A RANDOMIZATION TEST FOR THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING

by

David Rockoff

A Dissertation Submitted to the Faculty of the

GRADUATE INTERDISCIPLINARY PROGRAM IN STATISTICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2018

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by David Rockoff, titled A Randomization Test for the Detection of Differential Item Functioning and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Nicole Kersting Date: April 27, 2018

Katherine Barnes Date: April 27, 2018

Hao Zhang Date: April 27, 2018

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director: Nicole Kersting Date: April 27, 2018

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: David Rockoff

ACKNOWLEDGEMENTS

I would like to show my gratitude to my advisor Dr. Nicole Kersting for her continuous support, encouragement and patience. I would also like to thank my committee members, Dr. Katherine Barnes and Dr. Helen Zhang, for serving as my committee members and providing valuable suggestions. Much thanks is also due Dr. Deborah Levine-Donnerstein, who served as a committee member until her retirement; and David Magis, who provided R code and insight.

Special commendations go to my family for cheering me on year after year.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

A test question, in the context of educational or psychological measurement, is said to possess Differential Item Functioning (DIF) when the probability of a test-taker answering correctly depends in part on group membership, even after controlling for ability level. DIF has historically been viewed as item bias, an undesirable property. Several statistical methods exist for examining test items for DIF, but none performs optimally across a wide range of settings in distinguishing DIF items from non-DIF items.

In this study a randomization test is implemented for DIF detection. Computer simulations are conducted to compare the randomization test's performance to the performance of three existing DIF detection methods, namely the Mantel-Haenszel, Logistic Regression and Lasso methods. Methods are compared by way of receiver operating characteristic curves.

The Randomization Test strongly outperforms all other studied methods across most settings, by amounts that are statistically significant and large enough to be of substantive relevance, when the examinee groups are of equal ability and scale linking is not employed in the Randomization Test. The Randomization Test with linking generally performs slightly worse than the Lasso and Logistic methods, and slightly better than the Mantel-Haenszel method, by amounts that are statistically significant but perhaps not large enough to be of substantive relevance.

CHAPTER 1

INTRODUCTION

This study implements a randomization test to detect differential item functioning (DIF). The research evaluates the randomization test's performance relative to three existing DIF detection methods: Mantel-Haenszel, Logistic Regression, and the recently developed Lasso regression. The randomization test is implemented within an item response theory context, specifically the Rasch model, in which examinee abilities are parameters to be estimated. The other studied methods use total score as a measure of ability.

The standard statistical methods to evaluate for DIF have been shown lacking. This study will show that the use of the randomization test provides a better tool for measurement of DIF in some broad cases.

An exploration of DIF, its importance, extant methods for analyzing it, and the difficulties in analyzing it, establish the ground upon which the research follows. The use of the randomization test in DIF detection is motivated by the traditional methods' often untenable assumptions regarding the sampling distribution of DIF statistics, an issue which has not previously been adequately addressed. The current study addresses this issue in an item response theory framework although a similar approach could be taken utilizing total score methods.

1.1 Problem statement and background

Test items (questions) are designed so that responses can be used to make inferences about the level of some underlying latent (not directly observable) skill or psychological construct of interest possessed by the test takers. Examples of such latent traits may be mathematical ability, sociopolitical attitudes, or depression. Understanding the properties of individual test items, and of the test in its entirety, is critical for understanding the meaning of the scores derived from the responses.

A test item is usually intended to function the same for all respondents who have the same level of the latent trait, regardless of the subpopulations to which they belong. Conversely, a test item is said to contain differential item functioning (**DIF**) if examinees who have the same latent trait level, but are from different pre-defined subpopulations or groups, give systematically different responses to the item. The wording of a question on an algebra exam, for instance, may cause female examinees to have a lower probability of a correct response than males of equal algebra ability. In such a case, the question is measuring some extraneous examinee attribute (gender) in addition to the attribute of interest (algebra ability).

An example of a DIF item is this analogy question from the Verbal section of the 1997 Scholastic Aptitude Test:

DECOY : DUCK ::
(A) net : butterfly
(B) web : spider
(C) lure : fish
(D) lasso : rope
(E) detour : shortcut.

Males were 15-20% more likely than females to answer this correctly at any given level of verbal ability as measured by performance on the Verbal section as a whole. The question inadvertently

measures gender, or some extraneous variable associated with gender (such as knowledge of hunting and fishing), in addition to verbal analogy skills.

Statistically speaking, an item contains DIF if examinees who have the same latent trait level θ (ability or proficiency) but are from different subpopulations have a different probability of giving a certain response to the item. A dichotomously scored item (Correct/Incorrect) contains DIF if

$$P(\text{Correct response} \mid \theta,\ \text{Group} = \text{G}) \neq P(\text{Correct response} \mid \theta,\ \text{Group} \neq \text{G})$$

where G is an observable group membership variable.

The groups being compared are typically referred to as the reference group (baseline or majority) and the focal group or groups. The item is said to contain uniform DIF if the group differential is constant across all levels of the ability spectrum; otherwise it is said to contain non-uniform DIF.

The ability to determine whether an item contains DIF is important for at least two reasons. First, many life events are steered by performance on high-stakes standardized exams, such as college admission and employment. If an item is biased in the traditional sense, it systematically results in differential performances depending on race, gender, or some other characteristic not of direct interest, beyond what may be attributed to actual differences in ability. The result is that scores will be artificially lower for some subpopulations. Hence, DIF detection is important in reducing bias and providing evidence for an accurate interpretation of scores.

Second, DIF analysis is increasingly being used to detect item-level treatment effects. For example, educators have used DIF analysis procedures to investigate whether different instructional practices lead to different responses to particular items (see for example Klieme and

Baumert 2001). Viewed through this lens, DIF can be regarded as an item characteristic of direct interest rather than a nuisance property to be corrected or controlled for.

The present study focuses on dichotomous items, and comparisons between only two respondent groups. It should be noted however that DIF can occur in any type of test item, and with any number of respondent groups greater than one. For ease of understanding, educational testing terminology such as ability level and correct/incorrect is used, although the framework applies equally to psychological and other tests where some other trait besides ability is being measured, and items are either endorsed or not endorsed.

1.2 Existing DIF detection methods: Overview

Numerous statistical methods for detecting DIF have been developed throughout recent decades. Some DIF detection methods use total score as an estimate of examinee ability. In other methods – those that implement Item Response Theory (see Section 1.3) – abilities are parameters to be estimated. One requirement they all have in common is that examinees must be matched on ability level before any meaningful comparison can take place. A DIF statistic – some measure of the difference in item difficulty between groups – is then computed for each item, and a hypothesis test conducted. However, due to issues of circularity inherent in DIF analysis, described more fully in Section 1.4, the true null distribution of the DIF statistic may not be well approximated by the assumed distribution. The present study extends DIF analysis within the context of item response theory and addresses the issue of distributional assumptions that have not been addressed in existing DIF detection methods.

An overview of the DIF detection methods used in this study follows. Further details on these methods can be found in Section 3.1.

The most common method for detecting uniform DIF is the Mantel-Haenszel (MH) method (Mantel & Haenszel, 1959; Holland & Thayer, 1988), which utilizes a contingency table approach to compare groups. In the MH method, examinees are first stratified by total score. Within each stratum, the odds of success on an item is computed for each group. The MH test for conditional independence tests the hypothesis that the odds ratio is equal to 1 in all strata.

Another common DIF detection method is the Logistic Regression method (Swaminathan & Rogers, 1990). This method models the probability of correct response as a logistic function of item difficulty, respondent ability as measured by total score, and respondent group membership. If the group membership coefficient is significantly different from 0, the item is flagged as containing DIF.

Magis, Tuerlinckx and De Boeck (2014) created a novel DIF detection method involving Lasso regression (Tibshirani, 1996), a popular variable selection technique in which a penalty is applied to large regression coefficients. They based their Lasso on the common Logistic Regression method described above, but their method tests all items simultaneously.

1.3 DIF detection using a randomization test

The new approach explored in this study – the randomization test – is motivated by the desire to obtain a more accurate null distribution of the DIF statistic. Randomization tests (or "permutation tests") are a class of nonparametric tests in which group membership is randomly shuffled among experimental or observational units repeatedly to form the distribution of a test statistic under the null hypothesis. Sen (2014) notes that "[t]hey are most useful when we have insufficient information about the distribution of the data, are uncomfortable making assumptions about the distribution, or if the distribution of the test statistic is not easily

computed." In a randomization test, the null hypothesis is that group membership has no effect on response.

The randomization test at the heart of the current study is implemented within Item Response Theory (IRT). IRT is an increasingly popular framework for tests and questionnaires that involves the statistical relationship among item properties, examinee properties and item responses. Specifically, the present study utilizes the Rasch model (Rasch, 1960), one common IRT model for dichotomous items. Under the Rasch model, the probability that examinee i will correctly answer item j is given as

$$P(Y_{ij} = 1 \mid \theta_i,\ b_j) = \frac{exp(\theta_i - b_j)}{1 + exp(\theta_i - b_j)}\ \text{, where}$$

- $P(Y_{ij} = 1)$ is the probability of examinee i getting item j correct
- θ_i is examinee i's latent ability level
- b_j is the difficulty parameter for item j.

Ability levels and item difficulties are on the same (arbitrary) scale. They may best be thought of as z-scores. If it is the difficulty scale that is fixed, then an item with a difficulty level of 1.0 is one standard deviation harder than the average item, and an examinee must possess an ability level of 1.0 in order to have a 50% chance of getting such an item correct. Conversely, if it is the ability scale that is fixed, an examinee with an ability level of 1.0 is one standard deviation above the average examinee in ability on the skill being tested, and an item with a difficulty level of 1.0 is defined as one which such an examinee has a 50% chance of getting correct.

The θ_i's and b_j's are typically unknown in practice and must be inferred/estimated from the item responses. Parameter estimates are indeterminate unless constraints are placed, typically

by constraining either the ability or difficulty parameters to have mean of 0 and standard deviation of 1.

Under the Rasch model, an item contains uniform DIF if its difficulty parameter differs by examinee group. (Non-uniform DIF is not a consideration under the Rasch model since non-uniform DIF occurs when the slopes of the logistic curve differ between groups, and under the Rasch model all items are assumed to have slope 1.)

A randomization test in DIF analysis proceeds as follows: once responses have been generated, difficulty parameters are estimated separately for each examinee group, and the absolute difference calculated, giving an observed DIF parameter estimate for the item. Then the null, or reference, distribution of DIF parameter estimates is approximated by repeated random reassignment of group labels to the examinees and subsequent re-estimation of the DIF parameter, while holding the response vectors fixed. The observed DIF parameter estimate is then compared to the reference distribution. The two-tailed p-value is the proportion of DIF parameter estimates in the reference distribution that are greater in absolute value than the observed DIF parameter estimate.

Because randomization tests do not rely on any parametric form for the reference distribution, a randomization test should avoid the contamination issues inherent in current DIF methodology, described in the next section.

It should be noted that while there are many different IRT models, the current study is limited to data generated under the Rasch model, and the randomization test at the heart of this study involves parameter estimation assuming the Rasch model. While there are many other methods for DIF detection, this study's focus will be on the aforementioned ones as they are the

most commonly used both in practice and in research (in the case of MH and Logistic) or show promise as a relatively new method (in the case of Lasso).

1.4 Issues in DIF methodology

Traditional DIF detection methods involving statistical tests can be broadly classified into two categories: IRT-based and non-IRT-based. With IRT-based methods, examinee ability is a latent variable to be estimated along with item parameters, and groups are compared via differences in item parameter estimates. In non-IRT methods, total test score is used as a proxy for examinee ability, and groups are compared after controlling for total score. One requirement all DIF detection methods have in common is that examinees must be matched on ability level before any meaningful comparison can take place. Then a DIF statistic – some measure of the difference in item difficulty between groups – is computed for each item, and a hypothesis test conducted.

Circularity and contamination

A key challenge in DIF detection involves a sort of paradox: testing an item for DIF requires that all other items on the test be DIF-free. The very nature of DIF analysis requires that examinees be matched on ability before comparisons can be made. Ability estimates are usually based on internal criteria such as total test score or weighted score; but if any of the items contain DIF, then test scores will be biased, resulting in contaminated measures of ability. This contamination in turn leads to improper matching and thus contaminated DIF statistics.

A related challenge in IRT-based DIF analysis is scale indeterminacy. In most cases, item and person parameters are estimated jointly from response patterns, but the parameter scales are indeterminate unless constraints are placed on either the item parameters or the person

parameters. If no items are assumed *a priori* to be DIF-free, the usual approach is to estimate item parameters separately for each group. However, if an item's difficulty parameter estimates differ by examinee group, it is unclear whether the difference is due to DIF or to differences in group ability level distributions. A "linking" step must then be undertaken in order to place the groups onto the same metric and to control for group differences in ability. One common linking method involves constraining the mean item difficulty to be equal across groups; this is the linking method employed in the randomization test in the present study.

If the test contains any DIF items, however, parameter estimates will erroneously be based in part on these DIF items, resulting in improper linking. This circularity issue typically results in inflated Type I error rates: too many non-DIF items are classified as containing DIF (Magis, Beland, Tuerlinckx, & De Boeck, 2010).

This contamination issue results in unspecifiable null distributions for hypothesis testing. The reason is that when a single item is tested for DIF under the null hypothesis that the item is DIF-free, it is possible that other items contain DIF, which would contaminate ability estimates and consequently affect item difficulty estimates. Hence the exact sampling distribution of DIF estimates cannot be determined. Any parametric test of observed differences is therefore approximate unless all other items are known to be DIF-free, or unless estimates are based on a subset of items known to be DIF-free.

A potential solution to this dilemma is to base the matching or linking only on items known to be DIF-free ("anchor items"). Trouble still arises, however, if the anchor set consists of very few items: ability estimates may be unstable, resulting in unstable item parameter estimates for the tested items. In the (not implausible) event that no items are known *a priori* to be DIF-free, anchoring is still possible via a "purification" procedure, developed to identify items that

are "probably" DIF free. There is, however, no guarantee that only DIF-free items will be included in the anchor. (See Discussion for more on the issue of purification.)

Furthermore, most of the existing tests for DIF rely on sampling distributions that are only asymptotically known. Thus, even when the matching criterion can be assumed DIF-free, DIF tests may not be accurate if the number of respondents or items is small. A randomization test (Fisher, 1935) provides a simple way to approximate the sampling distribution of any test statistic and hence might offer a valuable approach to address this challenge.

Edgington and Onghena (2007, p. 289) have stated that historically, randomization tests were accepted as "the only true valid statistical tests of treatment effects, and that a parametric test […] was valid only to the extent that it provided similar p-values to those that would be given by a randomization test."

Multiple testing

An additional issue in DIF analysis is that of multiple testing. Because DIF analysis often involves testing many items, the probability of committing at least one Type I error – erroneously flagging as DIF at least one non-DIF item – may be undesirably high. Although adjustment for multiple testing "appears to be the exception rather than the rule […] in applied DIF literature" (Kim & Oshima, 2013), there have been studies investigating the efficacy of various adjustment procedures, such as Benjamini and Hochberg's (1995) procedure to control expected false discovery rates, Holm's procedure, and Bonferroni corrections. Holm's and Bonferroni's procedures are meant to control familywise Type I error rates.

The current study does not use any adjustments for multiple testing due to the employment of receiver operating characteristic (ROC) curves as the primary means of evaluating and comparing the effectiveness of DIF detection methods. ROC curves illustrate the

performance of a binary classifier system as the decision threshold varies, comparing sensitivity (ability to detect true positives) against specificity (ability to avoid false positives). Evaluation of a DIF detection procedure is fundamentally a binary classification issue: any given item is either DIF or non-DIF, and is classified as either DIF or non-DIF. This study evaluates how well the detection methods balance the ability to correctly classify DIF items as DIF with the ability to correctly classify non-DIF items as non-DIF.

This study's main goal is to evaluate how well each DIF detection method distinguishes between DIF items and non-DIF items, comparing false positive rate to true positive rate at a number of decision thresholds (e.g. significance levels or Lasso penalty parameters). False positive rate in this context refers to the conditional probability of classifying an item as DIF when it is non-DIF, which is equivalent to the per-test Type I error. Since the decision thresholds in an ROC curve are chosen by the researcher, adjustments for multiple comparisons are irrelevant; instead of comparing DIF detection procedures at each significance level from 0 to 1 in increments of 0.01, say, the researcher may use increments of 0.01/20 for a 20-item test, which does nothing to control Type I error but only yields slightly more precise ROC curves for comparison.

CHAPTER 2

LITERATURE REVIEW

The present study focuses on data simulation to compare DIF detection across a wide variety of settings. A germane literature review focuses on other studies where the researchers conducted simulations to compare DIF detection methods and/or to determine the factors that affect the ability to detect DIF.

The following sections summarize findings from existing simulation studies that compare DIF detection methods across a variety of conditions, as well as studies that investigate the effects of other factors on DIF detection. In order to maintain relevance to the current undertaking, this review is limited to studies in which data were generated under the Rasch model, items were dichotomous, and DIF detection was conducted with the Mantel-Haenszel (MH), Logistic and/or Lasso methods. While there also exists much DIF detection research involving empirical data such as large-scale national or international testing (see Klieme & Baumert, 2001 for example), such research will not be addressed here since it is only of tangential relevance to the current study.

Much of the simulation-based research in DIF detection has focused on the effects of the following factors:

- DIF detection method
- Percentage of items that contain DIF (DIF%)
- Sample size (number of examinees)

- Test length (number of items)
- Magnitude of DIF
- Impact (difference between groups in average ability)
- Type of DIF (uniform or non-uniform)

Understanding the effect of DIF detection method is the primary focus of the current research. The next five factors are also explored because they are manipulated to provide the specific conditions under which the performance of the different detection methods is evaluated.

In virtually all such research, the outcomes of interest are hit rate (ability to detect true DIF items) and false positive rate (erroneous classification of non-DIF items as containing DIF). Recently there has been increased usage of receiver operating characteristic (ROC) curves to measure the global performance of a DIF detection method (Magis, Tuerlinckx & De Boeck, 2014; Tutz and Schauberger, 2015; Schauberger and Tutz, 2016). In ROC curves, hit rates and false positive rates are combined into a single measure, area under the curve (AUC), for use as the outcome of interest. AUC is the primary outcome measure used in the present study.

A recurring theme across these studies is that many factors are involved in the effectiveness of DIF detection, and that the magnitude of the factors' effects varies greatly across detection methods and experimental conditions such as sample size, test length etc. Some general findings have been consistent throughout, however. For instance, the Mantel-Haenszel (MH) method is good at detecting uniform DIF, but fares poorly in detecting non-uniform DIF (Swaminathan & Rogers, 1990). For the most part, the IRT-based methods tend to have greater power as long as their stronger model assumptions are met (Sireci and Rios, 2013; Stout, 1990). DIF detection suffers when a high percentage of items contain DIF, groups differ in average ability, or DIF magnitude is small.

2.1 Studies comparing DIF detection methods

Simulation studies have shown that the relative performance of any DIF detection method depends in no small part on factors such as number of items containing DIF, number of examinees, and the type and magnitude of group differences. There are numerous studies that have directly compared performance of DIF detection methods. Yet, it is difficult to identify a single method that works best for all commonly examined contexts.

Swaminathan & Rogers (1990) introduced the Logistic method for DIF detection and compared it to MH in simulations involving test lengths of 40, 60, and 80 items, sample sizes of 250 and 500 examinees, and DIF% of 20%. Using a significance level of $\alpha = .01$, they found that hit rates were very similar between MH and Logistic. False positive rates were better controlled with MH (around the nominal level) than with Logistic (ranging from 1% to 6%, depending on setting).

Schauberger & Tutz (2016) conducted Monte Carlo simulations to compare a new DIF detection method with MH and Logistic. They found that the Logistic method tended to yield better hit rates than MH by around .04-.05 in tests with 500 examinees per group and 20 items, using a significance level of $\alpha = .05$. However, they also found that this method had higher false alarm rates by around .01-.02. ROC curves showed very little difference between MH and Logistic in terms of area under the curve. Magis, Tuerlinckx and De Boeck (2015) introduced a DIF detection method based on Lasso regression, and found it to be superior to Logistic in terms of AUC in most settings, especially with small sample sizes (100 examinees per group) and when groups were of unequal ability. They also found that the ROC curves for MH and Logistic were virtually indistinguishable from each other in most settings.

2.2 Studies examining the effect of other factors on DIF detection

There is no shortage of simulation studies investigating the effects of various settings on DIF detection in general. Some settings show up in many simulation studies. The effect of each factor on DIF detection varies considerably across models, DIF detection methods, and other factors.

Lopez (2012) notes that "even so-called industry standards differ widely enough in terms of implementation that researchers are still discovering issues that may have substantial effects on power and Type I error rates." There are very few broadly applicable guidelines with respect to sample size or test length. Several studies suggest a minimum of 200 examinees per group when using the MH method for DIF analysis (for example Mazor, Clauser, and Hambleton, 1992; Broer, Lee, Rizavi, and Powers, 2005).

Guilera, Gómez-Benito, Hidalgo & Sánchez-Meca (2013) helpfully make a point of distinguishing between variables whose values would be known to the researcher in an empirical DIF analysis, and variables whose values would be unknown. In the present study, sample size and test length fall into the former category, while DIF% and DIF magnitude fall into the latter. Impact would typically be considered unknown in an observational study such as in a traditional DIF analysis where items are investigated for bias for or against a particular demographic group. Impact may be considered known and equal to 0 in a randomized experiment, however, such as when examinees are randomly assigned to treatment groups and DIF analysis is undertaken to assess intervention effects.

A review follows of the pertinent simulation studies in which the researchers manipulated all or some of the same factors as in the current study, and in which the Rasch model was utilized, and the DIF detection methods were MH, Logistic and/or Lasso. The relative

importance of factors affecting DIF detection is summarized first. Then each factor is addressed individually.

Relative importance of factors affecting DIF detection

Guilera, et. al. (2013) conducted a meta-analysis of 55 DIF detection studies that utilized Monte Carlo simulations to analyze the false alarm rate and/or hit rate of the MH method under various conditions. Most of the studies in their analysis involve more complex models than the Rasch, but many of the findings apply just as well to Rasch.

They noted that false alarm rates using the MH method were most affected by three primary factors: DIF percentage, application of purification procedures, and item difficulty parameters. Their evaluation of DIF percentage showed that false alarm rates increase when more items are DIF. The application of purification procedures notably reduces false alarm rate. Their research also showed that items of close to average difficulty (difficulty parameter close to 0) helps contain false alarm rates.

Additionally, Guilera, et. al. found that MH hit rates were most affected by: sample size (power is increased greatly when each group contains at least 500 examines); item difficulty parameters (values close to 0 result in improved hit rates); application of purification procedures (this process improves hit rates); and DIF percentage (hit rates decrease when more items are DIF).

Percentage of DIF items

A key factor manipulated in the current study is the proportion of items containing DIF (DIF%). To date, most of the simulation-based DIF research has involved a relatively small percentage of items containing DIF. This is because DIF analysis has mainly been used in large-scale assessments to detect item bias, with tests containing many items, not many of which are

expected to contain DIF. Usually between 10% and 15% of items in standardized tests contain DIF (Herrera & Gomez, 2008). In contrast, when the purpose of testing is to gauge item-level treatment effects – as is becoming more common – it is expected that a larger proportion of test items will contain DIF.

It has been shown that the percentage of items containing DIF has a considerable impact on the ability to correctly identify which items contain DIF. In the meta-analysis of Guilera, Gómez-Benito, Hidalgo & Sánchez-Meca (2013), the authors noted that as the proportion of DIF items increased, the false alarm rate generally increased and hit rate decreased under the MH method, especially when more than 20% of the items contained DIF. An explanation for this occurrence is that detection rates are significantly affected by matching criteria contamination. Fidalgo, Mellenbergh, and Muñiz (2000) and other researchers have noted that the greater the number of items with DIF, the less accurate the test score will be, with the likelihood of increased erroneous detection rates.

Schauberger and Tutz (2016) conducted Monte Carlo simulations to compare a new DIF detection method with MH, Logistic and Lasso. In tests with 500 examinees, 20 items, and no impact, they found that false alarm rates did not change much with any detection method when DIF% increased from 20% to 40%, but hit rates rose notably, typically by around .01. They used a significance level of .05.

Guilera et al. (2013) found across many studies that hit rate is notably reduced when more than 20% of items are DIF items. Other researchers have found that false alarm rates increased as DIF% increased (Meade & Wright, 2012; Stark, Chernyshenko & Drasgow, 2005; Wang & Yeh, 2003).

DIF magnitude

Numerous studies have found that larger DIF effects are easier to detect. Schauberger and Tutz (2016) found that false alarm rate under MH and Logistic did not change much as DIF magnitude increased, but hit rate improved greatly. Numerous other studies have found that false positive rates increased as DIF magnitude increased. (See for example Meade & Wright, 2012; Stark, Chernyshenko, & Drasgow, 2005; Wang & Yeh, 2003)

Impact

Group difference in ability, sometimes referred to as "impact", also has been shown to affect DIF detection, and is a common presence in real-world DIF analysis in observational studies. Impact can be defined as the mean ability level of the focal group minus the mean ability level of the reference group. DIF is easier to detect when groups are equal in ability, i.e. when impact = 0. When groups are unequal in ability, there is a confounding between DIF and differential ability. Schauberger and Tutz (2016) found that when impact was changed from 0 to 1, false alarm rates for MH increased by about .005-.007, at a variety of DIF magnitudes. False alarm rates for Logistic increased by .010-.013. Hit rates declined markedly for both methods, especially with strong or very strong DIF.

Numerous other studies have found a degradation of DIF detection when ability levels differ by group. (Jodoin & Gierl, 2001; DeMars, 2010; Chen, Chen & Shih, 2013). Other studies have noted that unequal ability distributions lead to increased false alarm rates but little change in hit rates. (Guilera, Gómez-Benito, Hidalgo & Sánchez-Meca, 2013).

Sample size

The number of examinees is another factor known to affect DIF detection performance. Hit rate tends to increase along with sample size, as might be expected; perhaps unexpectedly, so does false alarm rate, although not as much as hit rate (Guilera, Gómez-Benito, Hidalgo & Sánchez-Meca, 2013). Numerous other studies have found that hit rates and false alarm rates increase along with sample size (Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Chen, Chen & Shish, 2013; Van De Water, 2014).

Test length

It might be expected that longer tests yield more stable ability estimates and thus better item parameter estimates. However, Guilera, Gómez-Benito, Hidalgo & Sánchez-Meca (2013) found that across many studies, tests of moderate length (20 to 40 items) yield lower power and lower false alarm rates than shorter or longer tests. Yet, test length has less of an effect on DIF detection than other factors. This is because the accuracy of the matching criterion is not affected as much by test length as by other factors, especially DIF% (Narayanan & Swaminathan, 1994; Guilera et al., 2013). Rogers & Swaminathan (1993) found that the hit rate for the MH procedure was not sensitive to test length, but the hit rate for the logistic procedure was.

2.3 Literature review conclusion

The cumulative insight revealed in the literature of DIF research is a muddle of multiple methods with different performances in different settings. A great deal of the difficulty in DIF detection stems from the circularity/contamination issue: the DIF test of a given item may be affected by the presence of other DIF items. This research differs from prior work in that it offers

a new DIF detection method that might better deal with the circularity/contamination issue. The presented study applies an old "trick" – the randomization test – to the current problem.

Much of the extant literature is limited to relatively small DIF percentages. The present study investigates tests with 50% DIF. In previous research not much distinction is made between the DIF-as-bias and DIF-as-treatment-effect paradigms. The present study emphasizes the distinction in the context of observational and experimental studies.

CHAPTER 3

METHODS

To assess the randomization test's effectiveness in DIF detection, Monte Carlo simulations were implemented using R. The Randomization Test was compared to the Lasso, Mantel-Haenszel (MH) and Logistic Regression methods. Comparisons were made under a variety of conditions in which several factors were manipulated.

3.1 DIF detection methods

The DIF detection methods used in the present study will be described in greater detail in this section. An overview of randomization tests in general, including some theoretical foundations and their applicability to DIF analysis, will begin the discussion, followed by an exploration of the Mantel-Haenszel and Logistic Regression methods. The section concludes with a description of the Lasso method as used by Magis, et. al. (2015).

Randomization test

Randomization tests are a broad and widely applicable class of non-parametric statistical test. In a randomization test, the distribution of the test statistic under the null hypothesis is found by calculation of the test statistic under repeated random reassignment of group classification, rather than basing it on assumptions about the distribution of the variable of interest. The

reference distribution in a randomization test performs the same role as the sampling distribution in a traditional hypothesis test.

Randomization tests are useful when the distribution of the test statistic under the null hypothesis is unknown or difficult to calculate. Edgington & Onghena (2007) claim that randomization tests are the only "true" valid tests, and that historically parametric tests were used only because prior to the advent of high powered computing, determining a parametric sampling distribution was much less cumbersome than generating a permutation distribution.

The hypotheses in a randomization test are not quite the same as in traditional hypothesis tests. Randomization tests involve a so-called "strong hypothesis", which states that the treatment or group membership would not have had an effect on any of the individuals in the study. In the context of the present study, that means that while the null hypothesis in the permutation test DIF detection method ostensibly is that an item's difficulty is the same for the focal group as for the reference group, strictly speaking, the null hypothesis is that group membership does not affect any of the tested students' ability to answer correctly after controlling for ability.

The term *randomization test* is often used interchangeably with the term *permutation test*, although Edgington & Onghena (2007) distinguish between the two based on the distinction between observational studies and randomized experiments. They hold that the proper term in an experimental setting is *randomization test*, while the correct term in an observational setting is *permutation test*. In the experimental setting, the authors claim that randomization tests are appropriate and justified as long as there is random assignment of treatments to experimental units, regardless of whether the experimental units are a random sample from some population.

In the observational setting, however, permutation tests are not valid unless the observational units constitute a random sample from some population.

Their argument rests on the concept of exchangeability: Under the null, the joint distribution of the observations must be the same for each permutation of group membership. In the present study, group membership is shuffled among examinee response vectors, as opposed to individual item responses. Exchangeability is satisfied if the likelihood of a given response vector for an examinee (say, 0010011001 on a ten-item test) is the same no matter which group the examinee is in, assuming there is no DIF. This holds when the two examinee groups have the same ability distribution, but not when ability distribution differs. Therefore, when ability distribution differs between groups, the randomization test does not yield valid p-values. Scale linking largely attenuates the detrimental effects of this lack of exchangeability because it helps approximate the item parameter estimates one would have obtained had the groups been equal in ability.

From a DIF analysis perspective, we may think of DIF in both the experimental and the observational setting. The experimental setting applies when DIF analysis is undertaken to gauge the effects of some treatment and there is random assignment of examinees to treatment groups. A randomization test is justified in this situation as long as there is random assignment of examinees to instructional group. The observational setting applies when DIF analysis is undertaken for the traditional purpose of detecting unfairness or "item bias", such as based on gender or race or some other attribute that is innate rather than assigned. A randomization test is justified in this situation as long as the examinees in the study are random samples from populations.

In the present study, randomization tests were conducted via the following heuristic:

1. Generate responses.
2. Estimate item difficulties for each group separately.
3. Linking step (skipped in some instances): Adjust the item difficulty estimates for the focal group to have the same mean across all items as the reference group. This helps control for ability differences between groups. (In the present study this is done by adding a constant to each item, but preliminary analysis indicates that more advanced linking produces better results).
4. Take the observed DIF parameter estimate for each item to be its estimated difficulty for the reference group minus its estimated difficulty for the focal group. This is the observed test statistic.
5. Creation of the reference distribution
    a. Randomly reassign group indicators to examinees (and their associated response vectors)
    b. Estimate item difficulties for each randomized "group" separately.
    c. Link scales (skipped in some instances)
    d. Calculate difference in item difficulty estimates between groups.
    e. Repeat (a)-(d) 100 times.
6. The 2-sided p-value for a given item is the proportion of DIF statistics in the reference distribution (5) that are greater than or equal to the observed DIF statistic in (4) in absolute value.

It should be noted that an exact randomization test would involve every single possible permutation to form the reference distribution. However, this is often unfeasible in practice due to the large number of possible permutations; for instance, in the current study, settings with 30 examinees per group would require $\frac{60!}{30!\,30!} = 1.18 \times 10^{17}$ permutations. Generally, a suitably large sample from the population of possible permutations will yield quite accurate results. The extant literature recommends a minimum of 1000 permutations in order to obtain an accurate p-

value for a given hypothesis test. Because the current study's purpose was not to obtain a highly accurate p-value for a single hypothesis test, but rather to obtain reasonably accurate p-values over a composite of numerous hypothesis tests, 100 permutations were used in order to keep computational costs to a minimum. (Use of 1000 permutations at several settings yielded no discernible increase in precision but took ten times as long).

Mantel-Haenszel method

The Mantel-Haenszel (MH) procedure may best be understood as a group of 2×2 contingency tables. For a given item, respondents with total score s can be categorized via a two-way classification of respondent group and item success, as represented by the following table:

	Number of examinees with correct response	Number of examinees with incorrect response	Total
Reference Group	A_s	B_s	n_{Rs}
Focal Group	C_s	D_s	n_{Fs}
Total	n_{1s}	n_{0s}	n_s

Within each stratum s, the odds ratio (reference group to focal group) of a correct response to the item is $\frac{A_s D_s}{B_s C_s}$. The null hypothesis in the MH test is that the odds ratio equals one in all strata. The test statistic is

$$\boldsymbol{MH} = \frac{(\,|\sum_s A_s - \sum_s \mathrm{E}(A_s)| - 0.5\,)^2}{\sum_s \mathrm{Var}(A_s)}$$

where $\mathrm{E}(A_s) = \frac{n_{Rs}\, n_{1s}}{n_s}$ and $\mathrm{Var}(A_s) = \frac{n_{Rs}\ n_{Fs}\ n_{1s}\ n_{0s}}{{n_s}^2(n_s - 1)}$. This test statistic is asymptotically chi-squared with 1 degree of freedom.

Logistic Regression method

With the Logistic Regression method (Swaminathan & Rogers, 1990), assuming two examinee groups, the probability of a correct answer to a dichotomously scored item is modeled as

$$P(Y_{ij} = 1) = \frac{\exp(\beta_{0j}+\beta_1 S_i+\beta_{2j} G_i)}{1 + \exp(\beta_{0j}+\beta_1 S_i+\beta_{2j} G_i)}, \text{ where}$$

- $P(Y_{ij} = 1)$ is the probability of examinee i getting item j correct
- S_i is the total score for examinee i
- G_i is an indicator variable for the group membership of examinee i

$$G_i = \begin{cases} 1, \text{if examinee is in the focal group} \\ 0, \text{if examinee is in the reference group} \end{cases}$$

If $\beta_{2j} \neq 0$, the item contains uniform DIF.

Note that there is a single β_1 for the entire test; the effect of ability on response is presumed not to differ by item.

Lasso regression

Lasso regression (Tibshirani, 1996) is a popular variable selection technique. In Lasso regression, instead of finding regression coefficients that minimize the error sum of squares or maximize likelihood, the objective function includes a penalty applied to large absolute values of standardized regression coefficients. The result is that regression coefficients are shrunk towards 0, with the smallest shrunk all the way to 0.

In logistic regression, Lasso estimates are defined as the set of β_j's that minimize $-l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$, the negative log likelihood plus a penalty for the size of the standardized regression coefficients. Magis, Tuerlinckx and De Boeck (2015) developed a Lasso-based

approach to DIF detection. Their parameterization is equivalent to the Logistic Regression method described above, but all items are tested simultaneously.

In this model the likelihood is

$$L(Y|\beta) = \prod_{i,j} \left(P\left(Y_{ij} = 1\right)\right)^{Yij} \left(1 - P\left(Y_{ij} = 1\right)\right)^{1-Yij}$$

where $P\left(Y_{ij} = 1\right) = \frac{\exp\left(\beta_{0j}+\beta_1 S_i+\beta_{2j} G_i\right)}{1+\exp\left(\beta_{0j}+\beta_1 S_i+\beta_{2j} G_i\right)}$. All item parameters and group effects are seen as contributing to the overall model fit, but only DIF parameters (the β_{2j}'s) are penalized. That is, their method attempts to find the set of item parameter estimates that minimizes $-l(Y|\beta) + \lambda \sum_{j=1}^{p} |\beta_{2j}|$.

Pitt and Myung (2002) state concisely that "[b]ecause a particular λ-value balances fit (as indicated by the log likelihood) and complexity (indicated through the number of nonzero DIF parameters), the selection of an optimal λ parameter can be framed as a model selection problem that has desirable generalization properties." Magis, et. al. found the best method for selecting an optimal λ in their Lasso-based DIF detection was a weighted information criterion, which is a weighted average of Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978).

3.2 ROC curves

Comparisons of DIF detection methods in the present study are based on receiver operating characteristic (ROC) curves, which illustrate the performance of a binary classifier system as its discrimination threshold is varied. For each DIF detection method, hit rate (ability to correctly identify DIF items) was plotted against false alarm rate (propensity to incorrectly classify non-DIF items as DIF items) for each of a number of α levels or, in the case of Lasso,

penalty parameters. The area under the ROC curve (AUC) for a given method is used as a global measure of its performance. Since a perfect classification mechanism would entail 100% hit rate and 0% false alarm rate, the optimal ROC has area 1. An ROC area of 0.5 indicates a worthless test. AUC can be interpreted as the probability that the test result from a randomly chosen DIF item is more indicative of DIF than the test result from a randomly chosen non-DIF item.

For all DIF detection methods except Lasso, a hypothesis test was conducted for each item under the null hypothesis of no DIF. Each test resulted in a p-value. Over 100 replications, 100 p-values were obtained for each item. Thus, for the MH, Logistic, and Randomization detection methods, classification rules were based on significance levels, on a gradient from 0 to 1 in increments of .005. At each increment, a test item was classified as DIF if the p-value was less than or equal to the significance level, and classified as non-DIF otherwise. Because it is known which items are truly DIF and which are non-DIF, it could be determined which items were correctly classified, and hit rates and false positive rates computed at each increment.

Another classification rule had to be employed for Lasso because the Lasso method does not involve hypothesis testing. Following the framework of Magis et al. (2015), DIF classification was conducted at varying levels of the Lasso penalty parameter λ. The minimum λ for the classification rule was chosen to be zero, which yields the same maximum likelihood estimates that would be obtained under usual logistic regression. The maximum λ for the classification rule was chosen to be the smallest λ at which all DIF parameter estimates were zero (i.e. all items classified as non-DIF). Classification rules were evaluated by taking λ at each of one thousand evenly spaced increments between 0 and this maximum.

Under each setting and for each detection method, the hit rate and false alarm rate were obtained for each replication at each α or λ, then averaged across the 100 replications. Thus, one ROC curve was generated for each detection method under each setting.

Within each setting, DIF detection methods were compared by their respective areas under the curve (AUC). Differences between the Randomization Test and each other method were analyzed using a permutation test for paired data with the two-sided alternative.

CHAPTER 4

RESULTS

4.1 Simulations

The following factors were manipulated in the simulations:

Factor	Levels
Percentage of DIF items ("DIF%")	10%, 20%, 50%
Sample size (number of examinees in each group)	30, 100, 500
Test length (number of items)	10, 20, 40
Magnitude of DIF	0.4, 0.8
Impact	-1, 0, 1

Settings were chosen based on those commonly used in prior simulation studies. Settings for the percentage of DIF items were based partly on what has been used in prior research; the 50% level was added because it is expected that DIF analysis will increasingly be used to investigate item-level intervention effects, where the tests contain many items that are expected to differ by group.

Three different sample sizes were considered: 30, 100, and 500 examinees per group. Group sizes of 100 and 500 were chosen because of their relative ubiquity in prior studies, and can be seen as representing medium-small and medium group sizes. The group size of 30 was included too because of the importance of applying DIF analysis to smaller samples.

Three different test lengths were considered: 10, 20 and 40 items. Two levels of DIF magnitude, defined as the differential item difficulty, were considered as well: 0.4 and 0.8,

representing moderate and fairly substantial DIF respectively. These values have been used in many prior simulation studies. In the present study, an item was imbued with DIF by increasing its difficulty for the focal group.

Three settings were used for Impact (group difference in mean ability): -1, 0 and 1. These are common in similar DIF detection studies as well. Many studies have shown that the effectiveness of some DIF detection methods depends in part on this factor.

A full factorial design was used, with each level of each factor crossed with each other for a total of $3 \times 3 \times 3 \times 2 \times 3 = 162$ settings. Within each setting, 100 random response sets were generated, then analyzed itemwise for DIF with each of the four DIF detection methods.

Following the framework of most DIF simulation studies, the desired number of examinees were simulated, each designated as belonging to either the reference group or the focal group. Examinees were then imbued with random ability levels generated from the $N(-1,1)$, $N(0,1)$, or $N(1,1)$ distribution as called for. A pre-determined number of test items were created, with difficulty parameters generated randomly and independently from the $N(0,1)$ distribution. Since DIF in the Rasch model is fundamentally a difference in item parameters between groups, a desired item was coerced to contain DIF by changing its difficulty parameter for the focal group.

Responses to each item for each examinee were generated independently according to a Rasch model, using the R irtoys package (Partchev, 2015). The basic algorithm for generating a response for a given respondent to a given item is

- Determine the probability the respondent will answer the item correctly, P_{ij}, based on the Rasch model.
- Generate a random Bernoulli variable with probability parameter equal to P_{ij}.

- If the Bernoulli variable = 1, the respondent answered correctly. If the Bernoulli variable = 0, the respondent answered incorrectly.

Items were then analyzed for DIF based on the generated responses, using four different methods: the MH, Logistic Regression, Lasso, and Randomization Test. For the MH and Logistic methods, the R difR package (Magis, Beland, Tuerlinckx & De Boeck, 2010) was utilized for DIF analysis. For the Lasso method, DIF analysis was conducted using R code provided by Magis (personal communication), which calls the glmnet package (Friedman, Hastie, & Tibshirani, 2010) for Lasso penalization. For the Randomization Test method, item difficulty parameter estimates were obtained with irtoys, and the test was implemented in R using code created by the author for this study (Appendices D and E).

For the Randomization Test, after initial item difficulty estimates were obtained, the generated response set was retained but group membership was randomly permuted to examinees, and the group differences in item difficulty parameters were estimated for this permuted response set. Repeated random assignment of group membership and parameter re-estimation yielded a reference distribution for the test statistic, against which the original observed DIF statistic was compared.

In simulations with no impact, two sets of DIF estimates were obtained – one with linking and one without. The linking involved a simple location shift of the item difficulty estimates for the focal group to have the same mean as the item difficulty estimates for the reference group.

4.2 Results

The results are divided into three main sections: DIF analysis in an experimental setting, DIF analysis in an observational setting, and the effect of DIF% on DIF detection.

In an experimental setting, as when DIF analysis will be used to test for item-level effects of some treatment or intervention, group membership is often randomized to examinees rather than being an innate characteristic. In such a case, it may be reasonably assumed that the groups do not differ from each other with respect to average ability. As a result, IRT-based DIF detection methods – including the Randomization Test that is the focus of this study – can be employed without the linking step (Embretson & Reise, 2000). Additionally, exchangeability required for the randomization test is preserved in the experimental setting.

In contrast, in the observational setting, such as when examining items for bias against a particular demographic group, the assumption that the groups are equal in ability is often questionable. Therefore IRT-based DIF detection methods require the linking step in order to place the groups on the same scale. Furthermore, in the observational setting, exchangeability is not preserved, so a randomization test yields invalid p-values. Putting the groups on the same scale through linking helps to approximate valid p-values.

DIF detection methods are compared via average area under the curve (AUC), a global measure of performance of a binary classifier.

Main findings: Experimental setting

The Randomization Test without linking yields superior results across almost all settings in the absence of impact, i.e. when the groups have equal ability distributions. Table 1 shows the mean AUC for each method for each level of each main factor. It can be seen that the Randomization Test consistently outperforms the other methods.

Differences in mean AUC between the Randomization Test and each of the other identified methods were analyzed using a permutation test for paired data with the two-sided alternative. Significant differences are indicated by asterisks in Table 1 below. Blue text indicates instances where the Randomization Test was the significantly better method. The Randomization Test is clearly superior across all settings. Its superior performance becomes more pronounced as DIF% increases.

Table 1. Mean AUC by method, by factor level
(No impact, no linking in Randomization Test)

		Items			Examinees		
	Overall	10	20	40	30	100	500
Rand	.794	.790	.796	.795	.629	.786	.965
Lasso	.669 ***	.669 ***	.672 ***	.667 ***	.564 ***	.654 ***	.790 ***
MH	.664 ***	.666 ***	.667 ***	.658 ***	.551 ***	.650 ***	.789 ***
Logistic	.672 ***	.668 ***	.673 ***	.676 ***	.569 ***	.657 ***	.791 ***

	DIF magnitude		DIF %		
	0.4	0.8	10%	20%	50%
Rand	.723	.864	.789	.798	.794
Lasso	.622 ***	.716 ***	.775 **	.738 ***	.495 ***
MH	.618 ***	.709 ***	.761 ***	.730 ***	.501 ***
Logistic	.623 ***	.721 ***	.775 *	.740 ***	.502 ***

$\cdot\ .05 \leq p < .10$ $\quad * .01 \leq p < .05$ $\quad ** .001 \leq p < .01$ $\quad *** p < .001$

An example of the superiority of the Randomization Test, when groups were of equal ability and no linking was performed, is shown in Figure 1. This graphic shows the ROC curves for the "medium" setting (20-item test, 20% DIF items, 0.4 DIF magnitude, 100 respondents per group, equal abilities).

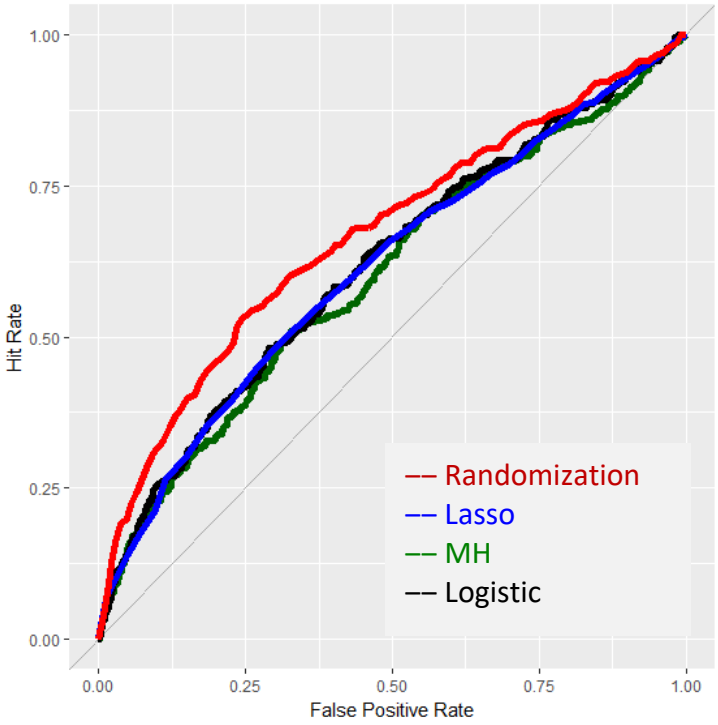

Figure 1. ROC curve by method (20-item test, 20% DIF items, 0.4 DIF magnitude, 100 respondents per group, no impact, no linking in randomization test)

The corresponding areas under the curve are given in Table 2.

Table 2. AUC by method (20-item test, 20% DIF items, 0.4 DIF magnitude, 100 respondents per group, no impact, no linking in Randomization Test)

Method	AUC
Randomization	.668
Lasso	.614
MH	.601
Logistic	.618

Results for two-way factor combinations are shown in Appendix B. A vast majority of results were statistically significant in favor of the Randomization Test. The few non-significant results occurred when DIF% was 10%.

Analysis of variance was conducted to determine which factors are most important to classification accuracy in the experimental setting. AUC was used as the response variable. All main effects were highly significant except Number of Items. AUCs increased along with sample size. While false alarm rates increased dramatically along with sample size, hit rates increased even more dramatically.

Six out of ten two-way interactions were found to be significant at the at the .001 level, as were three out of ten three-way interactions (See Appendix C for the ANOVA table). The two-way interactions that did not involve Number of Items were all significant at the .001 level. None of the two-way interactions involving Number of Items were significant.

The significant two-way interactions were mostly ordinal. The Randomization Test outperformed each other method at all levels of DIF%, with the effect becoming more pronounced as DIF% increased, especially when DIF% was 50%. The Randomization Test similarly outperformed each of the other methods at all levels of Examinees, with the effect becoming more pronounced as number of examinees increased. The Randomization Test outperformed each of the other methods at both levels (0.4, 0.8) of DIF Magnitude, with the effect more pronounced when DIF magnitude was 0.8. This was driven largely by the increased false alarm rates at 0.8 DIF magnitude for all methods except Randomization.

The DIF% × DIF Magnitude interaction is summarized by the finding that, while in general AUCs increased significantly when DIF magnitude increases from 0.4 to 0.8, the

increase was not as dramatic when DIF% was 50%. Similarly, while AUCs increased along with sample size, the increase was not as pronounced when DIF% is 50%.

The three-way interactions also were significant except those involving Items. Most of the significance is explained by the fact that having 50% DIF mutes the effects of the other factors for most methods but not for the Randomization Test.

It may be concluded that all methods except Randomization are severely impacted when DIF% reaches 50%. This indicates that DIF detection ability is highly dependent on the specifics of the test and examinees.

An additional ANOVA was performed to determine which factors affect the performance of only the Randomization Test. The results show that only DIF Magnitude, Number of Examinees, and their interaction were significant. The Randomization Test performed better as the number of examinees increased and DIF magnitude increased. The effect on AUC of increasing sample size from 100 to 500 was not as pronounced when DIF magnitude was 0.8; primarily this is due to very little opportunity for improvement at n=100 and DIF magnitude = 0.8, as the AUC was already .900.

Main findings: Observational setting

The Randomization Test with linking generally performed slightly worse than Lasso and Logistic, and slightly better than MH. Table 3 shows the mean AUC for each method for each level of each main factor. Note that impact is a factor in the observational setting, but not in the experimental setting.

Differences between the Randomization Test and each of the comparison methods were analyzed using a permutation test for paired data with the two-sided alternative. Significant differences are noted in Table 3. Instances where the Randomization Test was the significantly

better method are shown in blue. Instances where the Randomization Test had significantly poorer performance are shown in red.

Lasso tended to perform the best, while MH fared the worst. The Randomization Test was the worst-performing of the four methods when the mean ability level of the focal group was 0, and also when 50% of items contained DIF. However, none of the methods performed well when 50% of the items were DIF. AUCs in these settings were typically around .50, which indicates that the method is no better than a guess. Although many of the differences were statistically significant, they may not be large enough to be of practical import.

Lasso performed significantly better than the Randomization Test in almost all settings. Logistic performed significantly better than the Randomization Test in several settings as well. The Randomization Test performs significantly better than MH about half the time.

Table 3. Mean AUC by method, by factor level
(Linking in Randomization Test)

		Items			Examinees		
	Overall	10	20	40	30	100	500
Rand	.660	.661	.662	.656	.547	.651	.781
Lasso	.666 ***	.669 ***	.666 ·	.663 ·	.564 ***	.652	.783
MH	.652 **	.655	.655 ·	.645 *	.536 ·	.639 **	.780
Logistic	.662	.659	.664	.664 ·	.559 *	.646	.781

	DIF magnitude		DIF %			Impact		
	0.4	0.8	10%	20%	50%	-1	0	1
Rand	.618	.702	.761	.727	.491	.645	.664	.670
Lasso	.620 ·	.712 ***	.770 ***	.734 *	.494	.656 **	.669 ***	.673
MH	.607 ***	.697	.744 ***	.713 ***	.499 *	.641	.664	.650 ***
Logistic	.615	.710 *	.759	.727	.501 **	.653	.672 **	.662 *

· $.05 \leq p < .10$ * $.01 \leq p < .05$ ** $.001 \leq p < .01$ *** $p < .001$

As an example of the negligible differences in performance in many settings, Figure 2 shows the ROC curves for the setting with a 20-item test, 20% DIF items, 0.4 DIF magnitude, 100 respondents per group, impact = 1, and linking performed. Note that the different DIF detection methods are virtually indistinguishable. The differences in AUC were not statistically significant, based on twenty additional runs at this setting.

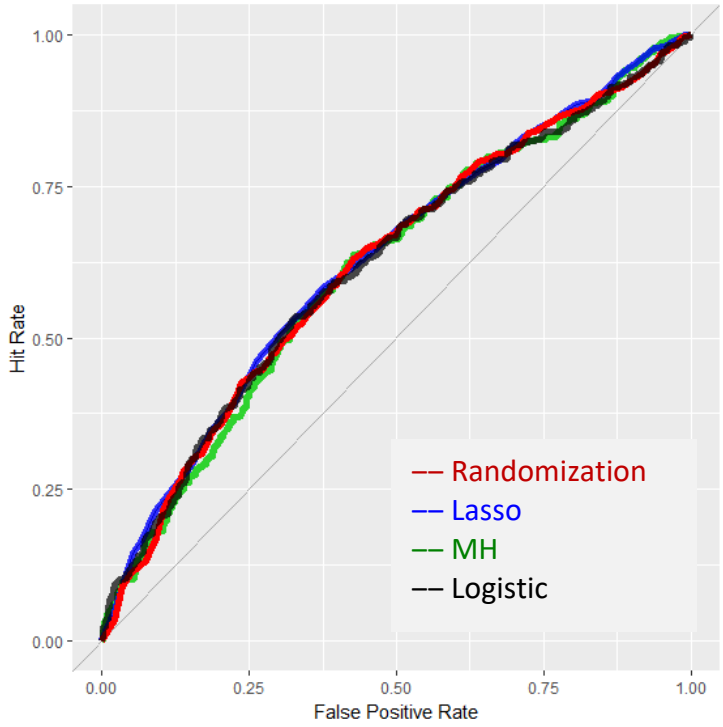

Figure 2. ROC curve by method (20-item test, 20% DIF items, 0.4 DIF magnitude, 100 respondents per group, impact = 1, linking in randomization test)

The corresponding areas under the curve are given in Table 4.

Table 4. AUC by method (20-item test, 20% DIF items, 0.4 DIF magnitude, 100 respondents per group, impact = 1, linking in Randomization Test)

Method	AUC
Randomization	.620
Lasso	.627
MH	.617
Logistic	.619

Results for two-way factor combinations can be found in Appendix B. Lasso was significantly better than Randomization in almost all setting combinations involving a short test (10 items) or a small sample size (30 examinees per group), although the difference in mean AUC was typically no larger than .020 or.025. Randomization was significantly better than MH in many settings, especially when impact = 1. Randomization was outperformed by all other methods when there were 500 examinees in each group and a DIF magnitude of 0.8, and also when 50% of items were DIF and impact was -1 or 0.

Analysis of variance was conducted to determine which factors are most important to classification accuracy in the observational setting. AUC was used as the response variable. All main factors and fourteen out of fifteen two-way interactions were found to be significant at the .001 level. Nine of the 20 three-way interactions were significant at the .001 level, and four others were significant at the .01 level. See Appendix C for the ANOVA table. The significance of many interaction terms is explained by the finding that when 50% of the items were DIF, that negated the effects of other factors. More specifically,

- All methods performed better with equal ability groups, except when DIF% was 50%.
- Performance for all methods improved greatly with increased sample size, except when DIF% was 50%.
- All methods performed much better when DIF magnitude was 0.8, except when DIF% was 50%.
- The impact of increasing DIF magnitude from 0.4 to 0.8 was most pronounced when there were 100 examinees.
- Randomization and Lasso were detrimentally affected more than the other methods by the leap to 50% DIF items.
- An increase from 30 to 100 examinees most helped Randomization and MH, but only on a 20- or 40-item test.

The effect of DIF percentage

The superiority of the Randomization Test without linking becomes more pronounced as the number of DIF items increases. One example can be seen in Figure 3, which shows the AUCs for each method at varying percentages of DIF items, in the setting with a 20-item test, a DIF magnitude of 0.4, 100 examinees in each group, and no difference in group ability. Additional simulations were performed with DIF percentages of 5, 15, 25, 30, 40, 60, and 70. The Randomization Test without linking does not suffer as more DIF items are added, while other methods (including the Randomization Test with linking) exhibit a marked degradation in ability to distinguish between DIF and non-DIF items.

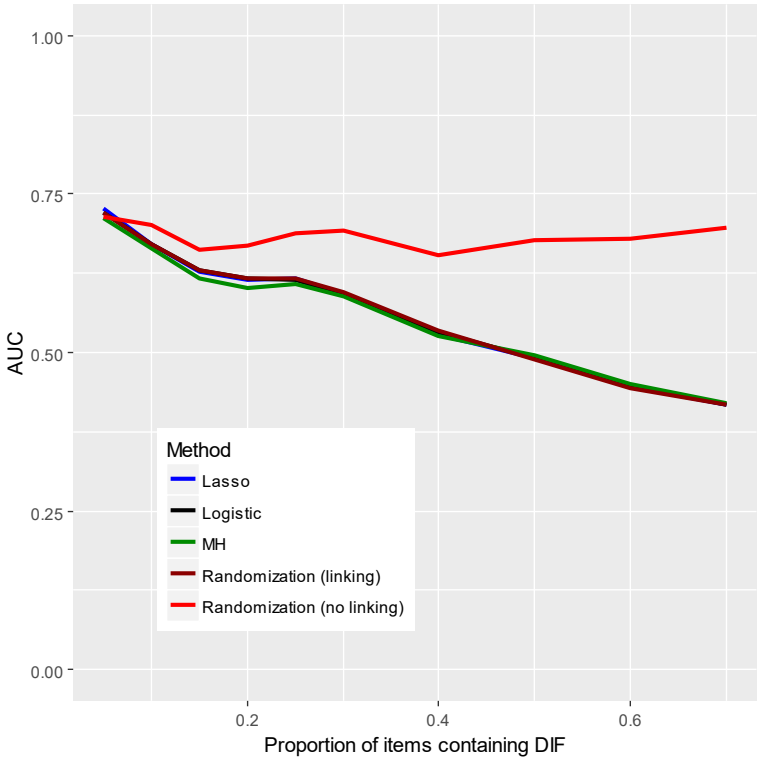

Figure 3. AUC by method, by DIF% (20-item test, 0.4 DIF magnitude, 100 respondents per group, no impact)

The same pattern was seen across many settings when there was no difference in ability. As DIF% increased, AUCs held steady for the Randomization Test without linking but decreased for the other methods, especially MH and Logistic with 500 examinees.

CHAPTER 5

DISCUSSION AND CONCLUSION

The Randomization Test performs considerably better than other methods if there is no difference in group mean ability and no linking is performed. The Randomization Test is recommended as a DIF detection method in the experimental setting as long as the Rasch model is appropriate for the test items in question.

Where equal ability cannot be assumed, and thus linking must be performed, the Randomization Test while less robust still performs on par with other methods. (This is the "observational setting"). The differences in mean AUC between the best- and worst-performing method are relatively small in any given setting – typically in the .01-.03 range. Lasso has a slight advantage over the other methods in ten-item tests. MH tends to underperform, while Randomization and Logistic perform better than MH but not as good as Lasso. The computational costs of using the Randomization Test make it not worthwhile in this case. Employment of more sophisticated linking procedures will likely improve the Randomization Test's performance.

In practice, in an observational setting one generally will not know whether respondent groups have equal ability distributions. Where DIF analysis is used to test for the effects of some treatment or intervention, however, there is often random assignment of individuals to treatment groups; in such a case, the assumption of equal ability distributions may be well-founded.

The findings of this study recommend the use of low significance levels when implementing the Randomization Test for DIF detection. Lower significance levels are observed to be associated with consistently smaller empirical false discovery rates in the current research.

5.1 Type I and Type II error

The outcome of interest throughout this study is area under the curve (AUC), which is a combination of false alarm (Type I error) rate and hit rate (power, or $1 -$ Type II error rate). It may prove informative to consider each of these components separately and in relation to each other.

In the observational setting, practitioners are usually examining items for potential bias. Items found to contain DIF are investigated more closely by content experts to determine the source of DIF. The item may be reworded or otherwise revamped in order to remove any biases. If the item is deemed beyond repair, the item may be removed from the test bank. A Type I error – declaring DIF where there is none – thus results in unnecessary time and labor costs associated with investigating the item and possibly developing a new item to replace it in the bank. On the other hand, a Type II error – failure to detect DIF – leads to the biased item being kept in the test bank, likely to the continuing detriment of some examinee group of interest. The latter may be considered the more serious error. Hence in the observational setting, practitioners may wish to guard against Type II error more than Type I error. That is, a high hit rate may carry more weight than a low false alarm rate.

In the experimental setting, practitioners are interested in whether some treatment or intervention has any effect on how difficult the items are for test takers. Type I and Type II errors have the usual connotation. A Type I error constitutes an item being declared unequally

difficult across treatment groups when it is in fact equally difficult. A Type II error constitutes an item being declared equally difficult across treatment groups when it is in fact unequally difficult. The relative costs of Type I and Type II errors depend largely on the purpose of the experiment.

The Randomization Test without linking controls the false alarm rate (Type I error) quite well compared to the other methods, particularly with large samples. With 500 examinees and no impact, false alarm rates for the Randomization Test remain close to the nominal level, while Lasso yields false alarm rates above .60 in many settings, and MH and Logistic yield rates above .25, especially when 50% of the items contain DIF.

The power (hit rate) for the Randomization Test without linking tends to be slightly better than MH but slightly worse than Logistic. The Randomization Test has much better Type I and Type II error rates than MH and Logistic with 50% DIF, 500 items or 0.8 DIF. The Randomization Test has much better Type I error than Lasso in those settings.

Lasso does not involve hypothesis testing and associated significance levels, therefore comparisons to the other methods are not entirely parallel. For Lasso, hit rates and false alarm rates were determined in each simulation based on which items were classified as DIF, as determined by the optimal penalty parameter using the Weighted Information Criteria. False positive rates are excessive when DIF% is 50%, especially with large samples and/or DIF magnitude – above .60 in many instances. When DIF% is 10%, false positive rates tend to be lower, particularly with smaller samples, smaller DIF magnitude, and longer tests—below .20 in many cases. Hit rates are much higher with large samples and/or large DIF magnitude – over .95 with 500 examinees and 0.8 DIF magnitude, compared to typically under .20 with 30 examinees and 0.4 magnitude. The ratio of hit rate to false alarm rate tends to be in the 1.5-3.7 range when

10% of items are DIF (especially high in the setting with 10% DIF, 0.4 magnitude, and 500 examinees). The ratio is in the 1.3-2.3 range when 20% of items are DIF, and in the 0.8-1.1 range when 50% of items are DIF.

5.2 Linking

When groups are equal in average ability, exchangeability as required by a randomization test is fulfilled, which results in the test performing strongly in the experimental setting. When groups are not equal in ability, exchangeability as required by a randomization test is not fulfilled, which results in the test performing poorly when linking is not undertaken. According to Good (2005), if the group distributions differ only in location, a mean transformation will preserve exchangeability: "Sometimes a simple transformation will ensure that observations are exchangeable. For example, if we know that X comes from a population with mean μ and distribution $F(x-\mu)$ and an independent observation Y comes from a population with mean v and distribution $F(x-v)$, then the independent variables $X' = X - \mu$ and $Y' = Y - v$ are exchangeable." Although linking does not involve transforming the response vectors directly, it does have the effect of treating the resulting estimates somewhat *as though* the responses had been so transformed, which likely ameliorates the lack of exchangeability.

On the other hand, when groups *are* equal in ability, linking adds random noise to the parameter estimation process, resulting in a diminution in DIF detection. As a result, when groups are equal in ability, linking during the Randomization Test negatively affects the outcomes. This effect is most pronounced when a large percentage of items contain DIF.

In the present study, a simple type of linking was utilized in the Randomization Test: the difficulty parameters for the focal group were shifted to have the same mean as the reference

group. More sophisticated linking procedures may help preserve exchangeability, thereby improving the Randomization Test's performance in the observational setting.

5.3 Future directions

This research, and the accumulation of inquiry in the field, offers numerous potential avenues for further exploration. While not exhaustive the following endeavors are worthy of consideration.

Methods and models

A possible path is the comparison of the Randomization Test to additional DIF detection methods. Particularly interesting would be a comparison with other IRT-based methods such as the likelihood ratio test (Thissen, Steinberg & Gerrard, 1986) and Lord's χ^2 (Lord, 1980).

Another exciting possible extension of this current research would be an analysis of the performance of the Randomization Test under more complex IRT models, such as the 2- and 3-parameter logistic models, which include item discrimination and pseudo-guessing parameters. These models can also incorporate nonuniform DIF, where an item's discrimination parameter varies by group.

This study has also encouraged an interest in the development of a Randomization Test for use in diagnostic classification models, in which the person parameters are presence or absence of attributes.

Purification

Yet another potential extension of this work is to incorporate scale purification (Lord, 1980; Candell & Drasgow, 1988; Segall, 1983) into the Randomization Test, to determine the extent to which it improves performance. Purification is necessitated by the circularity issues

discussed in Section 1.4, and is employed to reduce the likelihood that DIF analysis on a given item will be contaminated by other DIF items. As has been noted by researchers before (Guilera, Gómez-Benito, Hidalgo & Sánchez-Meca, 2013) using purification procedures can be a vital moderator that greatly aids in DIF detection, reducing false alarm rates and enhancing power.

Purification procedures follow this general process:

1. Calculate a provisional DIF statistic for each item, using all other items as an anchor.
2. Remove from the anchor set those items showing large DIF statistics.
3. Re-estimate IRT parameters – or in the case of non-IRT DIF analysis, recalculate total scores – based on the reduced anchor set, and test each item again for DIF.
4. Optional: Repeat steps 2 and 3 until there is no change in which items show apparent DIF.

Purification has been seen to ameliorate the complications brought about by the circularity issue, especially when the proportion of DIF items is relatively low (Candell & Drasgow, 1988).

Item purification in the MH and Logistic methods (using difR's built-in purification option) typically results in an AUC increase of between .010-.030 when 10% or 20% of items are DIF, based on additional runs in the present study. Purification procedures within the Randomization Test show modest potential: preliminary attempts across several settings are seen to boost AUC by around .005. Use of more sophisticated linking methods may result in still further improvement with respect to purification in the observational setting.

More than two groups

The Randomization Test for DIF detection, as with most DIF detection methods, can be extended to test for DIF among three or more groups. Such extensions already exist as the Generalized MH (Somes, 1986), Generalized Logistic (Magis et al., 2011), Group Lasso (Tutz and Schauberger, 2015), Generalized Lord's chi-square (Kim, Cohen & Park, 1995), and

Generalized Breslow-Day (Fujii & Yanagawa, 1990) DIF procedures. An appropriate test statistic would need to be developed as a measure of difference among three or more groups.

5.4 Conclusion

The Randomization Test provides superior DIF detection when groups can be assumed equal in ability and the Rasch model is appropriate. It is hoped the Randomization Test can provide practitioners an improved method by which to classify test items as DIF or non-DIF. This will improve the ability to assess whether an intervention has had an effect on item performance. It will improve test fairness when groups are equal in ability.

The results of the current research reveal some similar findings as prior work. In this study, as in the others noted, DIF detection is better when the DIF amount is larger (Schauberger & Tutz, 2016; Meade & Wright, 2012; Stark, Chernyshenko, & Drasgow, 2005; Wang & Yeh, 2003), when groups are of equal ability (Schauberger & Tutz, 2016; Jodoin & Gierl, 2001; DeMars, 2010; Chen, Chen & Shih, 2013; Guilera, Gómez-Benito, Hidalgo & Sánchez-Meca, 2013), or when there are more examinees (Guilera, Gómez-Benito, Hidalgo & Sánchez-Meca, 2013; Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Chen, Chen & Shish, 2013; Van De Water, 2014); and, test length does not make a substantial difference (Narayanan & Swaminathan, 1994; Guilera et al., 2013; Rogers & Swaminathan,1993).

The current research likewise echoes the commonly cited (Fidalgo, Mellenbergh & Muñiz, 2000; Schauberger & Tutz, 2016; Meade & Wright, 2012; Stark, Chernyshenko & Drasgow, 2005; Wang & Yeh, 2003; Guilera et al., 2013) diminishing effectiveness of DIF detection procedures as DIF% increases. The notable exception is the Randomization Test without scale linking when groups are of equal ability, which does not exhibit diminished

effectiveness as DIF% increases. Much of the extant literature is limited to relatively small DIF percentages. The current research investigates tests with 50% DIF.

In addition, in previous research not much distinction is made between the DIF-as-bias and DIF-as-treatment-effect paradigms. The traditional view of DIF-as-bias falls squarely into the realm of observational studies where group membership (gender, race) is innate. DIF-as-treatment-effect often falls under the purview of randomized experiments where group membership is randomly assigned (traditional lecture vs. flipped classroom; or old textbook vs. new textbook).

As the current study shows the Randomization test is much better than the other studied methods in cases where groups are of equal ability and no scale linking is performed. This corresponds to a situation where the researcher is interested in testing whether differentiated instruction has any effect on probability of getting items correct, and students are randomly assigned to an instruction group. Such a scenario is an excellent example of an experimental setting. In such a setting no scale linking is employed because the groups may be assumed equal in ability. The importance of this distinction between the experimental versus observational became profoundly more evident as the data derived from this study developed.

The ubiquity of differentiated instruction and standardized testing in all realms of modern life is without question. It is unreasonable to imagine, as populations grow and competition for education and employment increase, that this will lessen. The search for more efficient but truly accurate tests and measurement is intense. The research presented in this study offers a path for improving the validity of such undertakings.

APPENDIX A

Area under the curve for each setting

The following tables list the area under the ROC curve for each method, for each factor combination.

Rand-L: Randomization test with linking
Rand-NL: Randomization test without linking

Table A1. AUCs (10-item test, 30 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.542	.565	.478	.640	.593	.481
	Lasso	.551	.566	.472	.689	.626	.476
	MH	.505	.499	.503	.603	.575	.503
	Logistic	.517	.515	.498	.635	.605	.506
0	**Rand-L**	.544	.552	.478	.638	.626	.485
	Rand-NL	.558	.569	.544	.657	.722	.705
	Lasso	.567	.534	.466	.633	.653	.497
	MH	.531	.539	.496	.678	.641	.503
	Logistic	.549	.536	.488	.702	.658	.505
1	**Rand-L**	.532	.548	.491	.590	.671	.510
	Lasso	.551	.560	.491	.628	.691	.506
	MH	.531	.528	.481	.660	.613	.485
	Logistic	.553	.534	.489	.678	.636	.490

Table A2. AUCs (10-item test, 100 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.581	.626	.485	.846	.802	.480
	Lasso	.584	.629	.493	.854	.805	.474
	MH	.612	.606	.498	.810	.748	.502
	Logistic	.596	.612	.492	.814	.749	.504
0	**Rand-L**	.609	.615	.482	.879	.795	.490
	Rand-NL	.636	.680	.673	.902	.885	.901
	Lasso	.616	.619	.483	.901	.793	.497
	MH	.636	.620	.506	.830	.795	.501
	Logistic	.644	.606	.498	.838	.791	.500
1	**Rand-L**	.696	.587	.482	.898	.819	.488
	Lasso	.702	.574	.492	.913	.822	.493
	MH	.620	.573	.492	.861	.795	.494
	Logistic	.623	.572	.500	.867	.804	.492

Table A3. AUCs (10-item test, 500 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.892	.847	.519	.980	.940	.507
	Lasso	.907	.824	.507	.994	.941	.506
	MH	.821	.834	.503	.990	.967	.522
	Logistic	.821	.829	.499	.989	.967	.502
0	**Rand-L**	.922	.846	.487	.986	.954	.502
	Rand-NL	.931	.923	.940	.998	.998	.996
	Lasso	.947	.856	.501	.998	.973	.508
	MH	.866	.848	.505	.998	.985	.508
	Logistic	.871	.847	.507	.998	.985	.506
1	**Rand-L**	.898	.863	.502	.988	.961	.504
	Lasso	.922	.870	.501	.999	.975	.497
	MH	.839	.852	.502	.997	.980	.491
	Logistic	.844	.854	.508	.997	.980	.496

Table A4. AUCs (20-item test, 30 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.504	.522	.484	.576	.534	.445
	Lasso	.546	.532	.505	.621	.604	.500
	MH	.514	.502	.498	.614	.555	.490
	Logistic	.540	.548	.502	.646	.601	.504
0	**Rand-L**	.515	.529	.493	.666	.663	.481
	Rand-NL	.538	.581	.545	.692	.736	.698
	Lasso	.529	.546	.494	.684	.671	.486
	MH	.547	.537	.495	.657	.598	.492
	Logistic	.551	.537	.497	.657	.641	.485
1	**Rand-L**	.536	.556	.493	.654	.641	.535
	Lasso	.547	.568	.481	.685	.650	.500
	MH	.504	.526	.488	.641	.563	.480
	Logistic	.523	.533	.505	.664	.602	.489

Table A5. AUCs (20-item test, 100 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.642	.625	.494	.868	.780	.500
	Lasso	.630	.617	.488	.845	.773	.498
	MH	.615	.592	.513	.807	.732	.518
	Logistic	.633	.604	.509	.822	.742	.520
0	**Rand-L**	.672	.617	.489	.885	.810	.499
	Rand-NL	.701	.668	.678	.910	.898	.892
	Lasso	.670	.614	.492	.886	.810	.490
	MH	.664	.601	.496	.874	.804	.491
	Logistic	.671	.618	.492	.888	.811	.498
1	**Rand-L**	.671	.620	.486	.872	.807	.500
	Lasso	.670	.627	.489	.875	.812	.502
	MH	.620	.617	.507	.829	.791	.510
	Logistic	.629	.619	.510	.842	.804	.509

Table A6. AUCs (20-item test, 500 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.891	.842	.496	.985	.948	.495
	Lasso	.876	.816	.501	.978	.937	.490
	MH	.876	.808	.506	.991	.959	.521
	Logistic	.880	.808	.500	.992	.959	.507
0	**Rand-L**	.914	.852	.504	.987	.964	.500
	Rand-NL	.940	.935	.935	.995	.995	.995
	Lasso	.914	.851	.503	.990	.972	.501
	MH	.906	.853	.500	.996	.981	.515
	Logistic	.909	.857	.504	.996	.981	.517
1	**Rand-L**	.902	.844	.495	.986	.954	.499
	Lasso	.903	.842	.498	.995	.971	.496
	MH	.898	.836	.494	.996	.978	.483
	Logistic	.897	.836	.495	.996	.979	.485

Table A7. AUCs (40-item test, 30 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.513	.493	.443	.591	.481	.370
	Lasso	.564	.538	.479	.677	.608	.484
	MH	.507	.486	.476	.565	.544	.473
	Logistic	.549	.530	.506	.608	.613	.504
0	**Rand-L**	.546	.554	.501	.630	.611	.468
	Rand-NL	.574	.574	.571	.678	.685	.699
	Lasso	.566	.549	.508	.654	.621	.495
	MH	.527	.524	.495	.600	.573	.490
	Logistic	.559	.544	.502	.684	.651	.500
1	**Rand-L**	.596	.554	.510	.672	.667	.528
	Lasso	.564	.556	.496	.659	.653	.492
	MH	.506	.500	.482	.567	.568	.482
	Logistic	.546	.549	.497	.642	.616	.488

Table A8. AUCs (40-item test, 100 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.644	.621	.482	.834	.760	.468
	Lasso	.640	.620	.487	.821	.748	.486
	MH	.590	.591	.507	.794	.738	.504
	Logistic	.607	.607	.505	.833	.767	.510
0	**Rand-L**	.656	.605	.492	.867	.805	.497
	Rand-NL	.667	.673	.677	.907	.903	.904
	Lasso	.652	.601	.492	.863	.799	.496
	MH	.626	.617	.500	.857	.788	.501
	Logistic	.646	.611	.506	.884	.815	.501
1	**Rand-L**	.646	.606	.509	.867	.798	.516
	Lasso	.644	.607	.514	.871	.815	.502
	MH	.643	.591	.499	.814	.763	.499
	Logistic	.655	.599	.511	.842	.788	.493

Table A9. AUCs (40-item test, 500 examinees)

		DIF magnitude = 0.4			DIF magnitude = 0.8		
Impact		10% DIF	20%	50%	10% DIF	20%	50%
-1	**Rand-L**	.899	.839	.499	.985	.952	.498
	Lasso	.881	.818	.499	.979	.949	.500
	MH	.860	.799	.506	.987	.962	.521
	Logistic	.869	.805	.507	.989	.965	.516
0	**Rand-L**	.896	.851	.500	.987	.957	.498
	Rand-NL	.931	.942	.935	.995	.996	.997
	Lasso	.890	.847	.498	.989	.974	.503
	MH	.906	.852	.505	.994	.978	.512
	Logistic	.911	.858	.507	.994	.980	.514
1	**Rand-L**	.878	.841	.493	.986	.955	.505
	Lasso	.875	.836	.492	.993	.978	.498
	MH	.877	.817	.500	.995	.977	.485
	Logistic	.886	.822	.496	.996	.978	.487

APPENDIX B

Average AUC by method, by two-way combination of factors

Randomization test compared to each other method by way of permutation test for paired data.

Statistically significant differences marked.

- Blue indicates significant difference in favor of Randomization Test
- Red indicates significant difference in favor of other test

$\cdot$ $.05 \leq p < .10$ $\quad * \; .01 \leq p < .05$ $\quad ** \; .001 \leq p < .01$ $\quad *** \; p < .001$

Table B1. Average AUC by method, by two-way combination of factor levels: Experimental setting (No impact; no linking in Randomization Test)

Items	10			20			40		
Examinees	30	100	500	30	100	500	30	100	500
Rand	.626	.780	.964	.632	.791	.966	.630	.788	.966
Lasso	.558 *	.652 *	.797 *	.568 *	.660 *	.789 *	.566 *	.651 *	.784 *
MH	.565 *	.648 *	.785 *	.554 ·	.655 *	.792 ·	.535 *	.648 *	.791 *
Logistic	.573 ·	.646 *	.786 *	.561 ·	.663 *	.794 ·	.573 ·	.661 *	.794 *

Minimum possible p-value is .031

Items	10		20		40	
DIF magnitude	0.4	0.8	0.4	0.8	0.4	0.8
Rand	.717	.863	.725	.868	.727	.863
Lasso	.621 *	.717 **	.624 **	.721 **	.623 **	.710 **
MH	.616 **	.715 *	.622 **	.712 **	.617 **	.699 **
Logistic	.616 *	.720 *	.626 **	.719 **	.627 **	.725 *

Minimum possible p-value is .004

Items	10			20			40		
DIF %	10%	20%	50%	10%	20%	50%	10%	20%	50%
Rand	.780	.796	.793	.796	.802	.791	.792	.796	.797
Lasso	.777	.738 *	.492 *	.779 *	.744 *	.494 *	.769 *	.732 *	.499 *
MH	.757	.738 *	.503 *	.774	.729 *	.498 *	.752 *	.722 *	.501 *
Logistic	.767	.737 *	.501 *	.779	.741 *	.499 *	.780 *	.743 *	.505 *

Minimum possible p-value is .031

Examinees	30		100		500	
DIF magnitude	0.4	0.8	0.4	0.8	0.4	0.8
Rand	.562	.697	.673	.900	.935	.996
Lasso	.529 *	.599 **	.582 **	.726 **	.756 **	.823 **
MH	.521 **	.581 **	.585 **	.716 **	.749 **	.830 *
Logistic	.529 *	.609 *	.588 **	.725 *	.752 **	.830 **

Minimum possible p-value is .004

Examinees	30			100			500		
DIF %	10%	20%	50%	10%	20%	50%	10%	20%	50%
Rand	.616	.645	.627	.787	.785	.787	.965	.965	.966
Lasso	.606	.596 *	.491 *	.765 *	.706 *	.492 *	.955	.912 *	.502 *
MH	.590	.569 *	.495 *	.748 *	.704 *	.499 *	.944 •	.916 *	.508 *
Logistic	.617	.595 *	.496 *	.762 *	.709 *	.499 *	.947	.918 *	.509 *

Minimum possible p-value is .031

DIF magnitude	0.4			0.8		
DIF %	10%	20%	50%	10%	20%	50%
Rand	.720	.727	.722	.859	.869	.865
Lasso	.706 •	.669 **	.493 **	.844 **	.807 **	.497 **
MH	.690 *	.666 **	.500 **	.832 *	.794 **	.501 **
Logistic	.701 *	.668 **	.500 **	.849	.813 **	.503 **

Minimum possible p-value is .004

• $.05 \leq p < .10$ $\quad *\ .01 \leq p < .05$ $\quad **\ .001 \leq p < .01$ $\quad ***\ p < .001$

Table B2. Average AUC by method, by two-way combination of factor levels: Observational setting (Linking performed in Randomization Test)

Items	10			20			40		
Examinees	30	100	500	30	100	500	30	100	500
Rand	.554	.648	.783	.546	.658	.781	.540	.648	.779
Lasso	.564 *	.652 *	.790 *	.564 **	.655	.780	.565 *	.648	.778
MH	.549	.639	.778	.539	.643 *	.783	.520	.635 *	.780
Logistic	.561	.639	.778	.557	.651	.783	.561	.649	.782

Items	10		20		40	
DIF magnitude	0.4	0.8	0.4	0.8	0.4	0.8
Rand	.617	.706	.618	.705	.617	.695
Lasso	.622 ·	.716 ***	.620	.712	.619	.708 ·
MH	.605 ·	.705	.612 ·	.699	.603 *	.686
Logistic	.607 ·	.711	.619	.709	.618	.710

Items	10			20			40		
DIF%	10%	20%	50%	10%	20%	50%	10%	20%	50%
Rand	.759	.734	.492	.763	.728	.494	.761	.719	.488
Lasso	.775 ***	.739	.492	.769	.734	.495	.766	.729	.496
MH	.744	.722 ·	.500 *	.753	.713 *	.500	.734 **	.704 ·	.496
Logistic	.752	.727	.499 *	.763	.727	.501	.761	.728	.503 ·

Items	10			20			40		
Impact	-1	0	1	-1	0	1	-1	0	1
Rand	.656	.660	.668	.646	.669	.670	.632	.662	.674
Lasso	.661	.669 *	.677 **	.653	.672 *	.673	.654 *	.666	.669
MH	.644	.666	.655	.645	.667	.653 *	.634	.658	.643 **
Logistic	.647	.668	.662	.656	.673	.662	.655 *	.676 ***	.661 *

· $.05 \leq p < .10$ * $.01 \leq p < .05$ ** $.001 \leq p < .01$ *** $p < .001$

Examinees	30		100		500	
DIF magnitude	0.4	0.8	0.4	0.8	0.4	0.8
Rand	.521	.572	.583	.720	.749	.814
Lasso	.531 *	.598 **	.583	.720	.747	.818 *
MH	.508 *	.563	.576	.702 **	.736 **	.825 ***
Logistic	.526	.593 *	.581	.712	.738 *	.824 ***

Examinees	30			100			500		
DIF %	10%	20%	50%	10%	20%	50%	10%	20%	50%
Rand	.583	.575	.482	.757	.706	.491	.942	.901	.500
Lasso	.606 **	.596 **	.490	.758	.705	.493	.946	.902	.500
MH	.570	.548 *	.490	.728 ***	.687 ***	.502 **	.933	.904	.504
Logistic	.600 ·	.581	.497	.741 *	.696 **	.503 **	.935	.905	.503

Examinees	30			100			500		
Impact	-1	0	1	-1	0	1	-1	0	1
Rand	.514	.554	.571	.641	.654	.659	.779	.784	.781
Lasso	.558 ***	.564 **	.571	.638	.654	.662	.772 *	.790 **	.786 ·
MH	.523	.551	.534 **	.627 ·	.650	.640 **	.774	.789	.778
Logistic	.551 **	.569 *	.557 ·	.635	.657	.648 *	.772	.791 ·	.779

DIF magnitude	0.4			0.8		
DIF %	10%	20%	50%	10%	20%	50%
Rand	.694	.667	.491	.828	.787	.491
Lasso	.700 ·	.667	.493	.840 **	.801 **	.495
MH	.672 **	.650 ***	.498 *	.815 ·	.776	.499
Logistic	.685	.659 *	.501 ***	.833	.795	.501

· $.05 \leq p < .10$ * $.01 \leq p < .05$ ** $.001 \leq p < .01$ *** $p < .001$

DIF magnitude	0.4			0.8		
Impact	-1	0	1	-1	0	1
Rand	.610	.619	.624	.679	.709	.717
Lasso	.614	.622	.625	.699 •	.716 ***	.721
MH	.597 •	.618	.605 ***	.685	.709	.696 **
Logistic	.607	.623	.614 *	.699 •	.722 *	.709

DIF %	10%			20%			50%		
Impact	-1	0	1	-1	0	1	-1	0	1
Rand	.745	.767	.770	.709	.734	.739	.479	.491	.503
Lasso	.758 •	.775 **	.778 •	.719	.738	.745 *	.491 •	.495 •	.497 •
MH	.726 *	.761	.744 *	.694 •	.730	.715 **	.504 ***	.501 **	.492 *
Logistic	.741	.775	.760	.713	.740	.728 *	.505 ***	.501 ***	.497

• $.05 \leq p < .10$ * $.01 \leq p < .05$ ** $.001 \leq p < .01$ *** $p < .001$

APPENDIX C

Analysis of Variance

Table C1. Analysis of Variance: Experimental setting

	Source	df	SS	MS	F	p-value
***	DIF detection method (Method)	3	.6802	.2267	536.3	< .001
***	DIF percentage (Pct)	2	1.7069	.8534	2018.4	< .001
***	Examinees per group (N)	2	2.3622	1.1811	2793.3	< .001
***	DIF magnitude (Mag)	1	.5831	.5831	1379.1	< .001
	Test Length (L)	2	.0007	.0003	0.8	.46
***	Method · Pct	6	.6535	.1089	257.6	< .001
***	Method · N	6	.0846	.0141	33.3	< .001
***	Method · Mag	3	.0158	.0053	12.5	< .001
	Method · L	6	.0039	.0006	1.5	.21
***	Pct · N	4	.4862	.1215	287.4	< .001
***	Pct · Mag	2	.1411	.0705	166.8	< .001
•	Pct · L	4	.0044	.0011	2.6	.06
***	N · Mag	2	.0686	.0343	81.1	< .001
	N · L	4	.0025	.0006	1.5	.24
•	Mag · L	2	.0026	.0013	3.1	.07
	Method · Pct · N	12	.1648	.0137	32.5	.46
***	Method · Pct · Mag	6	.0301	.0050	11.9	< .001
	Method · Pct · L	12	.0036	.0003	0.7	.73
***	Method · N · Mag	6	.0154	.0026	6.1	< .001
	Method · N · L	12	.0063	.0005	1.2	.32
	Method · Mag · L	6	.0020	.0003	0.8	.58
***	Pct · N · Mag	4	.0302	.0075	17.9	< .001
	Pct · N · L	8	.0043	.0005	1.3	.30
	Pct · Mag · L	4	.0007	.0002	0.4	.80
	N · Mag · L	4	.0011	.0003	0.7	.63
	Method · Pct · N · Mag	12	.0030	.0002	0.6	.83
	Method · Pct · N · L	24	.0074	.0003	0.7	.78
	Method · Pct · Mag · L	12	.0039	.0003	0.8	.68
	Method · Mag · N · L	12	.0057	.0005	1.1	.39
	Pct · N · Mag · L	8	.0048	.0006	1.4	.24
	Residuals	24	.0101	.0004		

Table C2. Analysis of Variance: Observational setting

	Source	df	SS	MS	F	p-value
***	DIF detection method (Method)	3	.0181	.0060	49.3	< .001
***	DIF percentage (Pct)	2	8.7960	4.3981	35949.9	< .001
***	Examinees per group (N)	2	5.7521	2.8761	23508.7	< .001
***	DIF magnitude (Mag)	1	1.3171	1.3171	10765.6	< .001
***	Test length (L)	2	.0029	.0014	11.8	< .001
***	Impact (Imp)	2	.0423	.0212	173.0	< .001
***	Method · Pct	6	.0174	.0029	23.7	< .001
***	Method · N	6	.0143	.0024	19.4	< .001
***	Method · Mag	3	.0025	.0008	6.8	< .001
***	Method · L	6	.0038	.0006	5.1	< .001
***	Method · Imp	6	.0098	.0016	13.4	< .001
***	Pct · N	4	2.6100	.6525	5333.5	< .001
***	Pct · Mag	2	.6553	.3277	2678.3	< .001
***	Pct · L	4	.0029	.0007	6.0	< .001
***	Pct · Imp	4	.0161	.0040	32.8	< .001
***	N · Mag	2	.1553	.0777	634.7	< .001
***	N · L	4	.0037	.0009	7.5	< .001
***	N · Imp	4	.0046	.0011	9.3	< .001
***	Mag · L	2	.0036	.0018	14.9	< .001
***	Mag · Imp	2	.0037	.0018	15.1	< .001
*	L · Imp	4	.0015	.0004	3.1	.02
***	Method · Pct · N	12	.0085	.0007	5.8	< .001
·	Method · Pct · Mag	6	.0015	.0002	2.0	.08
	Method · Pct · L	12	.0025	.0002	1.7	.10
**	Method · Pct · Imp	12	.0047	.0004	3.2	< .01
***	Method · N · Mag	6	.0070	.0012	9.5	< .001
**	Method · N · L	12	.0050	.0004	3.4	.001
***	Method · N · Imp	12	.0130	.0011	8.9	< .001
***	Method · Mag · L	6	.0011	.0002	1.4	< .001
*	Method · Mag · Imp	6	.0023	.0004	3.1	.01
**	Method · L · Imp	12	.0051	.0004	3.4	.001
***	Pct · N · Mag	4	.0916	.0229	187.2	< .001
·	Pct · N · I	8	.0019	.0002	1.9	.08
***	Pct · N · Imp	8	.0066	.0008	6.7	< .001
·	Pct · Mag · L	4	.0011	.0003	2.2	.08
***	Pct · Mag · Imp	4	.0048	.0012	9.8	< .001
**	Pct · L · Imp	8	.0030	.0004	3.1	.007
***	N · Mag · L	4	.0035	.0009	7.1	< .001
***	N · Mag · Imp	4	.0063	.0016	12.9	< .001
*	N · L · Imp	8	.0026	.0003	2.7	.02
	Mag · L · Imp	4	.0003	.0001	0.7	.61

	Source	df	SS	MS	F	p-value
·	Method · Pct · N · Mag	12	.0028	.0002	1.9	.06
**	Method · Pct · N · L	24	.0078	.0003	2.7	< .01
*	Method · Pct · N · Imp	24	.0055	.0002	1.9	.03
	Method · Pct · Mag · L	12	.0018	.0001	1.2	.31
·	Method · Pct · Mag · Imp	12	.0028	.0002	1.9	.06
	Method · Pct · L · Imp	24	.0026	.0001	0.9	.61
·	Method · N · Mag · L	12	.0027	.0002	1.9	.07
·	Method · N · Mag · Imp	12	.0025	.0002	1.7	.09
***	Method · N · L · Imp	24	.0106	.0004	3.6	< .001
	Method · Mag · L · Imp	12	.0019	.0002	1.3	.24
**	Pct · N · Mag · L	8	.0038	.0005	3.9	< .01
***	Pct · N · Mag · Imp	8	.0052	.0007	5.4	< .001
***	Pct · N · L · Imp	16	.0092	.0006	4.7	< .001
*	Pct · Mag · L · Imp	8	.0028	.0003	2.8	.01
·	N · Mag · L · Imp	8	.0019	.0002	1.9	.08
·	Method · Pct · N · Mag · L	24	.0047	.0002	1.6	.08
	Method · Pct · N · Mag · Imp	24	.0034	.0001	1.2	.33
	Method · Pct · N · L · Imp	48	.0081	.0002	1.4	.14
	Method · Pct · Mag · L · Imp	24	.0029	.0001	1.0	.49
	Method · N · Mag · L · Imp	24	.0040	.0002	1.4	.18
·	Pct · N · Mag · L · Imp	16	.0033	.0002	1.7	.08
	Residuals	48	.0059	.0001		

APPENDIX D

R code for Randomization Test simulations

```
require(irtoys)

#############################################################################
#########################  Manipulated factors  #############################
#############################################################################
numitems <- 20               ## Test length
numDIF <- 4                  ## Number of DIF items
DIF.b <- .4                  ## DIF magnitude
n.ref <- 100                 ## Number of examinees in reference group
n.foc <- 100                 ## Number of examinees in focal group
ability.difference <- 0      ## Mean ability for focal group

nsims <- 100                 ## Number of reps
nperms <- 101                ## Number of randomizations for Randomization Test


Group <- c(rep(0,n.ref),rep(1,n.foc))
n <- n.ref+n.foc
DIFitems <- (numDIF>0)*1:numDIF
nonDIFitems <- (numDIF+1):numitems
refgroup <- 1:n.ref
focgroup <- (n.ref+1):n


######## Initialize matrices for results###############
pvals.randtest.equated <- pvals.randtest.not.equated <- matrix(ncol=numitems,nrow=nsims)


#############################################################################
#################################### Simulations ############################
#############################################################################
for (i in 1:nsims)  {
  a.ref <- rep(1,numitems)
  a.foc <- a.ref
  b.ref <- rnorm(numitems)
  b.foc <- b.ref + c(rep(DIF.b,numDIF),rep(0,numitems-numDIF))
  c.ref <- c(rep(0,numitems))
  c.foc <- c.ref
  parms.ref <- cbind(a.ref,b.ref,c.ref)
  parms.foc <- cbind(a.foc,b.foc,c.foc)
  parms <- rbind(parms.ref,parms.foc)

  thetas.ref <- rnorm(n.ref)
  thetas.foc <- thetas.ref[1:n.foc]+ability.difference

  responses.ref <- sim(ip=parms.ref, thetas.ref)
  responses.foc <- sim(ip=parms.foc, thetas.foc)
  responses <- cbind(rbind(responses.ref,responses.foc))
```

```
###Estimate difficulty parameter for reference group and focal group
  b.ref.est <- est(responses.ref,model="1PL",rasch="T",engine="ltm")$est[,2]
  b.foc.est <- est(responses.foc,model="1PL",rasch="T",engine="ltm")$est[,2]
  b.diff.est.not.equated <- b.foc.est - b.ref.est
  b.foc.est.equated <- b.foc.est + mean(b.ref.est) - mean(b.foc.est)   ####EQUATING
  b.diff.est.equated <- b.foc.est.equated - b.ref.est
  diffs.rand.equated <- diffs.rand.not.equated <- matrix(nrow=nperms,ncol=numitems)

  ### Randomizations.  Each row is one randomization. Each column is one item.
  for (j in 1:nperms) {
      ## Randomize group membership
    ref.rand <- sample(1:n,n.ref,replace=F)
    foc.rand <- seq(1:n)[-ref.rand]
      ## Estimate b parameters for both (randomized) groups
   b.ref.rand <- est(responses[ref.rand,],model="1PL",rasch="T",engine="ltm")$est[,2]
   b.foc.rand <- est(responses[foc.rand,],model="1PL",rasch="T",engine="ltm")$est[,2]
    b.diff.rand.not.equated <-  b.foc.rand-b.ref.rand
    diffs.rand.not.equated[j,] <- b.diff.rand.not.equated
    b.foc.rand.equated <- b.foc.rand + mean(b.ref.rand)- mean(b.foc.rand)
    b.diff.rand.equated <-  b.foc.rand.equated-b.ref.rand
    diffs.rand.equated[j,] <- b.diff.rand.equated
  }
      ## For each item: Calculate ranking of actual difference in estimated b
      ## compared to randomization differences in estimated b
      for (k in 1:numitems) {
        pvals.randtest.equated[i,k] <-
           mean(abs(b.diff.est.equated[k]) < abs(diffs.rand.equated[,k]))
        pvals.randtest.not.equated[i,k] <-
           mean(abs(b.diff.est.not.equated[k]) < abs(diffs.rand.not.equated[,k]))
      }
}

#############################################################################
######################################## Results ##############################
#############################################################################
Hits.randtest.equated <- FalsePos.randtest.equated  <- Hits.randtest.not.equated <-
  FalsePos.randtest.not.equated <- array()
sig <- seq(0,numitems,.01)/numitems

###########  Calculate Hit rates and False Positive rates  ###################
###########  for each alpha in 0, .0005, …,  1
####################
for (L in 1:length(sig)) {
  Hits.randtest.equated[L] <-
    mean(pvals.randtest.equated[,DIFitems] < sig[L],na.rm=T)
  FalsePos.randtest.equated[L] <-
    mean(pvals.randtest.equated[,nonDIFitems] < sig[L],na.rm=T)
  Hits.randtest.not.equated[L] <-
    mean(pvals.randtest.not.equated[,DIFitems] < sig[L],na.rm=T)
  FalsePos.randtest.not.equated[L] <-
    mean(pvals.randtest.not.equated[,nonDIFitems] < sig[L],na.rm=T)

Hits.randtest.equated <- c(Hits.randtest.equated,1)
```

```
Hits.randtest.not.equated <- c(Hits.randtest.not.equated,1)
FalsePos.randtest.equated <- c(FalsePos.randtest.equated,1)
FalsePos.randtest.not.equated <- c(FalsePos.randtest.not.equated,1)

######### ROC Curves #############
plot(FalsePos.randtest.equated ,Hits.randtest.equated,type="l",xlim=c(0,1),
  ylim=c(0,1),col="blue",lwd=2,xlab="False alarm rate",ylab="Hit rate")
abline(0,1,col="grey50")
lines(FalsePos.randtest.not.equated ,Hits.randtest.not.equated,col="red",lwd=2)

####### Area under the curve #########
height.randtest.equated <-
  (Hits.randtest.equated[-1]+Hits.randtest.equated[-length(Hits.randtest.equated)])/2
width.randtest.equated <- diff(FalsePos.randtest.equated)
AUC.randtest.equated <- sum(height.randtest.equated*width.randtest.equated)
height.randtest.not.equated <- (Hits.randtest.not.equated[-1]
   + Hits.randtest.not.equated[-length(Hits.randtest.not.equated)])/2

width.randtest.not.equated <- diff(FalsePos.randtest.not.equated)
AUC.randtest.not.equated <-
  sum(height.randtest.not.equated*width.randtest.not.equated)
AUC.randtest.equated;AUC.randtest.not.equated
```

APPENDIX E

R function for Randomization Test

```
############## Notes ##############
##  Modeled after functions in difR package.
##  Returns 2 p-values for each item (With and without equating).
##
## "Responses": Data matrix of 0s and 1s. Each row must be one respondent.
##  Each column must be one item.
## "Group": Vector of group membership, corresponding to rows of Responses.
##  0=Ref, 1=Foc.
## "nperms": Number of randomizations for randomization distribution.
###################################


difRand <- function(Responses, Group, nperms=100) {
  responses.ref <- Responses[which(Group==0),]
  responses.foc <- Responses[which(Group==1),]
  n <- nrow(Group)
  n.ref <- sum(Group==0)
  n.foc <- sum(Group==1)

  ############################ Parameter estimation ############################
  require(irtoys)
   ## Estimate item difficulties for reference group
    b.ref.est <- est(responses.ref,model="1PL",rasch="T",engine="ltm")$est[,2]
   ## Estimate item difficulties for reference group
    b.foc.est <- est(responses.foc,model="1PL",rasch="T",engine="ltm")$est[,2]
   ## Differences in estimated item difficulty, without equating
    b.diff.est.not.equated <- b.foc.est - b.ref.est
   ## Equating
    b.foc.est.equated <- b.foc.est + mean(b.ref.est) - mean(b.foc.est)
   ## Differences in estimated item difficulty, with equating
    b.diff.est.equated <- b.foc.est.equated - b.ref.est

  ############################### Randomizations ###########################
  diffs.rand.equated <- diffs.rand.not.equated <-
     matrix(nrow=nperms,ncol=ncol(Responses))
  ###Each row is one randomization. Each column is one item.
  for (j in 1:nperms) {
    ##Randomize group membership
     ref.rand <- sample(1:n,n.ref,replace=F)
     foc.rand <- seq(1:n)[-ref.rand]
    ##Estimate b parameters for randomized groups
     b.ref.rand <-
       est(Responses[ref.rand,],model="1PL",rasch="T",engine="ltm")$est[,2]
     b.foc.rand <-
       est(Responses[foc.rand,],model="1PL",rasch="T",engine="ltm")$est[,2]
     b.diff.rand.not.equated <-  b.foc.rand-b.ref.rand
     diffs.rand.not.equated[j,] <- b.diff.rand.not.equated
```

```
    b.foc.rand.equated <- b.foc.rand + mean(b.ref.rand)- mean(b.foc.rand)
   ####EQUATING
    b.diff.rand.equated <-  b.foc.rand.equated-b.ref.rand
    diffs.rand.equated[j,] <- b.diff.rand.equated
  }

  ##################### p-values ####################
  pvals.randtest.equated <- pvals.randtest.not.equated <- array()
  for (k in 1:ncol(Responses)) {
    pvals.randtest.equated[k] <- mean(abs(b.diff.est.equated[k]) <
      abs(diffs.rand.equated[,k]))
    pvals.randtest.not.equated[k] <- mean(abs(b.diff.est.not.equated[k]) <
      abs(diffs.rand.not.equated[,k]))
  }

  results <- matrix(nrow = ncol(Responses),ncol=2)
  rownames(results) <- matrix(paste("Item", 1:ncol(Responses),sep=" "))
  colnames(results) <- c("p-value: Not equated", "Equated")
  results[,1] <- pvals.randtest.not.equated
  results[,2] <- pvals.randtest.equated
  return(results)
}
```

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceeding of the second international symposium on information theory* (267–281). Budapest, Hungary: Akademiai Kiado.

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

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289-300.

Breslow, N. E. and Day, N. E. (1980). *Statistical methods in cancer research: Volume 1 - The analysis of case-control studies.* Lyon: International Agency for Research on Cancer (*IARC Scientific Publications* No. 32)

Broer, M., Lee, Y., Rizavi, S., and Powers, D. (2005). *Ensuring the fairness of GRE writing prompts: Assessing differential difficulty* (Research report RR-05–11). Princeton, NJ: Educational Testing Service.

Candell, G.L. and Drasgow, F. (1988). An iterative procedure for equating metrics and assessing item bias in item response theory. *Applied Psychological Measurement,* **12***,* 253-26.

Chen, J. H., Chen, C. T., and Shih, C. L. (2013). Improving the Control of Type I Error Rate in Assessing Differential Item Functioning for Hierarchical Generalized Linear Model When Impact Is Presented. *Applied Psychological Measurement*, **38**(1), 18–36.

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., Holland, P., and Thayer, D. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In Holland, P.W. and H. Wainer (Eds.), *Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates. 137-166.

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

Edgington, E., and Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Educational Testing Service. College Board Scholastic Aptitude Test. Princeton, N. J.

Embretson, S., and Reise, S. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.

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

Fisher, R. A. (1935). *The Design of Experiments* (9th Ed., 1971). New York: Hafner.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1), pp.1-22. URL http://www.jstatsoft.org/v33/i01/.

Fujii, Y., and Yanagaw, T. (1990). Homogeneity test with a generalized Mantel-Haenszel estimator for L2 × K contingency tables. *Journal of the American Statistical Association*, **85**(3), 744-748.

Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York: Springer Science+Business Media, Inc.

Guilera,G., Gómez-Benito, J., Hidalgo, M., and Sánchez-Meca, J. (2013). Type I Error and Statistical Power of the Mantel-Haenszel Procedure for Detecting DIF: A Meta-Analysis. *Psychological Methods*, **18**(4), 553–571.

Herrera, A. and Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality and Quantity,* **42**, 739–755.

Hidalgo-Montesinos, M. D., and Lopez-Pina, J.A. (2002). Two-Stage Equating in Differential Item Functioning Detection under the Graded Response Model with the Raju Area Measures and the Lord Statistic. *Educational and Psychological Measurement,* **62**, p. 32.

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

Houseman, E., Marsit, E., Karagas, M., and Ryan, L. (2007). Penalized Item Response Theory Models: Application to Epigenetic Alterations in Bladder Cancer. *Biometrics,* **63**(4), 1269-1277.

Howell, D. (2003). *Randomization tests.* Retrieved from www.uvm.edu/~dhowell/StatPages/ Randomization%20Tests.

Huggins, A.C. (2014). The Effect of Differential Item Functioning in Anchor Items on Populationv Invariance of Equating. *Educational and Psychological Measurement,* **74**, 627-658.

Jodoin, M. G., and Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education,* **14**, 329–349.

Kim, S.-H., Cohen, A., and Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, **32**, 261-276.

Kim, J., and Oshima, T. (2013). Effect of Multiple Testing Adjustment in Differential Item Functioning Detection. *Educational and Psychological Measurement*, **73**(3), 458–47.

Klieme, E., and Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education,* **16**(3), 385-402.

Lautenschlager, G., Flaherty, V., and Park, D. (1994). IRT Differential Item Functioning: An Examination of Ability Scale Purifications. *Educational and Psychological Measurement,* **54**, 21-31.

Li, H. H., and Stout, W. (1996). A New Procedure for Detection of Crossing DIF. *Psychometrika,* **61**(4), 647-677.

Liu, Q. (2011). Item purification in differential item functioning using Generalized Linear Mixed Models. Florida State University Libraries- Electronic Theses, Treatises and Dissertations. http://diginole.lib.fsu.edu/islandora/object/fsu%3A253928

Lopez, G. (2012). Detection and Classification of DIF Types Using Parametric and Nonparametric Methods: A comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST, and Logistic Regression Procedures. *Graduate Theses and Dissertations.* http://scholarcommons.usf.edu/etd/4131.

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

Magis, D., Beland, S., Tuerlinckx, F., and De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods,* **42,** 847-862.

Magis, D., and Facon, B. (2012). Angoff's delta method revisited: improving the DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, **65**, 302-321.

Magis, D., and Facon, B. (2012). Item Purification Does Not Always Improve DIF Detection: A Counterexample With Angoff 's Delta Plot. *Educational and Psychological Measurement,* **73**(2), 293–311.

Magis, D., Raiche, G., Beland, S., and Gerard, 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., and De Boeck, P. (2015). Detection of Differential Item Functioning Using the Lasso Approach. *Journal of Educational and Behavioral Statistics,* **20**(10), 1-25.

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

Mazor, K., Clauser, B., and Hambleton, R. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, **52**, 443-451.

Meade, A., and Wright, N. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, **97**(5), 1016–1031.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *J. R. Statist. Soc. B,* **70**(1), 53–71.

Narayanan, P., and Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, **18**(4), 315-328.

Partchev, I. (2015). irtoys: A Collection of Functions Related to Item Response Theory (IRT). R package version 0.2.0. https://CRAN.R-project.org/package=irtoys

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

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

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

Roussos, L., and Stout, W. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement,* **33**(2), 215-230.

Rogers, J., and Swaminathan, H. (1993). A Comparison of Logistic Regression And Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, **17**(2), 105-116.

Santelices, M.V., and Wilson, M. (2012). On the Relationship Between Differential Item Functioning and Item Difficulty: An Issue of Methods? Item Response Theory Approach to Differential Item Functioning. *Educational and Psychological Measurement,* **72**(1)*,* 5–36.

Schauberger, G., and Tutz, G. (2016). Detection of differential item functioning in Rasch models by boosting techniques. *British Journal of Mathematical and Statistical Psychology,* **69**, 80–103.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464. doi:10.1214/aos/1176344136

Scott, N.W., et al. (2007). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology,* **62**, 288-295.

Segall, D.O. (1983). *Test characteristic curves, item bias and transformation to a common metric in item response theory: A methodological artifact with serious consequences and a simple solution*. Unpublished manuscript, University of Illinois, Department ofPsychology, Champaign-Urbana.

Sen, S. (2014). *Permutation tests*. Retrieved from www.biostat.ucsf.edu/sen/statgen14/permutation-tests.html.

Shealy, R. and W. Stout (1993). A Model-Based Standardization Approach that Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DTF as well as Item Bias/DIF. *Psychometrika*, **58**(2), 159-194.

Sireci, S. and J. Rios (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, **19**(2-3), 170-187.

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *American Statistician,* **40**(2), 106-108.

Stark, S., Chernyshenko, O., and Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, **29**(3), 184-203.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, **55**, 293-325.

Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,* **27**, 361-37.

Thissen, D., Steinberg, L., and Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, **99**, 118-128.

Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–170). Hillsdale, NJ: Erlbaum.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, **58**, 267–288.

Tutz, G., and Schauberger, G. (2015). A Penalty Approach to Differential Item Functioning in Rasch Models. *Psychometrika*, **80**(1), 21–43.

Van De Water, E. (2014). A meta-analysis of Type I error rates for detecting differential item functioning with logistic regression and Mantel-Haenszel in Monte Carlo studies. Dissertation, Georgia State University. http://scholarworks.gsu.edu/eps_diss/113

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

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, **68**, 49–67.