lasso, the global estimation of the model provided for more optimal model-data fit. However, the penalty parameters (λ) were not as effective at modeling the noise introduced by the DIF items. That is, the penalty parameters should have been better at capturing meaningful group differences. Given the cleaner nature of simulated data analyzed in this study, it appeared that the parameters had the potential of modeling either pseudo-guessing behavior, or even slight misfit introduced by the non-unity discrimination parameters in the generation model, that was not fully captured by the other model parameters given the structural similarities with the Rasch model.

Simulee Characteristics

As mentioned previously, there is a relationship between the presence of impact and the location of the TIF target. This finding can be further explained by Figure 16. In short, the presence of impact causes the focal groups to be closer to the disparate TIF target. However, it is important to note that if the TIF target were disparate in such a way that the assessment would be targeting an ability level much higher than that of even the reference group, then the results would have been completely different (such an example could be a pre-test for a course). In such a case, the presence of impact would not have been somewhat advantageous to the methods because the presence of impact would have caused most of the impacted groups to have missed most of the items. The specifications of this simulation study led to situations where the presence of impact did not necessarily cause the methods to perform worse, because of the contribution of altering TIF targets simultaneously.

Assuming commensurate targets, a comparison of type I error rates between Table 8 (i.e., no impact) and Table 9 (i.e., impact present) shows that the presence of impact caused the LR lasso method to be too sensitive in detecting DIF by having led to much higher type I error rates. Also, as observed in Table 13, LR lasso hit rates suffered a smaller amount in the presence of impact. In fact, the change in hit rates between Table 12 (i.e., no impact) to Table 13 (i.e., impact present) was more drastic for LR lasso than GMH and GLR.

Interestingly, in viewing the same set of tables mentioned above, having unbalanced sample sizes actually lessened the effects of increased type I error rates. When there are ability differences in the focal groups (impact condition) there was an increase in type I error. Type I error did not increase as much when the groups were unbalanced. That may have occurred because, in the unbalanced condition, there were fewer people with the largest impact.

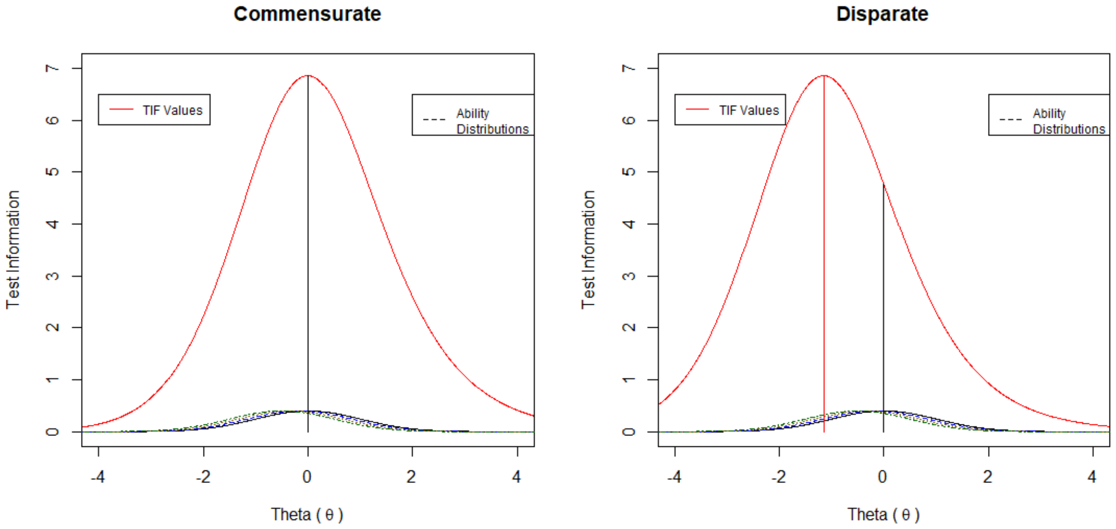


Figure 16. Correspondence of the Generating Ability Distributions and the TIFs by Presence of Impact.

Practical Scenarios

Five scenarios are presented within this section to provide in-depth illustrations of when practitioners may reasonably consider the merits of using the LR lasso method over GMH and GLR. While the scenarios are not intended to be exhaustive by any means, they provide a starting place for understanding why a more complex methodology may be considered for potential changes in a testing program.

Scenario One: A test is translated and administered in multiple languages, with no DIF or impact present across four linguistic groups.

Scenario one represents an ideal scenario where performance in detecting DIF is not confounded by multiple variables. This example is specified in Table 21 and could be understood as an achievement test that was translated into multiple languages and DIF is being tested across four linguistic groups. As such, the TIF is narrow near the central tendency of the simulees and no true DIF items with no impact observed across balanced samples. Given that the scenario has no DIF items (and thus DIF location is not applicable), Table 22 represents 500 replications. Table 22 shows that there were no true differences across methods (i.e., about two items misidentified per form).

Table 21. Specification of Condition Levels for Scenario One.

Condition	Level
Test Type	Commensurate
TIF Shape	Narrow
DIF Location	--
% of DIF Items	No DIF Items
Impact	None {.00, .00, .00, .00}
Sample Size	Balanced {500, 500, 500, 500}

Table 22. Simulation Results across 500 Replications for Scenario One.

Evaluation	GMH	GLR	LR lasso
Correct Classification Rate	37.96	38.18	37.97
Type I Error Rate	5.09 %	4.54 %	5.08 %

Scenario Two: An achievement test is given in a single language in K-12, and there truly are no DIF items on the given test form. However, impact is present across four ethnic groups.

Scenario two represents a more realistic scenario (when compared with scenario one) where performance in detecting DIF is confounded by the presence of impact. This example can be found in Table 23, and could be exemplifying a K-12 achievement test that is given to a diverse population of students and DIF is being tested across four ethnic groups. Similar to scenario one, the TIF is narrow near the central tendency of the simulees and no true DIF items observed across balanced samples. However, impact is present. Given that the scenario has no DIF items (and thus DIF location is not applicable), Table 24 represents 500 replications. Table 24 shows that LR lasso had slight degradation in type I error. The presence of simulee impact caused the focal groups to drift further from the maximum TIF, and subsequently reduced the variability of responses from those subgroups collectively. Also, the performance of GMH and GLR did not suffer in the presence of impact (compared with Table 22), but LR lasso did more so.

Table 23. Specification of Condition Levels for Scenario Two.

Condition	Level
Test Type	Commensurate
TIF Shape	Narrow
DIF Location	--
% of DIF Items	No DIF Items
Impact	Half Logit {.00, -.17, -.33, -.50}
Sample Size	Balanced {500, 500, 500, 500}

Table 24. Simulation Results across 500 Replications for Scenario Two.

Evaluation	GMH	GLR	LR lasso
Correct Classification Rate	37.96	37.97	37.54
Type I Error Rate	5.09 %	5.08 %	6.16 %

Scenario Three: DIF may be induced on a K-12 exam by new item types or revised content standards.

State education agencies may frequently attempt to shift curricular focus by adopting updated sets of content standards for instructional purposes. Revisions to such content standards usually require significant changes to a state’s general assessment system (and sometimes alternate assessment system, depending upon the changes required in extended content standards). Sometimes in the midst of such changes, there may be strong stakeholder interest for including new item types (e.g., technology enhanced items) on updated forms of the assessment. Scenario three captures how this study's simulation conditions could be specified to mimic such a situation.

Table 25 describes a test designed to have a narrow test target commensurate with the simulee population location, where impact across four simulee groups of equal size is observed but the test has two DIF items near the information target. Table 26 summarizes the findings across 250 replications. Overall, GMH and GLR had similar performance. LR lasso had a higher type I error rate, likely due to the presence of both DIF items and simulee impact.

Given the presence of two DIF items, Figure 17 contains two boxplots that capture the magnitude of the LR lasso penalty parameters (λ) depending upon whether or not the items that were flagged were truly generated as having DIF or not. An interesting trend emerges where items that truly exhibited DIF tended to have negative λ values, and false positive items tended to have positive λ values roughly half of the time. Given that the model estimates λ values agnostically of whether or not items truly exhibit DIF, it was likely not an issue of bias in the estimates. In fact, the finding was not surprising because the nature of false positives could be such that the regularization of the group-specific parameters could be tuned to where one or more focal groups appear to be favored on the items. In other words, the negative values on the true DIF items was signal that was consistent with the data generation model, whereas the false positive values are randomly distributed around zero, given the commensurate TIF target.

Table 25. Specification of Condition Levels for Scenario Three.

Condition	Level
Test Type	Commensurate
TIF Shape	Narrow
DIF Location	Near Information Target
% of DIF Items	5 % (i.e., 2 items)
Impact	Half Logit {.00, -.17, -.33, -.50}
Sample Size	Balanced {500, 500, 500, 500}

Table 26. Simulation Results across 250 Replications for Scenario Three.

Evaluation	GMH	GLR	LR lasso
Correct Classification Rate	37.98	37.77	36.82
Type I Error Rate	4.69 %	5.11 %	6.04 %
Hit Rate	93.00 %	90.80 %	61.80 %
Phi Correlation with Truth	.70	.66	.49

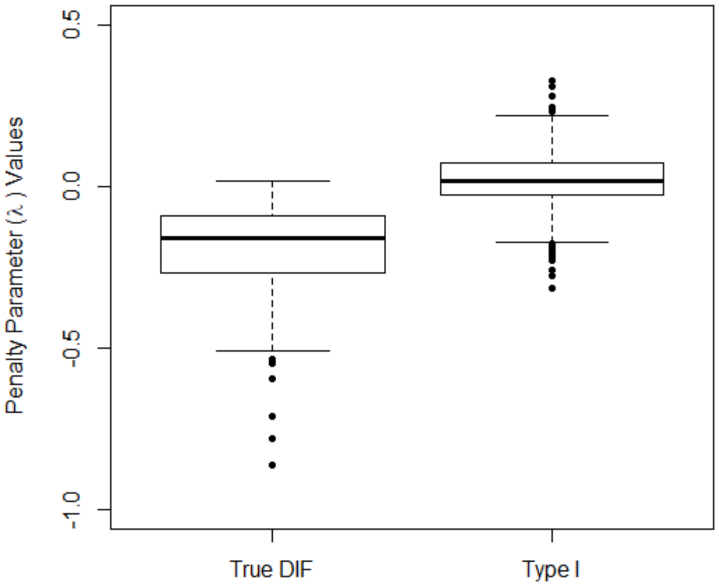


Figure 17. Magnitude of Penalty Parameters by Classification Type for Scenario Three.

Scenario Four: A new technology-enhanced item type could introduce construct-irrelevant variance that causes an item simultaneously to be more difficult and exhibit DIF.

The type of test described in Scenario Four could be defined in terms of the condition instantiations provided in Table 27. To describe the table, the scenario could be a test designed to provide measurement precision where examinees are located, but given some newer innovative item types (hopefully in field test positions), there appears to be DIF in the items while they are simultaneously more challenging items. As may be expected in much of K-12 testing, for example, there is observed impact and unbalanced sample sizes. Table 28 shows that LR lasso decreased in overall accuracy considerably (with the CCR and φ) when compared with GMH and GLR, which appeared to be an artifact of fewer items being flagged (with the decreased type I error rate and hit rate). Both the DIF location and unbalanced sample sizes could be suspected as influential, and a comparison with Table 5 shows that the very presence of DIF items caused the accuracy of LR lasso to drop consistently. Thus, it appeared that global model fit was impacted by the noise introduced by the two DIF items. This effect is better understood when considering that the DIF items were most likely too difficult for the focal groups to respond correctly, and still really difficult even for the reference group. The subsequent variance restriction was particularly taxing on LR lasso because the regularization was being performed on item scores that began to behave more like a constant. Thus, fewer items tended to be flagged for DIF, including those that may truly be exhibiting DIF. Similar to Figure 17, the penalty parameters in Figure 18 were congruent with the data generation model having a commensurate TIF target.

Table 27. Specification of Condition Levels for Scenario Four.

Condition	Level
Test Type	Commensurate
TIF Shape	Narrow
DIF Location	Offset above Information Target
% of DIF Items	5 % (i.e., 2 items)
Impact	Half Logit {.00, -.17, -.33, -.50}
Sample Size	Unbalanced {800, 600, 400, 200}

Table 28. Simulation Results across 250 Replications for Scenario Four.

Evaluation	GMH	GLR	LR lasso
Correct Classification Rate	37.83	37.76	36.98
Type I Error Rate	5.01 %	5.10 %	4.71 %
Hit Rate	91.60 %	89.80 %	43.40 %
Phi Correlation with Truth	.67	.66	.43

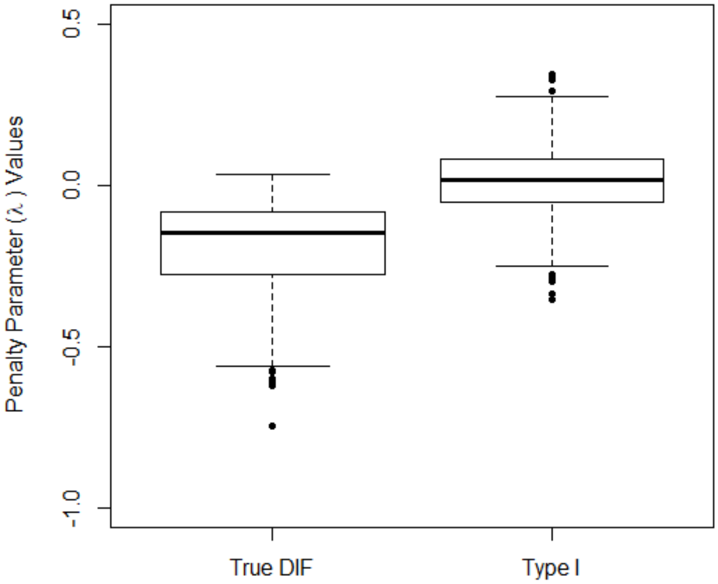


Figure 18. Magnitude of Penalty Parameters by Classification Type for Scenario Four.

Scenario Five: A cohort of freshman students take an end-of-course exam for an orientation course/seminar (e.g., University 101) at their university. The exam was written to be relatively easy for the majority of students, but there are DIF items on the exam (yet there is no examinee impact).

Scenario five contains a disparate test target in the midst of 10% of the exam being comprised of items that exhibit DIF. Table 29 contains further specifications of a narrow TIF shape, with all DIF items residing near the TIF target, for four groups with balanced sample sizes performing equally well. Table 30 summarizes the results across 250 replications. GMH and GLR performed similarly, but GMH had a slightly higher type I error rate. On the other hand, LR lasso performed appreciably worse than GMH and GLR. Scenario five arguably has fewer confounds with no impact, sample sizes are balanced, TIF shape is narrow, and DIF resides near the TIF target. Consistent with earlier results, having an increasing amount of DIF items disproportionately impacted LR lasso. This finding was especially true given the disparate test target.

Unlike Figures 17 and 18, the penalty parameter values in Figure 19 were similar between true DIF items and false positive items. While somewhat counterintuitive, the explanation provided for the disparate target plot in Figure 16 applies to this scenario. The impact observed across simulee groups shortened the gap between much of the simulee population and the TIF target, especially given the balanced sample size. Stated differently, there were more simulees near the TIF target that should have been offset had the groups been equal otherwise. The more apparent distinction between true DIF items and false positive items progressively disappeared because variance restriction was most noticeably observed in the reference group, while item score variability increased for the focal groups.

Table 29. Specification of Condition Levels for Scenario Five.

Condition	Level
Test Type	Disparate
TIF Shape	Narrow
DIF Location	Near Information Target
% of DIF Items	10 % (i.e., 4 items)
Impact	None {.00, .00, .00, .00}
Sample Size	Balanced {500, 500, 500, 500}

Table 30. Simulation Results across 250 Replications for Scenario Five.

Evaluation	GMH	GLR	LR lasso
Correct Classification Rate	37.37	37.54	35.65
Type I Error Rate	5.87 %	5.03 %	7.02 %
Hit Rate	92.90 %	88.90 %	61.50 %
Phi Correlation with Truth	0.74	0.74	0.51

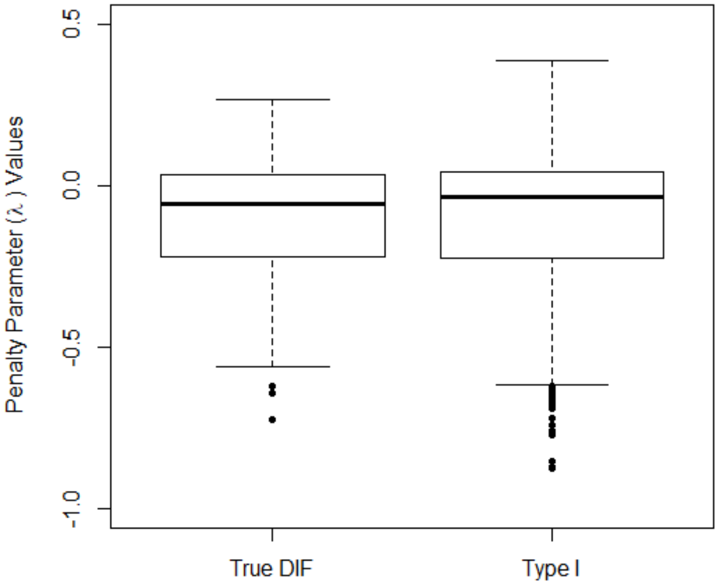


Figure 19. Magnitude of Penalty Parameters by Classification Type for Scenario Five.

Recommendations

The results suggested that the LR lasso method had inflated type I error overall with no additional benefit in power. In fact, even when type I error rates were comparable across methods, LR lasso had a lower hit rate in many instances (i.e., higher type II error rate). The sensitivity of LR lasso to detecting DIF items seemed to be substantially influenced by having an increased number of DIF items on a form. That is, the increasing presence of DIF items decreased the chances of flagging them accurately. This finding was not surprising when considering the global specifications in using LR lasso to fit a given data set. Noise introduced by DIF (and impact) could not be correctly partitioned into the penalty terms (λ) via the shrinkage estimator, and the noise was likely impacting the quality of the parameter estimates in other parts of the model more globally (e.g. model terms for item difficulty and test score). Furthermore, conditions which decreased variability in item scores (e.g., an item becoming too easy or too difficulty with respect to the examinees) also led to diminution in the lasso regularization, and the method increasingly failed to yield non-zero penalty parameters (λ) because the data began to behave more as a constant.

To be clear, the study results should not be interpreted as fully supporting that LR lasso is without merit in application. Across all conditions and replications that contained any simulated DIF items, LR lasso provided equal or superior hit rates in 40.57% of the simulated tests (or replications). Furthermore, LR lasso provided equal or lower type I error rates in 54.76% of the simulated tests, with 33.00% having lower type I error rates. However, LR lasso was more likely to flag an excessive number of items as having DIF, with 5.38% of replications across all conditions (both those free of DIF items and those containing DIF items) having eight or more items (20% or more of the total test length) falsely flagged for DIF.

As observed in multiple plots for scenarios with commensurate TIF targets, there was a proclivity for λ parameters to be positive roughly half of the time when items are falsely flagged for DIF. In theory, it is possible that a correction could be implemented to decrease type I error in one of two ways: (1) constrain the λ parameters during regularization to prevent positive values (which would be a non-trivial task and an arbitrary constraint), or (2) implement a *post hoc* correction of positive values that replaces them with zero. However, caution is strongly urged that trying to advantageously use such a trend (based upon simulated data alone) is without theoretical basis and would introduce risk to potentially increasing type II error because a small proportion of true DIF items have positive λ values. In other words, attempting to correct for type I error on the basis of positive λ values is not prudent and could have a slight adverse impact on type II error rates. A more statistically sound approach would be to use BIC instead of WIC to achieve a more parsimonious/conservative solution.

Considerations of Effect Size Measures

Currently, the measurement field does not have any effect size indices upon which to interpret multiple group DIF results. Largely, the decision for determining if an item exhibits DIF is based upon tests of statistical significance. To illustrate, the log-odds ratio (alpha) exists to describe DIF magnitude for two-group comparisons via MH, but a similar metric does not exist for GMH. The LR lasso method at least avoids the dependence upon significance testing within the context of multiple groups with the criterion of a non-zero group parameter. While not being a hypothesis test, per se, the criterion is based upon a difference from zero, which is conceptually similar to a hypothesis test and may not provide entirely different results (as was reflected in the simulation results too).

A few complications arise in deriving an effect size based measure to describe the findings of a test for DIF among multiple groups. First, having multiple focal groups requires a procedure that compares each focal group with the reference group. This comparison could be done separately (i.e., pairwise) or simultaneously (i.e., multivariate/matrix-based calculations). If a pairwise procedure would be implemented, then potentially useful data between focal groups amongst themselves would need to be considered in the analysis. Otherwise, possibly meaningful information would be ignored for the sake of a traditional interpretation, which is not a prudent use of having additional data available. If a matrix-based computation is considered, care must be taken to ensure a standardized result. It is possible that an index conceptually similar to Cohen’s *d* (or Mahalanobis distance, or the ETS delta scale) could be computed using a matrix of mean differences that are post-multiplied by an inverse covariance matrix. Though, unless the measure of effect size(s) could be reduced to a single value, interpretation becomes complex with the need for conjunctive or disjunctive criterion for discerning when an item may or may not be exhibiting DIF. And if reduced to a single index, then *post hoc* explorations would be needed. Additionally, there is no guarantee that any matrices based upon the assessment data would meet the requirements for definiteness and being non-singular needed for various methods.

Other options could exist as well. For instance, decomposition/factorization of matrix-based DIF information (and not like a singular value decomposition of scored assessment data like that done in more traditional dimensionality analyses) could eliminate the complexity of conjunctive or disjunctive criteria. However, simply having a flag for DIF does not indicate wherein the DIF lies among the multiple groups and *post hoc* procedures would be still required (much like those for ANOVA tests). While the GLR and LR lasso are suitable regression-based procedures, other regression-based procedures such as log-linear models could be used to analyze multivariate contingency table counts (assuming Poisson-distributed error terms). Other more traditional effect size measures for contingency tables, such as Phi (φ) and Cramer's *V*, do not have optimal statistical properties because ranges of the indices are influenced by distributions of the data.

Finally, the multiple groups DIF magnitude could be considered on a distance metric. For example, a distance measure could be used to explain relative differences between log odd ratios for each group. Doing so would capitalize on using contingency table information, while still maintaining an interpretation of log-odds that could be based on an effect size metric. Additionally, such a technique could allow for a more complex distance (e.g., Mahalanobis distance) to generalize to simpler distances for interpretation (e.g., normalized Euclidean distance) in the case of two groups. Moreover, such an approach could more globally consider relative differences in the log-odds ratios between groups across items, which concurrently allows for DIF to be interpreted as the relative difference in item difficulties between the multiple groups (similar to the philosophy advocated by Bechger and Maris, 2015), thereby avoiding the usual issue of using the total score (which is always related to the items which are being tested for DIF) for DIF detection. DIF magnitude could be classified according to the number of discrepant focal groups, the degree to which DIF exists for each focal group, as well as the direction of DIF.

Limitations and Future Research

As mentioned previously in this document, item purification was not performed with GMH and GLR. Inherently, the lasso constraints in LR lasso are a type of selection procedure, and comparisons with GMH and GLR using item purification adds additional selection procedures. As such, item purification would have presented itself to be a confound in this study given that the merits of using the lasso methodology needed to be compared against baseline performance (which is not having a selection/purification process in a multiple group setting), but perhaps it would not be confounding in future studies if the purpose of such studies examine specific scenarios where there may be expected differences between selection procedures.

Another limitation to the current study is that the generation models only included uniform DIF that is unidirectional/asymmetric with respect to groups. While this type of DIF is commonly observed, it is difficult to support using the results from this study to generalize more broadly to various types of DIF (non-uniform DIF and/or symmetric DIF). As an example, crossing non-uniform could be generated using the U3PL in a future study in such a way that the DIF-inducing discrimination and difficulty parameters are chosen to allow uniform and non-uniform DIF to yield comparable levels of bias (Swaminathan & Rogers, 1990; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). To clarify further, changes in discrimination requires a scaling adjustment to the difficulty parameters to yield nearly equivalent differences in ICCs across groups (if this definition of DIF is used to establish equivalent amounts of DIF). Magis, Tuerlinckx, and De Boeck (2015) note that LR lasso is more akin to the Rasch model than the U2PL because the examinee's total score used in the prediction is not a weighted sum score, so manipulating DIF through the discrimination parameter in the generating model would be a fruitful investigation to explore the effects on DIF detection imposed by the lasso constraints. This investigation could be coupled with further exploring convergence issues in terms of variability of the λ parameters.

The effect of non-normal ability distributions on LR lasso should also be explored (either for two groups or multiple groups). McLaughlin and Drasgow (1987) discussed that the sphere of generalization for the results of studies are often limited to samples containing normally distributed abilities, which is the case in this study. Moreover, strict normality is difficult to find in practice (Micceri, 1989). Given the warranted nature of non-normality, there are multiple ways it could be explored. If the data generation model is unidimensional, the true ability distributions could be based upon the standard Gaussian distribution and a rescaled Beta distribution with skew of -.75 (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005), and could be defined as 6*(rbeta(N, α=2.57, β=1) - .5) in order to correspond with roughly an 83% pass rate. If the data generation model is multidimensional, the true ability distributions could be based upon the multivariate standard normal distribution and a rescaled dirichlet as a multivariate generalization of the beta distribution suggested above for the unidimensional model.

Another limitation with the current study is that the generation model is unidimensional, when data are often impacted by multi-faceted sources of variability in reality. Therefore, data generation with a MIRT model could be explored (e.g., two-dimensional MIRT 3PL). Under such a study design, the first dimension is the construct intended to be measured, and the second dimension is a nuisance dimension that perhaps contributes the DIF. However, care should be taken to ensure that DIF is a property of the item parameters, and not necessarily created in the examinee ability values. Other variations of the MIRT 3PL with differing numbers of nuisance dimensions could be explored in this context, with differing levels of association assumed between both intended and nuisance dimensions.

Conclusion

Taken collectively, the results of the simulation study can be interpreted to support the claim that LR lasso failed to perform comparably with more established methods for multiple groups DIF detection across numerous instances but could potentially have merit in practical application in situations that have yet to be explored. While some limitations of LR lasso were noted within this study, there are a variety of other conditions which need to be explored before practitioners discard the method altogether for use in multiple groups contexts (a few such studies were suggested above). It may well be the case that the added complexity afforded by the regularization in estimating the group-specific model parameters through lasso constraints may confound the detection of the DIF items.

The merits of this study, ultimately, are two-fold. First, LR lasso simply did not perform on par with other, more traditional DIF-detection methods such as the GMH and GLR methods, even under rather ideal measurement conditions. This finding suggests that further investigation of the LR lasso method within the context of multiple groups may be equally discouraging. Second, methodologically speaking, this study demonstrates the need for DIF research to consider multiple factors. Those factors include: (a) the measurement properties of the scale relative to the test purpose and to the reference and focal group sampling distributions and (b) the specific characteristics of the DIF-impacted items relative to both the examinees and to the score scale properties. That is, it seems naïve and certainly an oversimplification of reality to merely consider sample sizes and various magnitudes of proficiency score differences between reference and focal groups. Nor should it be exclusively about the number of items chosen to have DIF. DIF research should be about a complex system of sampling, scale properties, and item design and psychometric characteristics that need to be considered simultaneously, regardless of which DIF-detection methods are being compared. In that respect, this study provides an important example of how to include many of those factors in a study and then tease out pragmatically relevant findings.

While the global specification of the LR lasso model seemed to be a promising attribute of the method, there was support from this study to believe that it could be at the crux of the classification errors that were observed. The method may be better supported under the conditions originally explored by Magis, Tuerlinckx, and De Boeck (2015), which included dichotomously scored data and two groups. Ultimately, while the accurate detection of DIF items is paramount to psychometrics, efficient and effective preventative measures during item and test development should have precedence and ideally lead to situations where many limitations of DIF methodology are never observed incipiently.

REFERENCES

Agresti, A. (2002). Categorical Data Analysis, Second Edition, New York: John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6): 716-723

American Board of Internal Medicine. (2015). First-Time Taker Pass Rates -Initial Certification. Retrieved from: http://www.abim.org/~/media/ABIM%20Public/Files/pdf/statistics-data/certification-pass-rates.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95-105.

Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement*, *34*, 807-816.Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277-300.

Baudrillard, J. (1994). Simulacra and simulation. Ann Arbor: University of Michigan Press.

Ben-Shakur, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 25*(1), 23-35.

Benítez, I., & Padilla, J.-L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, *8*(1), 52-68.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT Differential Item Performance for Nine Handicapped Groups. *Journal of Educational Measurement, 24*(1), 41-55.

Box, G. E. P., & Draper, N. R., (1987). *Empirical Model Building and Response Surfaces,* John Wiley & Sons, New York, NY.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* (2nd ed.). New York, NY: Springer-Verlag.

Camilli, G. (1979). A critique of the chi square method for assessing item bias. Unpublished manuscript, University of Colorado, Laboratory of Educational Research, Boulder.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioural Statistics*, *4,* 323-341.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253-260.

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

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*, 269-279.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Guilford Press.

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*(2), 149-170.

DeMars, C. E. (2010). Type I Error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*(6), 961-972.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational and Behavioral Statistics*, *18*(2), 131-154.

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: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355-368.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77,* 177-184.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. New Jersey: Erlbaum.

Fidalgo, Á. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research*, *5*, 43-53.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). DIF detection using several statistical procedures: Implications on the type I and type II error rate. *The Journal of Experimental Education, 73*, 23-39.

Finch, W. H., & French, B. F. (2007). Detection of Crossing Differential Item Functioning: A Comparison of Four Methods. *Educational and Psychological Measurement*, *67*(4), 565-582.

Finch, W. H. (2016). Detection of Differential Item Functioning for More Than Two Groups: A Monte Carlo Comparison of Methods. *Applied Measurement in Education*, *29*(1), 30-45.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*(3), 315-332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1-22.

Gomez-Benito, J., & Navas-Ara, M. J. (2000). A comparison of χ2, RFA and IRT based procedures in the detection of DIF. *Quality and Quantity*, *34*(1), 17-31.

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*(3), 314-329.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, *41*, 190-195.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Holland, P. W. (1985). *On the study of differential item performance without IRT. Proceedings of the 27th Annual Conference of the Military Testing Association* (Vol. 1, pp. 282-287). San Diego.

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

Howell, D. C. (2010). Statistical Methods for Psychology. 7th ed. Cengage Wadsworth Belmont, CA, USA.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, *76*, 297-307.

Jenson, A. R. (1980). Bias in mental testing. New York: The Free Press.

Kamata, A., & Vaughn, B. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal*, *2*, 48-69.

Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, *15*(3), 269-278.

Kim, J., & Oshima, T. C. (2013). Effect of Multiple Testing Adjustment in Differential Item Functioning Detection. *Educational and Psychological Measurement, 73*(3), 458-470.

Kok, F. G., Mellenbergh, G. H., & Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement*, *22*, 295-303.

Kristjansson, E., Aylesworth, R., McDowell, I, & Zumbo, B. D. (2005). A comparison of four methods for detecting DIF in ordered response items. *Educational and Psychological Measurement*, *65*, 935-953.

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests. *International Statistical Review / Revue Internationale De Statistique*, *46*(3), 237-254.

Lautenschlager, G. J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement*, *12*, 365-376.

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

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72*(5), 847-861.

Li, Z. (2015). Power and Sample Size Calculations for Logistic Regression Tests for Differential Item Functioning. *Journal of Educational Measurement, 51*(4), 441-462.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75,* 164-174.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning. *Behavior Research Methods*, *42*(3), 847-862.

Magis, D., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research, 46,* 733-755.

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, *11*(4), 365-386.

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of Differential Item Functioning Using the Lasso Approach. *Journal of Educational and Behavioral Statistics, 40*(2), 111-135.

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.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement*, *11*, 161-173.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105-118.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*, 289-311.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement*, *41*, 331-344.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, *16*, 381-388.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90-100.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257-274.

Oshima, T. C., Wright, K., & White, N. (2015). Multiple-Group Noncompensatory Differential Item Functioning in Raju's Differential Functioning of Items and Tests. *International Journal of Testing, 15*(3), 254-273.

Penfield, R, & Lam, T, (2000). Assessing differential item functioning in performance assessment: review and recommendations. *Educational Measurement: Issues and Practices*, *19*(3), 5-16.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*, 235-259.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353-370.

Penny, J., & Johnson, R. L. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel chi-square DIF index. *The Journal of Experimental Education, 67*(4), 343-366.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, *8*, 137-156.

Plake, B. S., Patience, W. M., & Whitney, D. R. (1988). Differential item performance in mathematics achievement test items: Effect of item arrangement. *Educational and Psychological Measurement*, *48*, 885-894.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.

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.

Reckase, M. (2009). Multidimensional Item Response Theory. New York: Springer.

Revolution R Enterprise version 8.0 (64-bit): an enhanced distribution of R Microsoft packages Copyright (C) 2015 Microsoft Corporation.

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

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

Ryan, K. E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal of Educational Measurement*, *28*(4), 325-337.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*, 143-152.

Scheuneman, J. D. (1987). An experimental exploratory study of causes of bias in test items. *Journal of Educational Measurement*, *24*, 97-118.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *American Statistician, 40*, 106-108.

Spray, J., & Carlson, J. (1986). Comparison of loglinear and logistic regression models for detecting changes in proportions. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Spray J. A. (1989). Performance of three conditional DIF statistics in detecting differential item functioning on simulated test. *ACT Research Report Series*, 89-7.

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

Stricker, L. J. (1981). A new index of differential subgroup performance: Application to the GRE Aptitude Test (GRE Board Professional Report 78-7P; ETS Research Report 81-13). Princeton, NJ: Educational Testing Service.

Stricker, L. J., & Emmerich, W. (1999). Possible determinants of differential item functioning: Familiarity, interest, and emotional reaction. *Journal of Educational Measurement*, *4*, 347-366.

Su, Y.-H., & Wang, W.-C. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*(4), 313-350.

Subkoviak, M. J., Mack, J. D., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement*, *21*, 49-58.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of differential solution strategies. *Journal of Educational Measurement*, *25,* 301-319.

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

Tian, F. (1999). Detecting differential item functioning in polytomous items. Unpublished doctoral dissertation, Faculty of Education, University of Ottawa.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*(1), 21-43.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement*, *18*, 15-25.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443-464.

Van der Flier, H. (1980). Vergelijkbaarheid van individuele testprestaties. Netherlands: Swets & Zeitlinger.

Van der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement*, *21*, 131-145.

van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*. New York: Springer Science+Business Media, Inc.

von Davier. M. (2009). Is There Need for the 3PL Model? Guess What?. *Measurement: Interdisciplinary Research and Perspectives*, *7*(2), 110-114.

Wald, A. (1939). Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *The Annals of Mathematical Statistics*, *10*(4), 299-326.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17,* 113-144.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479-498.

Welch, C., & Hoover, H. D. (1993). Procedures for Extending Item Bias Detection Techniques to Polytomously Scored Items. *Applied Measurement in Education*, *6*(1), 1-19.

Whitmore, M. L., & Schumacker, R. E. (1999). A Comparison of Logistic Regression and Analysis of Variance Differential Item Functioning Detection Methods. *Educational and Psychological Measurement*, *59*(6), 910-927.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, *3*, 23-40.

Wright, B. D. (1991). Rasch vs. Birnbaum. *Rasch Measurement Transactions*, *5,* 178-179.

Zwick, R., Donoghue, J. R. & Grima, A. (1993). Assessing Differential Item Functioning in Performance Tests. ETS Research Report Series, 1993: 1-42.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, *21*, 187-201.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10,* 321-344.