Predictive Modeling of Uniform Differential Item Functioning Preservation Likelihoods After Applying Disclosure Avoidance Techniques to Protect Privacy

Marlow Q. Lemons

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Educational Research and Evaluation

Gary Skaggs, Co-Chair

Bimal Sinha, Co-Chair

Yasuo Miyazaki

George Terrell

Laura McKenna

March 7, 2014

Blacksburg, Virginia

Keywords: differential item functioning, disclosure avoidance, data swapping

Predictive Modeling of Uniform Differential Item Functioning Preservation Likelihoods After Applying Disclosure Avoidance Techniques to Protect Privacy

Marlow Q. Lemons

Abstract

The need to publish and disseminate data continues to grow. Administrators of large-scale educational assessment should provide examinee microdata in addition to publishing assessment reports. Disclosure avoidance methods are applied to the data to protect examinee privacy before doing so, while attempting to preserve as many item statistical properties as possible. When important properties like differential item functioning are lost due to these disclosure avoidance methods, the microdata can give off misleading messages of effectiveness in measuring the test construct. In this research study, I investigated the preservation of differential item functioning in a large-scale assessment after disclosure avoidance methods have been applied to the data. After applying data swapping to protect the data, I attempted to empirically model and explain the likelihood of preserving various levels of differential item functioning as a function of several factors including the data swapping rate, the reference-to-focal group ratio, the type of item scoring, and the level of DIF prior to data swapping.

Dedication

I dedicate this dissertation to my younger brother, Marvin, and to the love of my life, Tony.

Acknowledgments

There are five individuals that I cannot go without acknowledging. To my co-chair Gary Skaggs, it was an honor to be your student for the past four years. I enjoyed every course that I enrolled in with you and every lecture within those courses. The discussions and debates that we shared will be memorable and invaluable. Thank you for believing in my potential and for mentoring me. To Bimal Sinha, I thank you for being a great mentor and for being patient with me in times of panic. You helped me produce a piece of literature that gives honor to the graduate program and the federal government. I hope that I made you proud. To Laura McKenna, you have been the pivotal piece to producing this resarch. In 2008, you had tens of applicants to choose from to select your intern. At that time, I had no experience with disclosure avoidance. If it were not for you selecting me as your intern and teaching me this topic, this dissertation would not exist. I thank God for letting our paths cross and, for that, I will always be your "minion". To Yasuo Miyazaki, I asked you to be a part of my committee because you are to me an expert in simulations and regression. I am ever grateful that you were on my committee to help me produce a study that did not cut any corners. To George Terrell, I thank you so much for stepping in as a committee member. I needed a statistician with a vast knowledge of categorical data analysis. I could not have chosen a better person than you. You are an awesome colleague and I look up to you so much for advice.

Finally, I would like to thank other individuals who impacted my life in my journey towards graduation. To Penny Burge, I cannot thank you enough for your strength, advising, and knowledge. You not only accepted me into the EDRE program, but you instilled in me the love, respect, and appreciation for qualitative research that will not go void. I make that promise to you. To my homeboy of homeboys Deepu George (grammar genius), Jiashen You (Latex programming genius), and Jason Thweatt (logic genius), thank your support and help. To Eric Smith (Head of the Department of Statistics at Virginia Tech), thank you for your financial and moral support to help me complete this degree to make our department complete. Finally, to the faculty and staff of the Department of Statistics... thank you for your support! There are no more cracks in the pot now.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Significance of Research Study

To address the significance of this research study, it is important to acknowledge that the idea was developed from a combination of topics applied during the apprenticeship component of the doctoral program and research conducted through the Center for Disclosure Avoidance Research at the United States Census Bureau (USCB).

Several reasons explain why this research study is important and significant. First, there is a need for data privacy. The advancement of technology has increased the capability to disseminate data and, unfortunately, the potential for its users to abuse the data by identifying individuals whose information was collected. Whether the data were disseminated through microdata or tabular statistics, a user could link his or her own set of information

or some level of prior knowledge of the population to it in order to identify individuals with unique characteristics. The purpose of data privacy is to preserve the "rights of an individual" by keeping his or her "information about him or herself from others" (Assembly of Behavioral and Social Sciences, 1979). In the context of educational assessment data, examinees of a particular race, gender, or disability, or examinees needing some sort of testing accommodation can be targets of data intrusion when these groups are small in size. Disclosure avoidance techniques are essential to protect the privacy of these individuals.

Second, there is need to preserve statistical properties in disseminated data. When it comes to data produced from educational assessments, that need still exists. Administrators of large assessments attempt to report findings of students' performances, item analysis, and, if possible, publish microdata containing records of individual responses. In regards to publishing data, administrators too wish to preserve some or all of the item analysis properties while preserving students' privacy. Prior to administration, assessment items are constantly tested for forms of bias like differential item functioning (DIF). DIF, in a short explanation, is a variety of statistical tools used to identify items that exhibit a significant difference in its performance when stratified by a characteristic of the examinees. A more formal definition of DIF is provided in the next chapter. When there is a significant difference in an item's performance by a factor not a part of the test construct, the need to allow accommodations is pivotal to provide fairness to subsets of examinees that need it (Camilli, 2006). Despite trial testing and past analysis that suggest otherwise, an item can show signs of DIF in any year that it is used. With that said, it is important to address that assessment

administrators should preserve the presence of DIF through data dissemination.

Finally, there is a need to investigate how upholding data privacy affects the ability to preserve statistical properties. This research study looks at one combination by investigating how DIF detection is affected when disclosure avoidance techniques are applied to education data. Currently, little to no research exists that explains the how these protection methods impact statistical measures of item bias. In government statistical journals, there are publications that investigated ways to preserve various statistical properties. For example, Nayak et al. (2011) extended the idea by Massell and Funk (2007) of their use of multiplicative noise masking to show that perturbed cell totals (that is, marginal frequencies altered to preserve privacy) are unbiased estimates of the original totals since they are symmetrically distributed. They found that moments and correlations from "noise-perturbed" data can "unbiasedly recover" the same moments and correlations found in the original data (Nayak et al., 2011). In education measurement journals, popularly published articles explored how various student accommodations or demographics affect item performance. For example, Bolt and Ysseldyke (2006) used DIF analysis to explore how a read-aloud accommodation affects item performance in mathematical versus a language arts assessments to detect performance differences due to accommodations. They identified DIF items in both types of assessments, and that the accommodation "did not allow for perfectly comparable measurement on either assessment" for which they defined as "measurement incomparability" (Bolt and Ysseldyke, 2006, p. 348). It is important to note that identifying the effect of testing accommodations is not the only type of DIF that one could test for, and that gender and

ethnicity DIF are more popularly tested factors. Nevertheless, the research in this dissertation will benefit the topic of DIF by uniquely addressing how methods to protect examinee privacy before disseminating the data can impact DIF.

1.2 Research Study Questions

The objective of this paper was to study the likelihood of preserving uniform DIF in an item after using data swapping as a disclosure avoidance technique to protect the data. In investigating this objective, three specific research questions were answered.

1. What is the rate at which uniform DIF is preserved in dichotomous and polytomous items after applying data swapping protection to the data?

2. Is there an association between the likelihood of preserving uniform DIF, the data swapping rate, the item scoring, the reference-to-focal group ratio, and the severity of the DIF originally detected?

3. Can the association between the likelihood of preserving uniform DIF, the data swapping rate, the item scoring, the reference-to-focal group ratio, and the severity of the DIF originally detected be explained using a generalized linear model?

The first question will be addressed using several sets of two-way contingency tables that analyze the behavior of item DIF before and after data swapping, with the sets of contingency tables being distinct by item scoring. The second research question will apply the frequencies

from these tables to tests of association and measures of correlation. The reader should understand that the results from these methodologies will reflect the relationship between DIF found before versus after data swapping since the data swapping rate, item scoring, and reference-to-focal group ratio are fixed factors. Generalized linear models, namely the cumulative logit model, will be used to address the third question.

Chapter 2

Literature Review

2.1 Differential Item Functioning

Creating assessments that do not bias against subgroups of examinees is the key (American Educational Research Association, 1999), and so DIF analyses are statistical procedures used to measure such "item bias" in "assessment instruments" (Anderson and DeMars, 2002). When examinees from different subgroups of a population have a different probability of answering an item correctly "after being matched on the ability of interest", that item is said to be affected by DIF (Clauser and Mazor, 1998). Items that are cited as being affected by DIF are reviewed by test developers and testing experts to determine what revisions are needed for that item (Camilli, 2006). This is so that all examinees with the same ability level, regardless of age, gender, ethnicity, learning or physical disability, or

testing accommodation have an equal likelihood of earning the same score on every item. The purpose of assessments is to measure "construct-level differences" only without any additional bias due to the demographics of the examinees (Kamata and Vaughn, 2004). To apply DIF methods, the sample is first divided into a reference group and a focal group stratified by matching ability level groups. Customarily, the stratification is based on, but not limited to, the total test score used. The reference group refers to those examinees with no special accommodation, condition, or characteristic in which no bias is assumed to be against or for them. The focal group, however, contains those examinees with such special situations where their responses are compared to examinees from the reference group. When a significant difference in the likelihood of performance for an item exists between the reference and focal groups for at least one stratum, DIF is said to be present in the item. This DIF is dependent on the assumption that the matching criterion used is a true measure of examinee ability. If this fails, Clauser and Mazor (1998) define the detected impact as "item impact" rather than DIF.

The DIF of an item can be explained in one of two ways. One way involves a difference in performance that is independent of ability level. More specifically, differences in performances exist at a particular ability level but there exists a commonality in the performance across all ability levels. This phenomenon is described as uniform DIF (Holland and Thayer, 1988; Camilli, 2006), and there are several DIF methods that can be used to detect this. These are described later in Chapter 2.3. Alternatively, DIF detection can be explained as a function of item performance and examinee ability levels. Not only are the odds ratios

distinct at different ability levels for an item, but there exists an increasing or decreasing trend in these odds ratios. This is an example of nonuniform DIF, and the concern is that the level of DIF is more or less severe at particular ability levels.

The methods used to detect DIF items vary. Kamata and Vaughn (2004) described how the Mantel-Haenszel Test (Holland and Thayer, 1988) is a popular contingency-table-based methodology for a dichotomously-scored item to detect a significant difference in the proportion of examinees who correctly answer that item. These authors also described the logistic regression method (Swaminathan and Rogers, 1990) which creates item-response theory models for both groups for an item (Kamata and Vaughn, 2004). These models represent the likelihoods of answering an item correctly by group and eventually create an inferential test statistic that measures the difference in their curves to determine if one group has higher or lower chances of performing well for that item. Another methodology called the SIBTEST (Shealy and Stout, 1993) compares the weighted difference in an item's performance between the focal and reference groups and divides this difference by the item's overall standard error. A slight advantage of this procedure is said to help identify and remove DIF items to produce a subset of remaining test items that are free of DIF. Regardless of which DIF methodology is used, one should explore his or her analyses with caution since DIF analysis is sample-size based. Groups with large sample size ratios pose a high risk for Type I error (Meyer et al., 2004; Wang and Su, 2004). This is primarily because the standard errors for the responses are significantly different.

The popularity of DIF has led to numerous publications that analyzed various types of

DIF over several large-scales assessments. Cohen et al. (2005) detected 22 of 29 DIF items that biased against ninth grade students with learning disabilities on the 2003 reading version of the Florida Comprehensive Assessment Test using the likelihood ratio test. Elbaum (2007) also compared overall test scores and item analysis for students with learning disabilities versus those without on the Stanford Achievement Test. Additionally, Elbaum (2007) examined how the read-aloud accommodation improved scores when offered to learning-disability and non-learning-disability students. The study identified several DIF items favoring those using the read-aloud accommodation which resulted in "improved performance scores" regardless of possessing a learning disability (Elbaum, 2007).

The usage of DIF methodologies has also led researchers to investigate how well they perform under special conditions. For example, (Aguerri et al., 2009) studied how one methodology can "erroneously" identify items for DIF when its assessment contains a small number of items. More specifically, the authors discovered that test analysts should be cautioned of creating short assessments containing items with high levels of discrimination and difficulty. A simulation study by Woods (2008) showed how non-normal distributions in examinee ability levels can affect the Type I error rate in one DIF methodology. (Meyer et al., 2004) examined how one DIF methodology, which is commonly used to compare large reference and focal group sizes, can still be used on polytomous items in which the focal group contains as little as seventy-six examinees.

Several literature reviews have become popular tools to understand the dynamics of DIF. Zumbo's description of the "generations of DIF" (2007) explains how DIF has evolved since

its emergence in the late 1980s. He described how the first generation as one that emphasizes the need and importance in assessment evaluation, the second generation as a period of popularizing DIF statistical methods involving contingency tables and regression models, and the third generation as the "praxis of DIF" (Zumbo, 2007) to which the state of DIF is present in. Penfield and Lam (2000) described the distinction between statistical DIF and substantial DIF. They described statistical DIF as DIF detected through statistical methods and substantial DIF as the "non-target constructs" that are the cause of the DIF detection (Penfield and Lam, 2000). Finally, Williams (1997) warned researchers not to equate item bias with DIF detection as being one and the same since item bias is investigated as a result of detecting DIF in an item.

2.2 Testing Accommodations and DIF

A student's learning process and the process of how a student demonstrates knowledge of what was learned are interdependent. The No Child Left Behind Act of 2001 (NCLB, 2002) mandated test developers to redesign the conditions of their testing environments and assessments in an effort to "measure student knowledge and skills in a content area" independent of a student's subgrouping (Cox et al., 2006). Doing so requires test administrators to provide various testing accommodations to its examinees. By definition, an accommodation is a "change in the testing environment, procedures, or presentation that does not alter what the test measures or the comparability of test scores" (Kim et al., 2009). In all, providing

testing accommodations for students that need it brings "balance to the testing process" and "improves the accessibility of the test items" (Finch et al., 2009).

Testing accommodations vary, and they can range from reading test items to a student to providing extra testing time. Nevertheless, testing accommodations are not meant to provide students with "an advantage", but rather to give individuals with learning or physical disabilities assistance so that their scores are comparable to those that do not require such assistance (Cohen et al., 2005). Some testing accommodations are required to meet federal standards such as the No Child Left Behind Act of 2001 (NCLB, 2002) and the Individuals With Disabilities Education Improvement Act of 2004 that "challenge our education system" (Salend, 2008). Others are not accommodations required by laws or act, but rather accommodations for students to perform at their best. Examples include creating an exam in a different language for students born with English as their second language, allowing students to voice respond to questions, or students needing assistance in writing their responses. But regardless of the testing accommodation, these adjustments allow for students with a testing disability to "meaningfully participate and demonstrate their skills and knowledge on tests" (Ketterlin-Geller et al., 2007).

There are several examples of studies that have investigated various forms of accommodation DIF. Finch et al. (2009) applied uniform and non-uniform DIF analyses to identify several items that distinguished significant item performance between accommodated and non-accommodated students with disabilities from grades three through eight. Kettler et al. (2005) identified performance differences between students requiring versus not requiring

testing accommodations due to disability. Also, Bolt and Ysseldyke (2006) examined how the read-aloud accommodation identified several item performance differences. These types of applied studies had three similar conclusions. First, they showed that item performance differences due to accommodations are limited to particular grade levels and, thus, DIF analysis can be applied to assessments at any grade level. Second, they concluded that item performance differences due to accommodations can span over mathematics and reading assessments. Finally, these studies identified items that favored students that did not use the accommodation.

2.3 Statistical Methods to Test DIF

Several publications outline various methods that can detect DIF. Polytomously-scored items require different DIF statistical tests. For example, logistic regression is a popular methodology for identifying item-performance bias. Proposed by Swaminathan and Rogers (1990), this nonparametric technique detects uniform and nonuniform DIF. More recently, a standardized item bias test, or SIBTEST, that creates a statistic that is "based on the ratio of weighted difference in proportion correct between reference and focal groups to its standard errors" (Shealy and Stout, 1993; Clauser and Mazor, 1998). This procedure is said to be just as robust as the logistic regression method even with relatively small samples sizes. For more details on how these methods work, we refer to Clauser and Mazor (1998). For this study, attention is given to the Mantel-Haenszel Test for dichotomously-scored items and

the Mantel Chi-Squared Test for polytomously-scored items since they are commonly used for these purposes (Clauser and Mazor, 1998). These methods are discussed below.

2.3.1 Statistical Methods for Dichotomously-Scored Items

An item is said to be dichotomously-scored when there are only two possible scores assigned to every examinee's response. Commonly, the two scores are zero and one where zero represents an incorrect response and one represents a correct response. With this scoring form, the average of the scores for a dichotomous item is equivalent to the proportion of examinees that gave a correct response. When separating the examinees into reference and focal groups, one can calculate and compare the two averages as well as perform inferential methods that determine whether the two averages show significance to conclude whether the type of group that an examinee is in affects the likelihood of answering the dichotomous item correctly.

There are several inferential methods that can be used to detect DIF. The Mantel-Haenszel Test is a nonparametric, inferential method proposed by Holland and Thayer (1988), to analyze dichotomously-scored items for DIF. For large sample sizes and contingency table cell counts, this test is "highly efficient for its statistical power" (Clauser and Mazor, 1998) despite being sample-size dependent. This procedure is popularly used to test for uniform DIF, but it is important to note that this is not the only procedure used for that purpose.

Table 2.1 represents a two-way contingency table summarizing the dichotomous scoring of an item between the reference and focal groups with ability levels in the k^{th} stratum.

Table 2.1: Contingency Table of Symbolic Notation for a Dichotomously-Scored Item

Group	1 = Correct	0 = Incorrect	Total
Reference	n_{11k} (N_{11k})	n_{12k} (N_{12k})	n_{1+k}
Focal	n_{21k} (N_{21k})	n_{22k} (N_{22k})	n_{2+k}
Total	n_{+1k}	n_{+2k}	n_{++k}

Thus, n_{11k}, n_{12k}, n_{21k}, and n_{22k} represent the frequencies and their expected values are represented in the parentheses respectively. Under the null hypothesis, examinees should have the same likelihood and odds of answering an item correctly regardless of the group which he or she is from and the stratified matching criterion. Evidence suggesting otherwise will become evident as the value of n_{11k} deviates further away from its expected count N_{11k}. The observed test statistic formula takes the form

$$Q = \frac{\left(\left| \sum_{j=1}^{K} (n_{11j} - N_{11j}) \right| - \frac{1}{2} \right)^2}{\sum_{j=1}^{K} Var(n_{11j})} \tag{2.1}$$

where

$$E(n_{11j}) = N_{11j} = \frac{n_{1+j} n_{+1j}}{n_{++j}},$$

and

$$Var(n_{11j}) = \frac{n_{1+j} n_{2+j} n_{+1j} n_{+2j}}{n_{++j}(n_{++j} - 1)}.$$

The distribution of this test statistic approximately follows a Chi-Squared distribution with one degree of freedom. An item with an observed test statistic value above 3.841

provides significant evidence at the five percent level of significance of DIF (5.024 is the critical test statistic at the 2.5% level of significance and 6.635 is the critical test statistic value at the 1% level of significance).

A measure of effect size for the Mantel-Haenszel Test is the odds ratio (Agresti, 2002; Camilli, 2006). Using the notation in Table 2.1, the odds ratio is calculated as the ratio of cross products. That is,

$$OR = \frac{\sum_k n_{11k} n_{22k} / n_{++k}}{\sum_k n_{12k} n_{21k} / n_{++k}}. \tag{2.2}$$

The construction of Equation (2.2) describes the magnitude of DIF towards the reference group compared to the focal group where the value of one represents equivalent performance on an item for the reference and focal groups. However, an odds ratio higher than one is evidence of a higher item performance for the reference group and the result of DIF favoring that group. An odds ratio less than one favors the focal group. This measure is typically converted into a log-odds statistic used for categorizing DIF. This is explained at the end of this subsection.

There are several conditions in which the Mantel-Haenszel Test would be the least useful in detecting item DIF. Although simple in form and popular in application (Clauser and Mazor, 1998), the test statistic formula for the Mantel-Haenszel Test is constructed to detect differences in group item performances and not to detect trends in effect sizes within an item. Hambleton and Rogers (1989) found that using this test for detecting nonuniform DIF

results in a significant loss of statistical power. Other studies measuring the efficacy of the Mantel-Haenszel Test have shown that behaviors such as small sample sizes and unusual distributions in ability levels between groups can result in an increase in its Type I Error (Meyer et al., 2004; Fidalgo et al., 2004). This has led to alternative measures to detect DIF such as the logistic regression procedure (Swaminathan and Rogers, 1990), the Likelihood Ratio Test (Thissen et al., 1988), and the Breslow-Day Test for non-uniform DIF (Camilli and Shepard, 1994; Breslow and Day, 1980).

Another popular test to detect item DIF is the Breslow-Day Test (Penfield, 2003). Similar to Mantel-Haenszel Test, this test statistic is sample size dependent and is of the same form as Equation 2.1. Contrary to the Mantel-Haenszel Test, which detects uniform DIF, the Breslow-Day Test detects nonuniform DIF. Nonuniform DIF can be alternatively described in reference to item response theory. Specifically, uniform DIF items are said to differ only by item difficulty between the two tested groups whereas nonuniform DIF implies that performance is now a function of an item's difficulty level, ability to guess, or an item's discrimination level.

The Mantel-Haenszel test statistic describes the magnitude of DIF, but it is difficult to interpret. Thus, the Education Testing Service (ETS) has created an alternative method for classifying this magnitude into 'low', 'medium', and 'high' levels of DIF using the estimated odds ratio values. Recall that an odds ratio compares the probabilities between the reference and focal students for an item. An item's odds ratio value that is greater than one suggests that the subjects in the reference group have a higher chance of performing well on that

item, odds ratio values less than one suggest that subjects in the focal group perform better on that item, and odds ratio values approximately equal to one say that both groups of students perform equally well on that item. The ETS scale calculates Δ_{MH_j}, which is a natural-log transformation of the odds ratio multiplied by a scalar factor of -2.35 (Camilli, 2006). Clauser and Mazor (1998) stated a description of ETS's item classification method:

> "Items classified in the first level, A, have a Δ_{MH_j} with an absolute value of less than 1.0 and/or have a value that is not significantly different from zero ($p > .05$). Items in the third level,C, have a Δ_{MH_j} with absolute value greater than 1.5 and are significantly greater than 1.0 (i.e. 1.0 is outside the confidence interval, around the estimated value). Items in the second level, B, are those that do not meet either of the other criteria. Items classified as A are considered to display little to no DIF and are considered appropriate for use in test construction. Items classified as B are used only if no A item is available to fill the content requirement of the test. Items classified as C are to be used only if the content experts consider them essential to meet the test specifications." (p.39)

2.3.2 Statistical Methods for Polytomously-Scored Items

An item that is polytomously scored allows an examinee to receive one of several possible scores rather than just two possible scores for his or her response. In the popular case, allowing several possible scores to be assigned for a response allows an examinee to receive

Table 2.2: Contingency Table for a Polytomously-Scored Item at the k^{th} Stratification Level

Group	0	1	$\cdots$	J-1	J	Total
Reference	n_{11k}	n_{12k}	$\cdots$	$n_{1(J-1)k}$	n_{1Jk}	n_{1+k}
Focal	n_{21k}	n_{22k}	$\cdots$	$n_{2(J-1)k}$	n_{2Jk}	n_{2+k}
Total	n_{+1k}	n_{+2k}	$\cdots$	$n_{+(J-1)k}$	n_{+Jk}	n_{++k}

no (0 points), partial (1,2,$\cdots$,$J-1$ points), or full credit (J points) for an item. With polytomously-scored items, the average no longer represents the percentage of examinees receiving full credit for an item. Rather, it represents the average score per examinee received for that item. Nevertheless, with two groups in question, one can compare these averages to determine whether DIF exists and its severity.

Table 2.2 provides an extension to the contingency table from Table 1 where polytomous scoring is allowed for an item at the k^{th} stratification level. Without loss of generality, consider an item where the test administrator assigns one of five possible scores for an examinee response. In this case, scores of one, two, or three points represent partial credit, zero point represents no credit, and four points represent full credit. Here, n_{2jk} represents the number of examinees in the focal group receiving j credit points in the k^{th} stratum whereas n_{1jk} represents that number of examinees in the reference group.

The Mantel Chi-Squared test statistic is a popular measure to detect the extent of DIF in polytomous items. This test statistic uses the total number of points accumulated from the examinees from the focal group across all stratification levels as its point estimate (Mantel, 1963). This point estimate is compared to the expected number of points that the examinees

should have if the distributions of scores are equivalent for the two groups. Extremely large or small total scores from what are expected is evidence to conclude that examinees in the focal group tend to have a higher or lower chance of performing as well as the examinees in the reference group. The Mantel Chi-Squared test statistic is calculated as

$$QM = \frac{\left(\sum_{i=0}^{T} F_i - \sum_{i=0}^{T} E(F_i)\right)^2}{\sum_{i=0}^{T} Var(F_i)} \tag{2.3}$$

where T represents the number of stratification levels, F_i is the total of the scores earned by examinees of the focal group, defined as

$$F_i = \sum_{j=0}^{J} x_j n_{2jk},$$

x_t is the matching criterion score, $E(F_i)$ represents the mean number of points under the null hypothesis of no bias detected defined as

$$E(F_i) = \frac{n_{2+k}}{n_{++k}} \sum_{j=0}^{J} x_j n_{+jk},$$

and $Var(F_i)$ represents the variance of F_i defined as

$$Var(F_i) = \frac{n_{1+k} n_{2+k}}{n_{++k}^2 (n_{++k} - 1)} \left[\left(n_{++k} \sum_{j=1}^{J} x_j^2 n_{+jk} \right) - \left(n_{++k} \sum_{j=1}^{J} x_j n_{+jk} \right)^2 \right]$$

where $x_j = j$. The distribution of QM is approximately Chi-Squared with one degree of

freedom. Therefore, similar to Equation 2.1, items with observed test statistic values above 3.841 provide significant evidence at the five percent level of significance of test accommodation DIF (5.024 is the critical test statistic at the 2.5% level of significance and 6.635 is the critical test statistic value at the 1% level of significance).

The Liu-Agresti common odds ratio (LACOR), ψ_{LA}, is a popular estimator used to measure effect size when DIF detection is observed with an item. According to Penfield and Algina (2003), for J scores, this statistic uses the first J-1 sets of cumulative frequencies across the matching criterion in calculating an odds ratio. Table 2.3 represents the notation for these cumulative frequencies. LACOR is calculated as

$$\widehat{\psi}_{LA} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J-1} A_{jk}D_{jk}/n_{++k}}{\sum_{k=1}^{K}\sum_{j=1}^{J-1} B_{jk}C_{jk}/n_{++k}} \tag{2.4}$$

where $A_{jk} = n^*_{1jk}$, $B_{jk} = n_{1+k} - n^*_{1jk}$, $C_{jk} = n^*_{2jk}$, and $D_{jk} = n_{2+k} - n^*_{2jk}$. LACOR values that deviate from the value of one are evidence of possible DIF and measure the magnitude of the effect. To place the LACOR on a symmetric scale (Penfield and Algina, 2003), one could observe the Liu-Agresti common 'log' odds ratio (LACLOR) by calculating the log of the inverse of the LACOR. In this case,

$$\widehat{\alpha}_{LA} = ln\left(\widehat{\psi}^{-1}_{LA}\right).$$

Since $ln\left(\widehat{\psi}^{-1}_{LA}\right) = 0$, when $\widehat{\psi}_{LA} = 1$, test items where the absolute LACLOR deviate from

Table 2.3: Cumulative Frequency Notation for an Item at the k^{th} Stratification Level

Group	0	1	$\cdots$	J-1	J
Reference	n^*_{11k}	n^*_{12k}	$\cdots$	$n^*_{1(J-1)k}$	
Focal	n^*_{21k}	n^*_{22k}	$\cdots$	$n^*_{2(J-1)k}$	

zero is evidence of test accommodation DIF (Penfield, 2003). However, absolute LACLOR values greater or less than the value of zero suggest possible item bias against the focal or reference groups respectively. For the derivation and distribution of the LACOR, see Liu and Agresti (1996). The Educational Testing Service (ETS) provides a scale that categorizes the LACLOR of an item into one of three DIF levels. Specifically, they suggested that absolute LACLOR values within 0.43 from zero show 'negligible' DIF, absolute LACLOR values between 0.43 and 0.64 show 'moderate' signs of DIF, and absolute LACLOR values of at least 0.64 show 'large' signs of DIF (Zieky, 1993). Dorans and Schmitt (1993) addressed a more popular method that is also congruent with the ETS scale in which one could observe the standardized effective size. The standardized effective size is equal to the difference between the focal and reference means divided by the standard deviation of the scores for the two groups combined. Polytomous items with an absolute standardized effective size within 0.17 of zero are said to have negligible DIF and are notated as "AA" items. Otherwise, polytomous items with an absolute standardized effective size higher than 0.17 but within 0.25 are said to have moderate signs of DIF and are notated as "BB" items, and items past the "BB" range items are said to have large signs of DIF and are notated as "CC" items.

2.3.3 Sample Size and DIF

Although the Mantel-Haenszel and Mantel tests are convenient asymptotic procedures to detect item DIF, its dependence on the sizes of the reference and focal groups is pivotal for them to effectively detect item DIF. This dependence has led resarchers to study the impact of group sizes and their ratios on this pheonomenon. Mazor et al. (1992) conducted simulations to investigate how small in size the reference and focal groups have to be before the Mantel-Haenszel statistic loses its ability to identify DIF items. These researchers intentionally created several DIF items of various DIF severity while using equal sample sizes of 2,000, 1,000, 500, 200, and 100 examinees in each group. They found that when these groups possed equal mean ability levels, only 18% of DIF items were successfully identified by the Mantel-Haenszel statistic when 100 examinees were used, almost 30% of DIF items were detected when groups contained 200 examinees each, almost 40% of DIF items were identified for groups containing 500 examinees, 61% of DIF items were detected with 1,000 examinees, and 74% of DIF items were detected with 2,000 examinees. Another study by Herrera and Gómez (2008) investigated how several ratios between the reference and focal group sizes, and considered the case where the reference sample size was large (1500 examinees) and small (500 examinees). These researchers found that the Mantel-Haenszel statistic was prone to producing a 8% false positive DIF- item rate for large group ratios and when the sample sizes were small rather than large. Finally, a similar study by Narayanan and Swaminathan (1994) also tested several group ratios as large as 10:1 to model uniform DIF detection rates. They found that the detection rate increased as the sample size increased.

The former studies validated the idea that these DIF methods can experience increased Type I error rates (Mazor et al., 1992; Shepard et al., 1985; Spray, 1989; Roussos and Stout, 1996). As a result, other researchers have suggested alternative methods involving the use of exact rather than asymptotic distributions to detect item DIF (Bolt, 2002) or outlier DIF methods that are robust against sample size (Magis and de Boeck, 2012).

2.4 Disclosure Avoidance

Agencies that produce, disseminate, and distribute large sources of data are often, if not required, held to standards of privacy protection of the individuals measured. For example, government agencies are held to federal laws like Title 13 and Title 26 which authorize these agencies to provide a level of protection on the data. Nongovernmental agencies adhere to institutional review board standards that give them the moral responsibility respect the rights of the research subjects. In either case, a level of disclosure avoidance is considered.

Disclosure avoidance, also referred to as statistical disclosure control, is the process of protecting the privacy of respondents while publishing data or summaries of data. One must take into account the importance of respondent privacy to further understand the 'why' of this definition. Disclosure is the action where individuals "recognize or learn something that they did not know already" about a respondent through released data (Hunderpool et al., 2012). This is possible when a data proprietor releases any form of information about the respondents used. In most cases, this is through the release of microdata or statistics

produced from the data where disclosure can take place. In short, customers of data become intruders of data when they are able to "associate data record to a target person" with a high likelihood (Paass, 1988).

Hunderpool et al. (2012) described four ways in which data is reported: *tabular data*, *dynamic databases*, *microdata*, and *statistical analyses*. Tabular data, the most simplistic, involves reporting "static aggregate information" that is readily accessible for users (Hunderpool et al., 2012). This method of delivering data has been used for decades by several large agencies where cross-tabulations are provided on several attributes for users to access. Dynamic databases allow users to submit a 'query' to a database and receive summarized aggregate statistics. For example, the USCB is creating a Microdata Analysis System that is "designed to allow users to perform various statistical analyses" including "regression, cross-tabulation, and generation of correlation coefficients" (Lucero and Zayatz, 2010). This service is said to provide users with more "accurate information than what is provided from the microdata files" which are perturbed to protect confidentiality (Lucero and Zayatz, 2010). Microdata, a recent tool which involves using some or all of the data records captured at the person level (Ruggles and Ruggles, 1975), is said to "increase the flexibility and availability of information a user" (Samarati, 2001). This flexibility, consequently, makes it capable for data users to link several sets of microdata together in an effort to identify individuals and compromise disclosure avoidance. One way to avoid this is to alter the values for some of the records or remove variables from the microdata that would pose a disclosure risk. We will see in the next section one example to alter microdata.

The need and importance of disclosure avoidance existed since the "dawn of information technology in the 1960s" (Wacks, 2010, p.111). As the ability to use technology to disseminate data increases, so does the risk for disclosure. The way to prevent disclosure in data is to apply appropriate disclosure protection techniques. For the past four decades, several techniques have been introduced and incorporated. Brief descriptions of two popular techniques are discussed.

Synthetic data (Rubin, 1993) is fictitious data created from bootstrapping to replace the real data. Although it is considered to be fabricated, synthetic data is useful since it preserves the statistical behavior from the real data while capturing anonymity. However, there is some realness to the data since it bootstraps from record values in the real data. What makes synthetic data just as competitive to other options is its ability to impute missing observations (Raghunathan et al., 2003). One can also develop synthetic data using predictive models where a model exists and fits well with the original data. For example, Zayatz (2007) explained how some of the Census products use synthetic data to "target records that have potential disclosure risk" by synthesizing the demographic variables "that are causing the risk".

Cell suppression is a common method used when providing tabular data. With cell suppression, attribute classes are withheld from publication or concatenated with other classes (Hunderpool et al., 2012) in an effort to explicitly conceal sensitive information about a respondent (Doyle et al., 2001). In most cases, classes with small frequencies are often concatenated with neighboring classes so that the final displayed frequencies are large enough

not to identify information for any individual. Because cell suppression does not guarantee enough protection, several extensions have been made to supplement this method. One of the most popular tactics involve increasing the value of small frequencies in tabular data, or adding noise (Kelly et al., 1992). Although this prevents an intruder from separating the "noise from the signal" (Willenborg and de Waal, 2001), the drawback of this method is that artificial error has been introduced to the tabular data (Kelly et al., 1992).

Regardless of which type of data disseminated or the disclosure technique used, a disclosure review board reviews and approves the data before it is disseminated. For example, the United States Census Bureau's Disclosure Review Board reviews all data that is publicly released by the agency, and is responsible for making sure that data proprietors providing a level of disclosure protection to the data where no subset of respondents' right to privacy is compromised (Zayatz, 2007). Other agencies, like the National Center for Education Statistics, are responsible for the release of data as well as the procedures used to protect the data (Institute of Education Sciences, 2013). Both review boards carry one commonality; they are responsible for data owners to adhere to a code of ethics, laws, or standards to protect the populations they wish to describe. Specifically, they provide a secondary check at the potential disclosures that can exist from publishing data.

2.5 Data Swapping

Data swapping (Dalenius and Reiss, 1982) is one of several techniques used in disclosure avoidance, and is considered to be a member of the "post randomization" family of disclosure techniques (Willenborg and de Waal, 2001). It uses permutations in the process of creating its data; that is, the order of the data is rearranged rather than discarded. The main objective is to exchange the information of a small percentage of records in the data in an effort to protect those records with unique responses. The term, and its usage, was first introduced in 1978 by Dalenius and Reiss, but the probabilistic justification of its use was addressed the year before. Like the other protection techniques, the development of this technique was centered over the growing concerns to preserve the values of summary statistics, publish as much of the values of the original data, and yet "introduce uncertainty about sensitive data values" to an intruder (Fienberg and McIntyyre, 2004).

Although agencies apply slight variations to data swapping relative to their missions, there are several commonalities in these variations that should be addressed in order to provide the reader with a general understanding of how data swapping works. First, a target percentage of records to swap is determined. This percentage is usually set between one to five percent, with three percent being the most popularly used (Shlomo et al., 2010; Zayatz, 2007). Second, a key is created that is determined by the data proprietors. The key is a list of variables in which records with unique combinations of these variables are flagged as unique records and are eligible to be swapped during the data swapping process. Although not required,

the unique records can be individually ranked based on the number of variables that are uniquely sensitive. Records with higher ranks have a higher chance of being swapped. It is also important to note that not every flagged record has to be swapped. If the number of flagged records exceeds the target percentage, a deselection process is applied to remove flagged records from being swapped. Similarly, records that are not flagged can also be chosen to pair with records that are flagged. This may, as well, deselect flagged records. Nevertheless, when all of the selected records have been paired, only the values from the key are swapped. As a result of these steps, a new set of data is formed that is different from the original data and can be disseminated for others to use.

Several worldwide government agencies incorporate data swapping as a means to protect their data. For example, the decennial Census data begins with an unedited data file (CUF) and cleans the data to create an edited data file (CEF). Several disclosure avoidance techniques are applied to the data, with "targeted" data swapping being one of those techniques (Shlomo et al., 2010). After identifying unique households that pose a risk of disclosure in the data while allocating small percent of records to swap, the selected households are paired with each other and the swapping of geographic information is performed to create a Census decennial file (CDF) that is used to create the statistical tables made available for public use (Zayatz, 2007). Before publishing, the Census Bureau conducts evaluations that compare the CEF and CDF files to measure the performance of preserving statistical properties. Some properties that are expected to be preserved are empirical distributions and standard errors; but other statistics preserved include medians and means. The Office for National

Statistics in the United Kingdom also uses data swapping in which they swap between two and five percent of their census data. In 2001, their swapping results in exchanging the information of over 327,000 individuals in almost 125,000 households (Shlomo et al., 2010). The procedures that they use to swap are similar to those of the USCB.

One could view data swapping to be a renewably-efficient data technique (Willenborg and de Waal, 2001), meaning that it uses all of the values from the original unperturbed data. This is not to say that data swapping is a pollution-efficient disclosure protection technique (Willenborg and de Waal, 2001). Like its alternatives, data swapping introduces data noise that can pose a risk of reduced data utility. Shlomo et al. (2010) suggest to use a rather small percentage of swaps since a "higher swapping rate protects more unique cells", but increase the chances of adding unnecessary noise to the data. They caution that using a high swapping rate can "overly protect" the pre-swapped data when imputation and adjusted parameters were used to clean the data.

In summary, making assessment data available to users through microdata is a flexibility because it satisfies the growing need for data proprietors to share information easily and quickly (Samarati, 2001). In doing so, it is important to protect the individuals that help create the data while preserving the properties collected in the data. Preserving potential item bias, using DIF, is one example. The methodology that will be used in this research study will contribute to the insight of measuring how often it is preserved when applying a commonly used disclosure avoidance technique.

Chapter 3

Methods

3.1 Research Study Methodology

It was hypothesized that the likelihood of preserving test-setting accommodation DIF was conditional on several independent variables such as the percentage of records swapped, the type of item scoring, the ratio between the reference and focal group sizes, and the magnitude of the DIF prior to swapping.

1. *Item Scoring.* The study used dichotomously-scored and polytomously-scored items, and DIF preservation was tested for both cases.

2. *Initial DIF Level.* Three levels of DIF (None, B, and C) were considered for the dichotomously-scored items, and three levels of DIF (None, BB, and CC) were considered for the polytomously-scored items.

3. *Swap Rate.* Three swapping rates (5%, 3%, and 1%) were considered, with the 3% swapping rate representing the baseline privacy protection of the data, 1% represents underprotection, and 5% represents overprotection. This is concordant with the common practice of data swapping described in Chapter 2.5.

4. *Sample Size Ratio.* Four reference-to-focal sample size ratios (10:1, 5:1, 2:1, and 1:1) were considered, with the 1:1 ratio representing the baseline. These ratios were considered since focal groups typically carry small sample sizes resulting in large ratios in comparison to the reference group (discussed in studies described in Chapter 2.2).

The methodology of this study can be summarized into four steps: (1) *simulating pre-swapped data*, (2) *assessing the DIF magnitude of the DIF pre-swapped data*, (3) *performing the data swapping to create the post-swapped*, and (4) *measuring the magnitude of the DIF in the post-swapped data*. The data that have been captured after completing these steps will be used to compute likelihoods of preserving various item DIFs after applying data swapping to protect the data, quantify the association with the several factors that will be used, and develop a generalized linear model to explain the likelihoods.

3.2 Research Study Simulation

A program using the SAS 9.3 statistical programming language was used to perform several simulations containing 10,000 iterations. Each iteration involved creating a fictitious dataset of 50,000 examinees with item responses to twenty dichotomous each worth zero or points

and twenty polytomous items each worth up to four points. The number of examinees and the number of test items reflected that of a typical statewide standardized tests. The ability levels (θ) for each examinee was sampled from a standard normal distribution with mean of zero and variance of one.

Item responses of zero or one were assigned for the twenty dichotomous items using the three-parameter model (Harris, 1989). According to this model, the probability of an examinee receiving full credit for an item given θ is

$$P\Big(Y_j = 1|\theta\Big) = c + \frac{1-c}{1+exp\{-1.7a_j(\theta - b_j)\}}. \tag{3.1}$$

The item difficulty (b_j) and discrimination (a_j) parameters were randomly generated from the standard normal distribution and the lognormal distribution with mean of 0 and variance of 0.2, respectively. The guessing parameter (c) for all of the dichotomous items was constant at 0.2. Each dichotomous item response was determined by comparing the probability of receiving full credit to a number randomly sampled from a uniform distribution between $[0, 1]$. If the probability was greater than or equal to the randomly generated uniform number, full credit was assigned. Otherwise, no credit was assigned.

The generalized partial credit model (Muraki, 1992) was used to generate the item responses for the polytomously-scored items in which integer scores between zero (no credit) and four points (full credit) were given. According to the model, the probability of assigning item response k points given θ is equal to

$$P(Y_j = k|\theta) = \frac{\exp\left\{\sum_{i=0}^{k} a_j(\theta - b_{ji})\right\}}{\sum_{l=0}^{4} \exp\left\{\sum_{i=0}^{l} a_j(\theta - b_{ji})\right\}}, \quad k = 0, 1, 2, 3, 4 \tag{3.2}$$

where $j = 1, 2, ..., 19, 20$. Similar to the dichotomous case, item discrimination parameters were randomly generated from a lognormal distribution with mean of 0 and variance of 0.2. To create the item step difficulties for a polytomous item j, four numbers were randomly sampled from the standard normal distribution $(z_{j0}, z_{j1}, z_{j2}, z_{j3})$ and then arranged in ascending order $(z^*_{j0}, z^*_{j1}, z^*_{j2}, z^*_{j3})$ where $z^*_{j0} = b_{j0},\ z^*_{j1} = b_{j1},\ z^*_{j2} = b_{j2},\ z^*_{j3} = b_{j3}$, and $b_{j0} < b_{j1} < b_{j2} < b_{j3}$. The probabilities from the partial credit model were used as input arguments into a SAS random number generator called *rantbl* to select an item response between zero and four based on these probabilities (Cody, 2010).

One medium- (BB) and one high-level (CC) polytomous DIF item were generated using the constant pattern described by Wang and Su (2004). Under this pattern, each of the item step difficulty parameters for the focal group was increased by $s > 0$ units and used to calculate the probabilities of the polytomous item responses. Thus, for the i^{th} polytomous item, the k^{th} item step difficulty parameter is calculated as

$$b^*_{ik} = b_{ik} + s_p, \quad k = 0, 1, 2, 3, 4.$$

For polytomous items, $s_p = 0.90$ was added to each of the difficulty thresholds to produce the BB-level DIF and $s_p = 1.20$ for the CC-level DIF items. These values were determined

using simulations, and were fairly in range of the constants used by Wang and Su (2004) in their simulation study. For those dichotomous items where s_d represents the additional difficulty, $s_d = 0.127$ was added to create the B-level DIF items and $s_d = 0.175$ for the C-level DIF items. These were created based on simulations that I conducted to maximize the number of iterations that would contain these types of DIF as intended. B-level DIF was detected 92.5% of the time when $s_d = 0.127$ was used and C-level DIF was detected 91.6% of the time when $s_d = 0.175$ was used.

The four DIF items became the *treatment* items for the analysis. Additionally, one dichotomous and one polytomous item of the remaining 36 items acted as the *baseline* items. These items used identical parameters for the reference and focal groups and did not contain statistically significant DIF in them.

A school identification number represented the geography variable in the data. Each examinee was randomly assigned one of 580 unique possible school identification numbers. This variable was needed since data swapping involves exchanging examinee records across schools and not within the same school.

Finally, seven flag variables were randomly generated to identify examinees who were in most need to be swapped. The seven flag variables represented whether an examinee had a unique first language, ethnicity, physical disability, gender, learning disability, or remedial or gifted course schedule. A value of one for a flag variable meant that the examinee was the only one within his or her school with that particular demographic characteristic or was one of two examinees with that characteristic. A value of zero was recorded otherwise. It

is important to note that an examinee could be flagged for none, one, or more than one of the seven demographics. Table A.1 contains a list of all of the variables generated via the simulation.

Several statistics were computed before swapping the data on the baseline and treatment items. The Mantel-Haenszel test statistic, its p-value, and effect size were calculated for the dichotomous items, and the Mantel test statistic, its p-value, and the LACLOR were computed for the polytomous items. These measures were discussed in the previous chapter. Recording these statistics was essential to understanding the status of these items before implementing the data swapping procedure to protect the data. After data swapping, the same measured were collected again in order to observe the effect that data swapping had on these items.

3.3 Data Swapping

Several procedures were used to perform the data swapping on the simulated data. First, a total risk score (TRS) was calculated on the sum of the flag variable values plus one. That is,

$$TRS = f_{clang} + f_{ethnic} + f_{disabil} + f_{cgender} + f_{clapmath} + f_{tasmath} + f_{gifted} + 1.$$

A higher total risk scores increased an examinee's chance of being selected for swapping. A high total risk score for an examinee did not necessarily guarantee that the examinee would be selected for swapping. The total risk score was then multiplied by a number randomly

sampled from the uniform distribution to create a *selection score* that was used to select the examinees for swapping. The examinees were sorted by selection scores and the swapping candidates were chosen based on top selection scores and the swapping rate.

Before the swapping took place, each examinee was paired with another examinee that was also selected for swapping. In doing so, the first requirement was that the paired examinees had different school identification numbers. Afterwards, the examinees were paired based on equivalent gender followed by equivalent ethnicity. All remaining examinees were paired by the TRS. In cases where there was a discrepancy, the selection scores were regenerated for all examinees and the selected examinees were paired again. The decision to pair on gender and ethnicity was based on the purpose of minimizing the likelihood of experiencing a swap failure and the amount of demographic information lost due to the data swapping.

When the data swapping took place, only the school identification number and group variables were interchanged between the paired examinees. Exchanging all of the information would not be an example of data swapping, but rather a reordering of the unswapped data. After swapping the data, the same statistics computed on the baseline and treatment items were computed again to assess how the magnitude of the DIF changed due to the data swapping.

3.4 Details of the Simulation

A total of twenty-four distinct simulations were performed to test the combination of the three swapping rates (1%, 3%, and 5%) and four reference-to-focal size ratios (10:1, 5:1, 2:1, and 1:1) at the two distinct item scorings. The results from the simulations were used to populate several contingency tables for the analysis. The 3% swapping rate and 1:1 reference-to-focal size ratio represented the baseline scenario for a couple of reasons. First, data proprietors often use a three percent swapping rate to protect their data since it represents just enough protection without introducing to much noise in the data (Duncan et al., 2001; Doyle et al., 2001; Zayatz, 2007). Therefore, using a 3% swappng rate as a baseline would allow the reader to understand how larger swapping rates that introduce more noise than needed or lower swapping rates which introduce less noise than needed affect DIF preservation likelihoods. Furthermore, larger group ratios mimic real data (Shealy and Stout, 1993) and baselining the 1:1 ratio allows the reader to gain a better understanding of how these larger group ratios affect DIF preservation.

3.5 Plan of Analysis

3.5.1 Estimation

Several contingency tables were constructed to estimate the rate at which uniform DIF is preserved in dichotomous and polytomous items after applying data swapping protection to

the data. Graphs were also constructed to show emphasis on the likelihoods of preserving the same level of DIF after data swapping for the combinations of the swapping rates and reference-to-focal group ratios. Because the likelihoods were relatively close in value, the likelihoods were mapped using the log of the likelihood's odds, or *logits* so that the reader could get a better visualization. Finally, additonal contingency tables were constructed to show the mean effect size for DIF levels after data swapping conditional on the level of DIF before data swapping. The purpose of these tables was to show possible relationships between the behavior of the mean effect size and the independent variables of the study.

3.5.2 Association

A test of association and a correlation coefficient were used to determine whether association existed and to measure the association between the data swapping rate, item scoring, reference-to-focal group ratio, and the severity of DIF before and after data swapping. The Cochran-Mantel-Haenszel (CMH) correlation test (Cochran, 1954) using standardized midranks was considered because of the ordinal nature of the variables. This test is similar to that of the Cochran-Armitage Test of Trend (Armitage, 1955) and carries a test statistic calculated as

$$Q_{CS} = \frac{(n-1)[\sum_{i=1}^{3}\sum_{j=1}^{4}(c_i - \mu_{\mathbf{c}})(a_j - \mu_{\mathbf{a}})n_{ij})]^2}{[\sum_{i=1}^{3}(c_i - \mu_{\mathbf{c}})^2 n_{i+}][\sum_{j=1}^{4}(a_i - \mu_{\mathbf{a}})^2 n_{+j}]},$$

where $\mathbf{c} = (c_1, c_2, c_3)$ represents the standardized rank scores for the pre-swap DIF levels

and $\mathbf{a} = (a_1, a_2, a_3, a_4)$ are the standardized midrank scores for the post-swapped DIF levels. For information on how the standardized midranks are calculated, see Stokes et al. (2003). Standardized midrank scoring was considered over regular integer scoring since its does not require the user to create any score system for qualitative responses. Note that Q_{CS} follows a Chi-Squared distribution with one degree of freedom, and extremely large values for Q_{CS} are evidence to conclude that such a correlation exists.

Somer's D (Somers, 1962) was used to compliment the Mantel-Haenszel Test by quantifying the association between pre- and post-swapped DIF levels. This statistic is calculated as

$$D_{C|R} = \frac{P - Q}{w_r},$$

where $w_r = n^2 - \sum_{i=1}^{3} n_{i+}^2$, P represents the number of pairs in which the pre-swapped DIF level was higher than the post-swapped DIF level, and Q represents the number of pairs in which the pre-swapped DIF level was lower than the post-swapped DIF level. Somers' D is a measure of asymmetric behavior, and so this coefficient ignores ties on the independent variable (Goodman and Kruskal, 1954). Several other nonparametric correlation coefficients like Kendall's Tau and the Gamma coefficient could have been used to measure the association. Somer's D was preferred over them since it is the least conservative, takes into consideration the ordinal nature of the variables, and it adjusts for ties between the row and column variables.

Confidence intervals, with 95% confidence levels, were then applied to account for the

sampling error associated with this statistic. The confidence intervals were computed as

$$CI_{95\%} = r_D \pm 1.96 \times ASE_{r_D},$$

where r_D represents the value of Somer's D and the ASE_{r_D} is the asymptotic standard error of r_D. Confidence intervals not containing the value of zero show significant evidence of an association between the aforementioned variables.

3.5.3 Modeling

The cumulative logit model is one of several generalized linear models (GLM) that can be used to predict the likelihood of an item containing a level of DIF as a function of the pre-swap DIF level, the swapping rate, and the reference-to-focal group ratio. This model was preferred over the loglinear GLM for two reasons. First, the loglinear GLM is commonly used to predict frequencies rather than likelihoods. Second, the loglinear GLM works best with frequency tables with equal row totals. When the row totals are different, an offset parameter is used in the model as an adjustment factor. Interpreting the offset parameter to the research at hand would be both difficult and irrelevant. The cumulative logit model in its multivariate form is

$$logit[P(Y \leq j)] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x},$$

where

$$logit[P(Y \leq j)] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right], \quad j = 1, 2, 3, 4.$$

The dependent variable represents the level of DIF preserved after data swapping. For the dichotomous case, $Y = 1$ represents A_1-level DIF preserved after data swapping, $Y = 2$ represents A_2-level DIF, $Y = 3$ for B-level DIF, and $Y = 4$ represents C-level DIF. Similarly, AA_1, AA_2, BB, and CC would be the analogous levels for the polytomous case respectively. Because there are two sets of post-swapping DIF levels based on two separate scales, separate cumulative logit models were made.

Four dichotomous and four polytomous cumulative logit models were fitted to measure the effects of the explanatory variables on DIF preservation likelihoods. Models 1 through 3 are main effects models in which the DL variable, the DL and SR variables, and the DL, SR, and RF variables were used respectively. The results from these models are provided in Tables 4.4 and 4.7. Model 4 is the full model containing all main and interaction effects. Its results are found in Tables 4.5 and 4.8. Each of these tables contains the regression model coefficients, the Akaike Information Criterion (AIC), the c-statistic, and the correct classification rate (CCR). The AIC (Akaike, 1974) was considered as the measure to compare models and is calculated as

$$AIC = 2k - 2\ln(L)$$

where L represents the value that maximizes the likelihood function for the cumulative logit model. Models with the lowest AIC are siad to be the best parsimonious model (Sakamoto et al., 1986). The c-statistic was used to measure model predictive power, and values higher than 0.7 suggest that the model's ability to predict is not by mere random chance (Hosmer and Lameshow, 2000). The CCR is equal to the proportion of all simulations where the model correctly predicted the level of item DIF after data swapping.

Chapter 4

Results

This research study attempted to estimate the rate at which uniform DIF is preserved in dichotomous and polytomous items after applying data swapping protection to the data, determine whether an association exists between the likelihood of preserving uniform DIF, the data swapping rate (SR), the item scoring (IS), the reference-to-focal group ratio (RF), and the severity of the DIF originally detected (DL), and determine whether the likelihood of preserving uniform DIF can be explained using a generalized linear model. The results were organized by the three research questions.

4.1 DIF Preservation Rate Estimation

Tables B.1 to C.12 describe the outcomes of the simulations before (BDS) and after data swapping (ADS). The BDS frequencies show how many of the 10,000 iterations successfully

created the intended items, whereas the ADS frequencies show the change in the item DIF after swapping. It is important to note that the sum of the row frequencies in the ADS cells always match the frequency to which the intended DIF level and the BDS DIF level were concordant. Consider the frequencies for B-level DIF found in Table B.1, for example. The intention was to simulate 10,000 iterations in which one B-Level DIF dichotomous item was created. However, this only happened in 7,031 of the 10,000 iterations. The frequencies for the ADS cells for that row sum to 7,031, which was not the frequency in which B-Level DIF was intended, but rather the frequency where B-Level DIF was actually created.

It is important to note that since A-level DIF is interpreted as "little to no DIF", the dichotomous tables in Appendix B contain two A-level DIF columns. The "A_2" level represented those simulations in which the data swapping produced a significant Mantel-Haenszel statistic to conclude that significant differences in group performances existed between the reference and focal groups, but the effect size difference was little. The "A_1" level represented those simulations in which the Mantel-Haenszel statistic was not found to be significant and the effect size was neglgible, hence the "no DIF". Both columns represent cases of A-level DIF, but their separation was meant to show the impact that data swapping has on creating noise such that a dichotomous item containing no-DIF data can contain little DIF after swapping. Similarly, the "AA_1" and "AA_2" levels in the polytomous tables show the distinction between non-significant versus significant Mantel test statistics although the effect sizes for both cases are small.

The twenty-four tables in Appendices B and C provide results for all combinations of the

RF and SR for the dichotomous and polytomous items being analyzed. Collectively, Tables B.1 to B.12 show that the intended levels of DIF were successfully created 92.11% of the time, whereas Tables C.1 to C.12 show that the intended levels of DIF were successfully created 90.33% of the time.

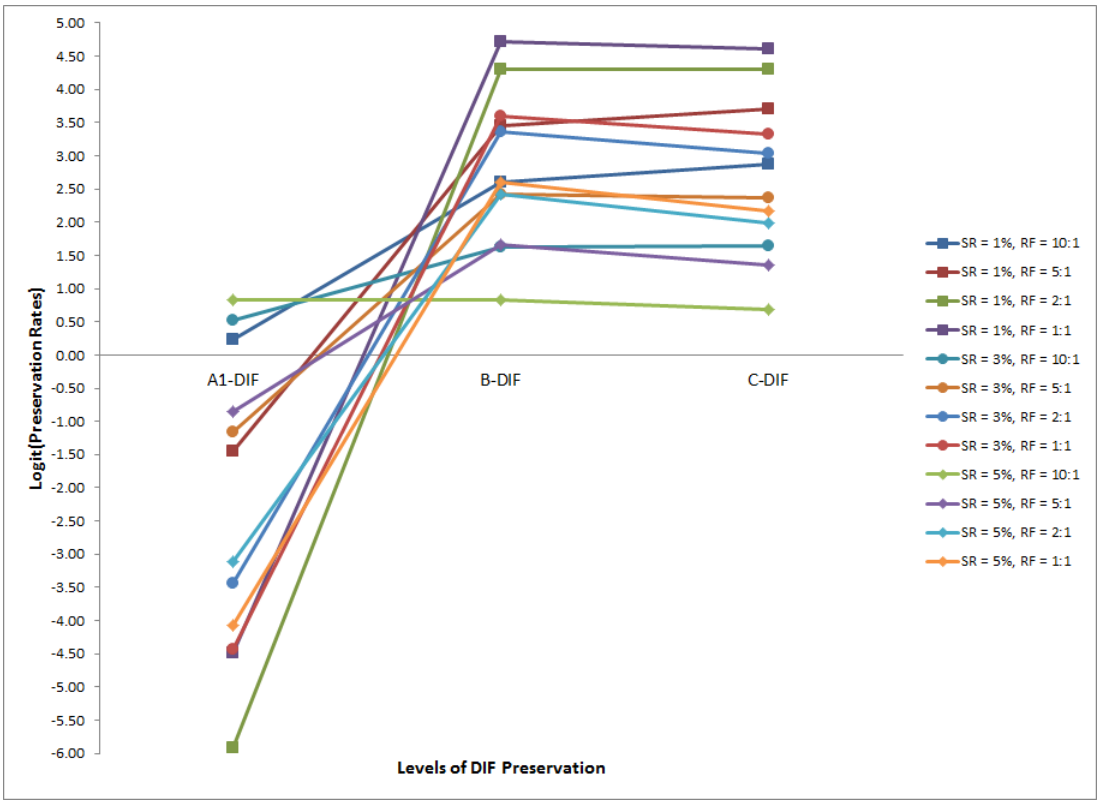

Figure 4.1: Comparison of DIF Preservation Rates-Dichotomous

Figure 4.1 provides an illustration of the dichotomous preservation likelihoods for the combinations of the SR and RF after data swapping. Logits, which are the log of the odds, were graphed on the vertical axis for the reader to get a better visualization of the likelihoods. The 3% SR and 1:1 RF combination represents the baseline scenario. For preserving A_1-DIF, two other scenarios were found to yield lower DIF preservation likelihoods than the baseline.

One scenario, where the SR was 1% and the RF was 1:1, had a logit of -4.490 (equivalent to a proportion of 1.11%). The propotion was calculated by dividing the 111 simulations resulting in a A_1-DIF level after data swapping by the 10,000 simulations in which A_1-DIF was created before data swapping. The other scenario involved a 1% SR and a 2:1 RF for which the logit was -5.912 (proportion equal to 0.27%). The highest three scenarios had preservation likelihoods that were at least 4.5 logits away from the baseline; namely where SR was 5% and RF was 10:1 (proportion equal to 69.67%), SR was 3% and RF was 10:1 (proportion equal to 62.58%), and SR was 1% and RF was 10:1 (proportion equal to 55.84%). A list of the likelihoods for the dichotomous case can be found in Table D.1 of the Appendix.

Preservation rates for B-level DIF were between 0.828 and 4.713 logits, which are much higher and less variable than the A_1-DIF preservation rates. Yet, several scenarios were found to be lower than baseline scenario. The three lowest, in particular, were where the SR was 5% and the RF was 10:1 (logit equal to 0.828), SR was 3% and the RF was 10:1 (logit equal to 1.627), and SR was 5% and the RF was 5:1 (logit equal to 1.667). Only two scenarios were found to have higher B-level preservation rates: SR of 1% and the RF of 1:1 (logit equal to 4.713) and SR was 1% and the RF was 2:1 (logit equal to 4.291). It is interesting to note that not only were the scenarios that were greatest now were among the poorest in preserving no-DIF, but also the scenario that was the poorest now was the best in preserving no-DIF.

C-level preservation rates were between 0.682 and 4.605 logits, which were much higher than the A_1-DIF preservation rates and slightly lower than the B-level preservation rates.

The baseline scenario had a logit of 3.320, which was equivalent to a preservation rate of 96.51%. Similar to the behavior found with B-level DIF preservation, the SR of 1% and the RF of 1:1 and SR was 1% and the RF was 2:1 were found to yield the highest preservation rates. Also, the scenario where the SR was 5% and the RF was 10:1 was still the worst in preserving C-level DIF after data swapping. Overall, Figure 4.1 show that it is much more difficult to preserve dichotomous items containing no DIF than items with some level of DIF before data swapping. One explanation is that data swapping adds noise to the data which increases the variability in the item scoring. As a result, this further distorts the expectation that examinees with higher ability levels leads to a higher chance of receiving a correct score than examinees with lower ability levels.

Figure 4.2 provides an illustration of the polytomous preservation likelihoods after data swapping, using logits to show a better visualization of the likelihoods. Preservation likelihoods were between 2.893 (94.75%) and 4.191 (98.51%) logits at the AA_1-DIF, between -0.299 (42.57%) and 6.213 (99.80%) logits for BB-level DIF, and between -0.918 (28.54%) and 8.517 (99.98%) logits for CC-level DIF. A list of the likelihoods can be found in Table D.3 of the Appendix. Compared to Figure 4.1, the behavior of the preservation likelihoods for the polytomous case is completely different. The preservation rates for AA_1-DIF are now higher and contain less variability. The preservation likelihoods for the BB- and CC-level DIF are more spread apart than they were for the dichotomous case. Additionally, these likelihoods are at several ranges of values where clusters of scenarios can be identified.

For preserving AA_1 DIF, the baseline had a logit of 3.112 (which is comparable to a

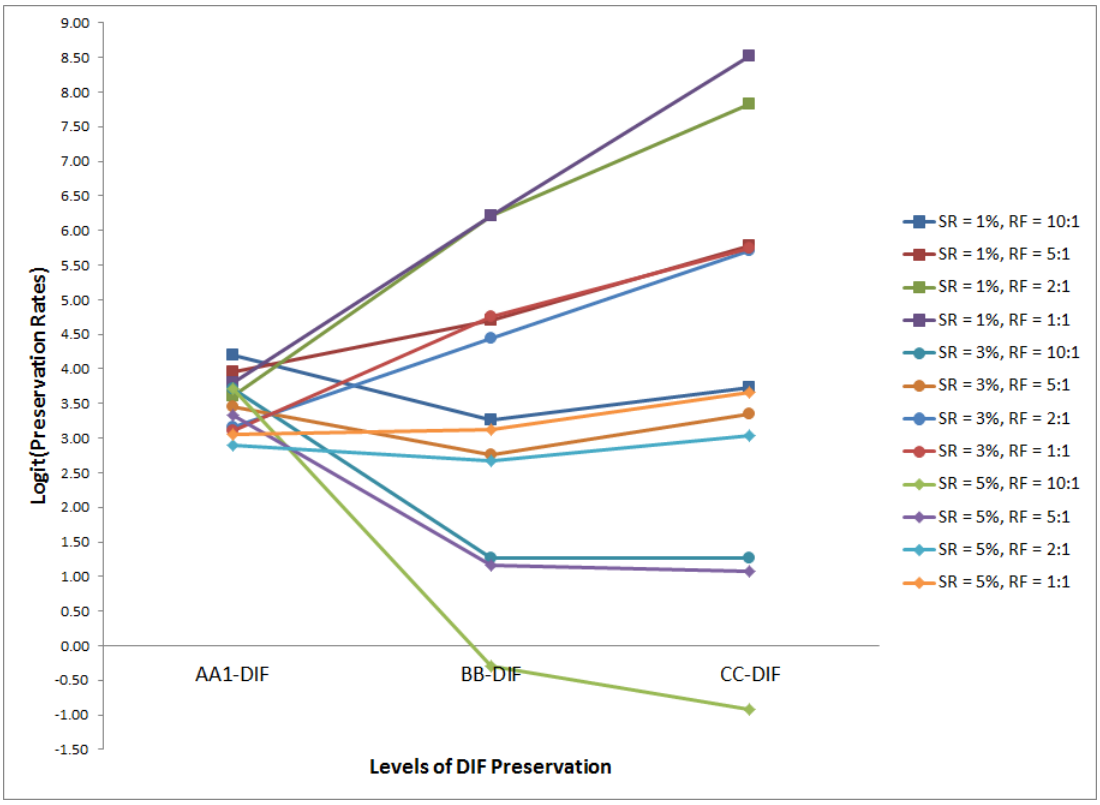

Figure 4.2: Comparison of DIF Preservation Rates-Polytomous

likelihood of 95.74%. The scenarios in which a 5% SR and a 2:1 RF were used (logit equal to 2.893) plus where a 5% SR and a 1:1 RF were used (logit equal to 3.046) were the only two scenarios that produced smaller preservation likelihoods. The scenarios with the largest preservation likelihoods both involved using a swapping rate of 1%, but different level for RF (10:1 had a logit of 4.191 and 5:1 had a logit of 3.960).

Although the variability in the preservation likelihoods for CC-level DIF was larger than the BB-level, their conclusions regarding which scenarios were best and worst were similar. Figure 4.2 clearly identified that using a 5% SR on data containing a 10:1 ratio yields the lowest DIF preservation rates for a polytomous item (BB-level logit of -0.299 and CC-

level logit of -0.918). In fact, this was the only scenario that produced BB- and CC-level preservation rates below 50%. The scenario in which SR was 1% and RF of 2:1 tied with with the scenario in which 1% and RF of 1:1 for the largest BB-level DIF preservation rate (logit equal to 6.213). For CC-level DIF preservation, the scenario in which 1% and RF of 1:1 was more effective than 1%-2:1 scenario (logit of 8.517 versus 7.824).

Table 4.1: DIF Preservation Decreases between Swapping Rates

		Dichotomous		**Polytomous**	
DIF Level	**R:F Ratio**	1% to 3%	1% to 5%	1% to 3%	1% to 5%
B (dichotomous)/ BB (polytomous)	1:1	1.76%	4.24%	0.65%	0.30%
	2:1	2.02%	4.80%	0.97%	0.29%
	5:1	5.04%	7.75%	5.03%	3.08%
	10:1	9.48%	13.98%	18.42%	19.63%
C (dichotomous)/ CC (polytomous)	1:1	2.50%	6.79%	3.39%	2.20%
	2:1	3.23%	7.56%	5.28%	4.26%
	5:1	6.22%	11.80%	17.82%	22.19%
	10:1	10.69%	17.49%	35.32%	49.47%

The results from Figures 4.1 and 4.2 suggest to data proprietors to use a data swapping rate of 1% to produce the highest DIF preservation rates when a dichotomous item contains B- or C-level DIF. For some data proprietors, the decision of the swapping rate to use is mandated by law. For such situation, the purpose of Table 4.1 is to show the decrease in the preservation likelihoods when using a swapping rate other than 1% so that the reader can consider the risk of using a swapping rate higher than 1%.

The mean effect size was also used as a measure to explain the item DIF transition due to data swapping. Tables E.1 to E.12, using Equation (2.2), explained the phenomenon for the dichotomous items. It is important to emphasize that these tables are row-conditional

tables, meaning that the columns reflect the DIF outcomes after data swapping conditional on the DIF created before swapping reflected in the rows. For example, the value of 1.56 in Table E.1 reflect the mean effect size of those iterations in which A_2-level DIF was created after swapping given that B-level DIF was created before data swapping. Tables F.1 to F.12 were created and interpreted in the same manner where Equation (2.4) was used. These tables suggest that a trend exists between data swapping and effect size, namely effect size decreases when items transition from A_1 and AA_1 DIF to higher levels of DIF after data swapping and effect size increases when items transition from some level of DIF to an even higher level after data swapping.

Estimates from these tables show that the effect size ranges between 0.74 and 0.86 when A_1-DIF is preserved after data swapping, between 1.69 and 1.73 when B-level DIF is preserved, and 2.09 and 2.24 when C-level DIF is preserved. For the polytomous items, data swapping has little impact on the effect size since the ranges were found to be between 0.98 and 0.99. Effect size ranges for BB-level and CC-level DIF were 1.20 to 1.21 and 1.30 to 1.32 respectively. These ranges contain much less variability and show a less severe impact on the focal and reference groups than what happens for dichotomous items. Overall, the ranges for the A_1 and AA_1 DIF levels suggest that data swapping shifts the effect size into levels that favor the focal group for dichotomous items more than polytomous items.

4.2 Tests for Association

Frequencies from the 'After Data Swapping' columns of Tables B.1 through C.9 were used to determine whether an association existed between the preservation likelihoods and the level of DIF before data swapping at all combinations of the swapping rates and reference-to-focal group ratios.

Table 4.2: Cochran-Mantel-Haenszel Correlation Test Values

		Item Scoring	
RF	SR	Dichotomous	Polytomous
1:1	1%	27092.65***	25786.18***
1:1*	3%*	26704.27***	25662.43***
1:1	5%	25575.83***	25008.02***
2:1	1%	26543.01***	26220.97***
2:1	3%	26063.98***	26087.75***
2:1	5%	24782.33***	25109.40***
5:1	1%	23909.75***	27784.55***
5:1	3%	22828.11***	26889.00***
5:1	5%	21319.42***	23753.47***
10:1	1%	20996.31***	27946.20***
10:1	3%	19898.00***	25133.86***
10:1	5%	18821.35***	23381.38***

$p_v < 5\%$,$^{**}p_v < 1\%$, $^{***}p_v < 0.1\%$

Table 4.2 contains the results of the Mantel-Haenszel Test. Strong associations were present for all combinations of swapping rate and reference-to-focal group ratios. When aggregating over all swapping rates and reference-to-focal ratios, strong associations were still present for the dichotomous ($Q_{CS} = 278,104, p_v =< .0001$) and polytomous ($Q_{CS} = 298,478, p_v =< .0001$) case.

Somer's D was used to quantify the association between pre- and post-swapped DIF levels

in Tables B.1 through C.9. Combining frequencies over all swapping rates and reference-to-focal ratios, the Somers' D coefficient and asymptotic standard error for the dichotomous case were 0.9057 and 0.0005 respectively. This resulted in a 95% confidence interval of (0.9047,0.9067), showing significant evidence that a correlation exists between initial DIF level and the level of DIF after data swapping. Similar conclusions resulted for the polytomous case in which the overall Somer's D coefficient and asymptotic standard error estimates were 0.9589 and 0.0003, resulting in a 95% confidence interval of (0.9583,0.9595).

Table 4.3: Somer's D Tests for Association

Swapping	**R:F**	**Dichotomous**		**Polytomous**	
Rate	**Ratio**	$D_{C\|R}$	ASE_D	$D_{C\|R}$	ASE_D
1%	1:1	0.9556	0.0012	0.9999	<0.0001
1%	2:1	0.9502	0.0013	0.9998	<0.0001
1%	5:1	0.9256	0.0016	0.9988	0.0002
1%	10:1	0.8969	0.0020	0.9917	0.0005
3%	1:1	0.9459	0.0014	0.9984	0.0003
3%	2:1	0.9405	0.0014	0.9984	0.0003
3%	5:1	0.9040	0.0018	0.9871	0.0007
3%	10:1	0.8738	0.0022	0.9374	0.0013
5%	1:1	0.9201	0.0016	0.9883	0.0007
5%	2:1	0.9103	0.0017	0.9802	0.0009
5%	5:1	0.8695	0.0020	0.9223	0.0015
5%	10:1	0.8487	0.0023	0.8876	0.0017

Table 4.3 provides the estimates and standard errors for Somers' D when considering the swapping rate, reference-to-focal ratio, and item scoring. Correlations for the dichotomous case were between 0.8487 and 0.9556. These correlations were much smaller in magnitude and contained more variability than the correlations for the polytomous case. For the polytomous case, correlations were between 0.8876 and 0.9999. Conditional on item scoring, the highest

correlations between pre- and post-swapped DIF levels were found where a 1% swapping rate and a 1:1 ratio was used. The lowest correlations were also located where a 5% SR and a 10:1 RF was used. Nevertheless, all twelve combinations for swapping rate and reference-to-focal ratio revealed a statistically significant association.

4.3 Generalized Linear Model Fitting

Two cumulative logit models were produced to explain DIF likelihood preservation in dichotomous and polytomous items separately. The results follow for each model:

4.3.1 Dichotomous Model Results

Table 4.4 shows the results of the cumulative logit model for dichotomous items. Model 1 is the main effects using the DL as a predictor. Including this variable in the model reduced the AIC from 851,749.18 to 342,325.35. According to this model, the DL was found to have a significant negative effect on DIF preservation likelihoods when a dichotomous item contains B- and C-level DIF before data swapping. A significant positive effect reflects a statistically significant increase in the logit of preserving DIF levels, while a significant negative effect reflects a statistically significant decrease in the logit. B-level dichotomous items have a 0.012 increase on the odds (or multiplicative effect (Agresti, 2002)) of post-swapped DIF preservation when compared to dichotomous items containing A_1-level DIF before swapping. The multiplicative effect is even smaller for items containing pre-swapped

Table 4.4: Main Effects Cumulative Logit Models: Dichotomous

	Model 1		**Model 2**		**Model 3**	
Estimates	β	se_β	β	se_β	β	se_β
Intercept (α)						
α_{A_1}	-1.24***		-1.32***		-2.11***	
α_{A_2}	1.97***		1.97***		1.53***	
α_B	9.94***		10.09***		10.12***	
Pre-Swap DIF Level (DL)						
DL_B	-4.45***	0.01	-4.56***	0.015	-5.02***	0.05
DL_C	-12.17***	0.05	-12.43***	0.05	-13.41***	0.05
Swapping Rate (SR)						
SR_1			-0.42***	0.01	-0.43***	0.01
SR_5			0.55***	0.01	0.59***	0.01
Reference-Focal Ratio (RF)						
RF_2					0.08***	0.01
RF_5					0.80***	0.01
RF_{10}					1.98***	0.01
AIC	342,325.35		335,234.63		308,697.15	
c	0.926		0.940		0.956	
CCR	81.58%		81.58%		83.14%	

*$p < 5\%$,**$p < 1\%$, ***$p < 0.1\%$

C-level DIF. The c-statistic for Model 1 was 0.926, which represents strong predictive power. Model 1 was found to predict post-swapped DIF levels correctly 81.58% of the time, with the best prediction performances occurring when the SR was 1% and the RF was 1:1 (95.10%) and when the SR was 1% and the RF was 2:1 (94.81%).

Model 2 includes the DL and SR as predictors to model DIF preservation likelihoods. The inclusion of the SR decreased the AIC from 342,325.35 to 335,234.63. A significant negative effect was still present with the DL. DIF preservation rates are negatively affected when dichotomous items are swapped at the 1% SR than at the 3% SR, but are positively affected

when the 5% SR is used. Dichotomous items containing B-level DIF have a 0.01 multiplicative effect on DIF preservation odds and, similar to Model 1, C-level DIF dichotomous items have an even smaller multiplicative effect. Swapping at a 1% rate results in a multiplicative effect between 0.642 and 0.773, but swapping at 5% results in a multiplicative effect between 1.70 and 1.77 with 95% confidence. The c-statistic for Model 1 was found to be 0.940, which was stronger than the c-statistic from Model 1. However, the predictive power rate and the combinations for the SR and RF where the model predicted best remained equal to that of Model 1.

Model 3 experienced a decreased in the AIC compared to Model 2 (335,234.63 versus 308,697.15). Similar effects were found in the DL and the SR after including the RF main effect. The RF was found to have a positive effect on DIF preservation likelihoods at all levels when compared to the 1:1 RF. Consistency was found in the positive effects, namely the 10:1 and 2:1 ratio had the largest and smallest positive effect on DIF preservation likelihoods respectively. The DL multiplicative effect at the B- and C-levels were 0.007 and 0.001 and the multiplicative effects due to the swapping at 1% and 5% were 0.649 and 1.807 respectively. Swapping data containing an RF of 10:1 results in a 7.266 multiplicative effect, while swapping data containing a 5:1 or 2:1 RF results in a multiplicative effect of 2.235 and 1.081 respectively. Model 3 had a higher c-statistic (0.953) than Models 1 and 2, and a slightly higher correct classification rate (83.14%). The best prediction performances occurred at the same locations as those from the previous two models.

Model 4 contains the full-interaction model, containing main and interaction effects to

Table 4.5: Interaction Effect Cumulative Logit Model: Dichotomous

	Model 4			Model 4			Model 4	
Estimates	β	se_β	**Estimates**	β	se_β	**Estimates**	β	se_β
Intercept (α)			2-Way Interactions			3-Way Interactions		
α_{A_1}	-2.18***		$DL_B \times SR_1$	-0.53***	0.09	$DL_B \times SR_1 \times RF_2$	0.12	0.12
α_{A_2}	1.51***		$DL_B \times SR_5$	0.73***	0.08	$DL_B \times SR_1 \times RF_5$	-0.15	0.12
α_B	10.14***		$DL_C \times SR_1$	-1.23***	0.12	$DL_B \times SR_1 \times RF_{10}$	-0.22	0.11
Pre-Swap DIF Level (DL)			$DL_C \times SR_5$	1.03***	0.07	$DL_B \times SR_5 \times RF_2$	0.02	0.10
DL_B	-4.94***	0.06	$DL_B \times RF_2$	0.06	0.08	$DL_B \times SR_5 \times RF_5$	-0.32***	0.09
DL_C	-13.46***	0.08	$DL_B \times RF_5$	0.12	0.07	$DL_B \times SR_5 \times RF_{10}$	-0.27**	0.09
Swapping Rate (SR)			$DL_B \times RF_{10}$	-0.78***	0.07	$DL_C \times SR_1 \times RF_2$	0.15	0.16
SR_1	-0.06	0.03	$DL_C \times RF_2$	0.21**	0.08	$DL_C \times SR_1 \times RF_5$	0.25	0.14
SR_5	0.13***	0.03	$DL_C \times RF_5$	0.19**	0.07	$DL_C \times SR_1 \times RF_{10}$	0.44**	0.13
Reference-Focal Ratio (RF)			$DL_C \times RF_{10}$	-0.74***	0.07	$DL_C \times SR_5 \times RF_2$	-0.11	0.09
RF_2	0.07*	0.03	$SR_1 \times RF_2$	-0.13**	0.04	$DL_C \times SR_5 \times RF_5$	-0.44***	0.09
RF_5	0.77***	0.03	$SR_1 \times RF_5$	-0.31***	0.04	$DL_C \times SR_5 \times RF_{10}$	-0.50***	0.08
RF_{10}	2.41***	0.03	$SR_1 \times RF_{10}$	-0.37***	0.04			
			$SR_5 \times RF_2$	0.01	0.04			
AIC	303,153.14		$SR_5 \times RF_5$	0.29***	0.04			
c	0.958		$SR_5 \times RF_{10}$	0.32***	0.04			
CCR	84.41%							

*$p < 5\%$,**$p < 1\%$, ***$p < 0.1\%$

explain the behavior of DIF preservation likelihoods. There was a slight decrease in the AIC compared to Model 3 (from 308,697.15 to 303,153.14). The behaviors of the main effect estimates for the DL and RF variables in Model 3 were present in Model 4. DIF preservation likelihoods were not negatively affected statistically when using a 1% SR over a 5% SR. This therefore suggests that, with all factors constant, underprotecting the data does not significantly affect DIF preservation likelihoods. Patterns were also found in the two-way and three-way interaction effects. Significant negative effects were present in the DL and the SR factors, particularly when dichotomous items contain B- or C-level DIF in data are swapped at a 1% SR. The DIF preservation likelihoods at these conditions were found to be significantly lower than dichotomous items containing A_1-level DIF and swapped at 3%. When a 5% SR is used, a significant positive effect in DIF preservation likelihoods was found. When considering the interaction between the DL and the RF, DIF preservation likelihoods were negatively affected when B- or C-level dichotomous items are swapped with data containing a 10:1 RF. It is only positively affected when a dichotomous C-level DIF item is swapped with data containing 2:1 or 5:1 RFs. Several conditions between the interaction of the SR and RF variables also affect DIF preservation likelihoods. Negative effects were present with data that was swapped at 1% and contained 2:1, 5:1, or 10:1 ratios, and the likelihoods at these combinations were lower than at the baseline of a 3% SR with a 1:1 RF. Significant positive effects were present when data is swapped at the 5% SR but contains a RF of 5:1 or 10:1. With the interaction of all three factors, there were four combinations in which statistically negative effects were present. According to the model, negative effects

were present in dichotomous B- and C-level items created by 5:1 and 10:1 RFs swapped at a 5% SR. This means that the DIF preservation likelihoods are smaller at these conditions than at the baseline treatment. A positive effect was found only where a C-level DIF dichotomous item comes from data with a 10:1 RF and swapped at 1% SR. Model 4 produced a c-statistic of 0.990 and a correct classification rate of 93.08%, which demonstrates the strong predictive power amongst the four models.

Table 4.6 contains the predicted probabilities of preserving the various DIF levels after data swapping as a function of the SR, RF, and DL. The predicted probabilities are row conditional since the cumulative logit model is a conditional model on all of its predictors. The highest probabilities on each row were bolded for the reader to better see the concordances and discordances between pre- and post swap DIF levels. For example, 71.38% represents the highest probability among the four post-swap DIF levels under the condition that the data has a 1% SR, a 1:1 RF, and the dichotomous item contained A_1-level DIF before data swapping. This represents a discordance since the dichotomous item initially contained A_1-level DIF before data swapping. However, the bolded 97.26% and 99.01% on the second and third rows represent concordances since they represent the highest likelihoods where the pre- and post-swapped DIF are the same. Only two combinations for the SR and RF were found to have complete concordances at all three pre-swapping DIF levels (SR= 3%, RF= $10:1$ and SR= 5%, RF= $10:1$). All of the other combinations for the SR and FR produced discordant results when A_1-level DIF was created before data swapping and concordant results when B- and C-level DIF was created before data swapping. These strong findings suggest

Table 4.6: Predicted DIF Preservation Likelihoods: Dichotomous

			DIF After Swapping			
SR	**RF**	**DIF Before Swapping**	A_1	A_2	B	C
1%	1:1	A_1	9.65%	**71.38%**	18.97%	<0.01%
1%	1:1	B	0.04%	1.72%	**97.26%**	0.98%
1%	1:1	C	<0.01%	<0.01%	0.99%	**99.01%**
1%	2:1	A_1	9.14%	**70.95%**	19.91%	<0.01%
1%	2:1	B	0.05%	1.94%	**97.14%**	0.87%
1%	2:1	C	<0.01%	<0.01%	1.33%	**98.67%**
1%	5:1	A_1	14.49%	**72.65%**	12.86%	<0.01%
1%	5:1	B	0.07%	2.61%	**96.68%**	0.64%
1%	5:1	C	<0.01%	<0.01%	2.41%	**97.59%**
1%	10:1	A_1	45.21%	**51.85%**	2.94%	<0.01%
1%	10:1	B	0.13%	4.72%	**94.81%**	0.35%
1%	10:1	C	<0.01%	<0.01%	5.40%	**94.60%**
3%	1:1	A_1	10.17%	**71.74%**	18.08%	<0.01%
3%	1:1	B	0.08%	3.05%	**96.32%**	0.55%
3%	1:1	C	<0.01%	<0.01%	3.49%	**96.51%**
3%	2:1	A_1	10.83%	**72.10%**	17.07%	<0.01%
3%	2:1	B	0.09%	3.48%	**95.95%**	0.48%
3%	2:1	C	<0.01%	<0.01%	4.56%	**95.44%**
3%	5:1	A_1	19.69%	**71.06%**	9.26%	<0.01%
3%	5:1	B	0.20%	7.14%	**92.44%**	0.22%
3%	5:1	C	<0.01%	<0.01%	8.63%	**91.37%**
3%	10:1	A_1	**55.77%**	42.28%	1.94%	<0.01%
3%	10:1	B	0.41%	13.78%	**85.70%**	0.11%
3%	10:1	C	<0.01%	<0.01%	16.08%	**83.91%**
5%	1:1	A_1	11.38%	**72.32%**	16.29%	<0.01%
5%	1:1	B	0.19%	6.88%	**92.69%**	0.23%
5%	1:1	C	<0.01%	<0.01%	10.27%	**89.72%**
5%	2:1	A_1	12.25%	**72.56%**	15.18%	<0.01%
5%	2:1	B	0.23%	8.07%	**91.51%**	0.20%
5%	2:1	C	<0.01%	<0.01%	12.12%	**87.88%**
5%	5:1	A_1	27.01%	**66.66%**	6.33%	<0.01%
5%	5:1	B	0.45%	14.83%	**84.62%**	0.10%
5%	5:1	C	<0.01%	<0.01%	20.42%	**79.57%**
5%	10:1	A_1	**66.35%**	32.40%	1.25%	<0.01%
5%	10:1	B	1.01%	28.06%	**70.89%**	0.04%
5%	10:1	C	<0.01%	0.01%	33.58%	**66.41%**

that dichotomous items are able to maintain the same level of DIF before and after data swapping when DIF is contained in the item, and that DIF can be created in an item after data swapping when DIF is not present beforehand.

4.3.2 Polytomous Model Results

Table 4.7 shows the estimates for main effects cumulative logit models. Model 1 contains the effect of the DL on DIF preservation in polytomous items. Including this predictor in the model reduced the AIC from 796,958.61 to 185,996.58. Although this appears to be a significant drop in the AIC, it was found that polytomous items with BB- or CC-level DIF do not significantly decrease DIF preservation likelihoods compared to polytomous items containing AA_1-level DIF. One possible explanation is that the distribution in the number of simulations preserving the same level of DIF versus different DIF before and after data swapping were similar across Tables C.1 through C.12. The c-statistic for Model 1 was 0.960, which represents a strong predictive power. Additionally, Model 1 was found to predict post-swapped DIF levels correctly 91.33% of the time. Unlike the predictive power from Model 1 for the dichotomous case, this model predicted well at several combinations of the SR and RF. Model 1 had the weakest prediction when the SR was 5% and the RF was 10:1 (54.86%).

Including the SR into the model lowered the AIC from Model 1 (185,996.58 versus 164,733.68). Levels for the DL continued to be not significant, suggesting that the DIF preservation likelihoods do not depend on the level of DIF before data swapping. A strong

Table 4.7: Main Effects Cumulative Logit Models: Polytomous

	Model 1		**Model 2**		**Model 3**	
Estimates	β	se_β	β	se_β	β	se_β
Intercept (α)						
α_{AA_1}	3.47***		3.74***		2.92***	
α_{AA_2}	21.27		21.51		21.77	
α_{BB}	43.45		44.52		46.97	
Pre-Swap DIF Level (DL)						
DL_{BB}	-23.42	56.92	-24.19	53.81	-26.45	79.26
DL_{CC}	-45.54	86.07	-47.14	84.30	-51.58	126.30
Swapping Rate (SR)						
SR_1			-1.02***	0.02	-1.07***	0.02
SR_5			1.45***	0.02	1.77***	0.02
Reference-Focal Ratio (RF)						
RF_2					0.18***	0.03
RF_5					1.61***	0.03
RF_{10}					3.25***	0.03
AIC	185,996.58		164,733.68		132,073.15	
c	0.960		0.979		0.989	
CCR	91.33%		91.33%		93.08%	

*$p < 5\%$,**$p < 1\%$, ***$p < 0.1\%$

negative effect on DIF preservation likelihoods was found when data is swapped at an SR of 1% versus 3%, while a strong positive effect was found when data is swapped at an SR of 5% versus 3%. Although Model 2 had a slightly larger c-statistic than Model 1 (0.979 versus 0.960), Model 2 yielded the same correct classification rate as Model 1.

Model 3 used the main effects of the DL, SR, and RF to predict DIF preservation likelihoods. The AIC for Model 3 was smaller than the AIC from Model 2 (132,073.15 versus 164,733.68), suggesting that more variability in DIF preservation was explained by including the RF into the model. The DL was found to not affect DIF preservation likelihoods,

which was similar behavior found in the previous models. The SR continued to be a strong predictor for explaining DIF preservation likelihoods, with the higher swapping rate yielding higher DIF preservation likelihoods. All three levels for the RF were found to have strong positive influences on DIF preservation. Data containing higher ratios, namely 2:1, 5:1, and 10:1 ratios, tended to have higher preservation likelihoods than data containing 1:1 ratios. Including the RF main effect in Model 3 increased predictive power since the c-statistic and correct classification rate increased to 0.989 and 93.08% respectively.

Model 4 involved the main and interaction effects of the DL, SR, and RF predictors. Little change in the AIC was observed after including the interaction effects in the model (123,437.27 versus 132,073.15 from Model 3). This, therefore, suggested very few significant interaction effects compared to Model 3. Similar to the previous models, the DL had a nonsignificant negative effect on DIF preservation while the RF had a significant positive effect. It was surprising to observe a significant positive effect in DIF preservation only when a 1% swapping rate was used compared to a 5% rate. It is believed that the inclusion of the interaction effects in the model changed the behavior of this phenomenon. There were negative effects in DIF preservation when BB- and CC-level polytomous items were swapped at a 1% SR and positive significant effects when swapped at the 5% SR. There were also positive significant effects when BB- and CC-level items were created from data with a 5:1 or 10:1 RF versus a 1:1 ratio. There were no significant effects detected between the interaction of the SR and RF predictors, but significant positive and negative effects were found between the interaction of the DL and SR variables. These significant interactions offset the ironic

Table 4.8: Interaction Effect Cumulative Logit Model: Polytomous

	Model 4			Model 4			Model 4	
Estimates	β	se_β	**Estimates**	β	se_β	**Estimates**	β	se_β
Intercept (α)			2-Way Interactions			3-Way Interactions		
α_{AA_1}	3.11***		$DL_{BB} \times SR_1$	-2.14***	0.27	$DL_{BB} \times SR_1 \times RF_2$	-0.09	0.38
α_{AA_2}	22.13		$DL_{BB} \times SR_5$	1.70***	0.15	$DL_{BB} \times SR_1 \times RF_5$	-0.30	0.31
α_{BB}	47.69		$DL_{CC} \times SR_1$	-3.46***	0.73	$DL_{BB} \times SR_1 \times RF_{10}$	-0.35	0.30
Pre-Swap DIF Level (DL)			$DL_{CC} \times SR_5$	2.15***	0.21	$DL_{BB} \times SR_5 \times RF_2$	0.32	0.20
DL_{BB}	-26.89	86.92	$DL_{BB} \times RF_2$	0.28	0.17	$DL_{BB} \times SR_5 \times RF_5$	0.02	0.18
DL_{CC}	-53.43	128.80	$DL_{BB} \times RF_5$	1.65***	0.15	$DL_{BB} \times SR_5 \times RF_{10}$	-0.15	0.18
Swapping Rate (SR)			$DL_{BB} \times RF_{10}$	2.90***	0.15	$DL_{CC} \times SR_1 \times RF_2$	0.89	0.91
SR_1	0.69***	0.11	$DL_{CC} \times RF_2$	-0.01	0.26	$DL_{CC} \times SR_1 \times RF_5$	0.54	0.77
SR_5	-0.06	0.09	$DL_{CC} \times RF_5$	2.05***	0.21	$DL_{CC} \times SR_1 \times RF_{10}$	0.52	0.75
Reference-Focal Ratio (RF)			$DL_{CC} \times RF_{10}$	3.88***	0.20	$DL_{CC} \times SR_5 \times RF_2$	0.78**	0.29
RF_2	0.04	0.09	$SR_1 \times RF_2$	-0.23	0.15	$DL_{CC} \times SR_5 \times RF_5$	0.26	0.24
RF_5	0.34***	0.09	$SR_1 \times RF_5$	-0.18	0.15	$DL_{CC} \times SR_5 \times RF_{10}$	0.03	0.23
RF_{10}	0.59***	0.09	$SR_1 \times RF_{10}$	-0.20	0.16			
			$SR_5 \times RF_2$	-0.19	0.12			
AIC	123,437.27		$SR_5 \times RF_5$	-0.06	0.13			
c	0.990		$SR_5 \times RF_{10}$	0.06	0.13			
CCR	93.08%							

*$p < 5\%$,**$p < 1\%$, ***$p < 0.1\%$

behavior found in the offset of the SR main effect. Additionally, only one combination of the interaction was found to have a significant effect on DIF preservation (DL =CC, SR =5%, RF =2:1).

Table 4.9 contains the predicted likelihoods for DIF levels after data swapping. This table is read similar to that of Table 4.6 where the predicted probabilities are row conditional. A considerable number of the combinations for the SR and the RF experienced concordant results regarding the DIF before and after swapping. The likelihoods were at least 90%, but they decreased as the levels of the SR and RF increased. There were two combinations for the SR and RF that experienced discordant results. At the SR of 5% and RF of 10:1, polytomous items containing BB- or CC-level DIF had a greater chance of experiencing different levels of DIF after data swapping than similar levels. For polytomous items containing BB-level DIF, there was a higher likelihood of producing an item with AA_2-level DIF after data swapping (57.43% versus 42.57%), resulting in a relative risk of 34.9%. Polytomous items containing CC-level DIF had a higher likelihood of producing polytomous items containing BB-level DIF after data swapping (71.46% versus 28.54%), producing an even higher relative risk (150.4%). These relative risks show the magnitude between the likelihoods of preserving the correct level of DIF versus a less severe level of DIF. Nonetheless, the results from this table show that higher levels of the SR and RF can reduce the level of DIF in polytomous items.

Table 4.9: Predicted DIF Preservation Likelihoods: Polytomous

			DIF After Swapping			
SR	**RF**	**DIF Before Swapping**	AA_1	AA_2	BB	CC
1%	1:1	AA_1	**97.81%**	2.19%	<0.01%	<0.01%
1%	1:1	BB	<0.01%	0.20%	**99.80%**	<0.01%
1%	1:1	CC	<0.01%	<0.01%	0.02%	**99.98%**
1%	2:1	AA_1	**97.37%**	2.63%	<0.01%	<0.01%
1%	2:1	BB	<0.01%	0.20%	**99.80%**	<0.01%
1%	2:1	CC	<0.01%	<0.01%	0.04%	**99.96%**
1%	5:1	AA_1	**98.13%**	1.87%	<0.01%	<0.01%
1%	5:1	BB	<0.01%	0.90%	**99.10%**	<0.01%
1%	5:1	CC	<0.01%	<0.01%	0.31%	**99.69%**
1%	10:1	AA_1	**98.51%**	1.49%	<0.01%	<0.01%
1%	10:1	BB	<0.01%	3.69%	**96.31%**	<0.01%
1%	10:1	CC	<0.01%	<0.01%	2.36%	**97.64%**
3%	1:1	AA_1	**95.74%**	4.26%	<0.01%	<0.01%
3%	1:1	BB	<0.01%	0.85%	**99.15%**	<0.01%
3%	1:1	CC	<0.01%	<0.01%	0.32%	**99.68%**
3%	2:1	AA_1	**95.90%**	4.10%	<0.01%	<0.01%
3%	2:1	BB	<0.01%	1.17%	**98.83%**	<0.01%
3%	2:1	CC	<0.01%	<0.01%	0.33%	**99.67%**
3%	5:1	AA_1	**96.94%**	3.06%	<0.01%	<0.01%
3%	5:1	BB	<0.01%	5.93%	**94.07%**	<0.01%
3%	5:1	CC	<0.01%	<0.01%	3.39%	**96.61%**
3%	10:1	AA_1	**97.60%**	2.40%	<0.01%	<0.01%
3%	10:1	BB	<0.01%	22.11%	**77.89%**	<0.01%
3%	10:1	CC	<0.01%	<0.01%	21.99%	**78.01%**
5%	1:1	AA_1	**95.46%**	4.54%	<0.01%	<0.01%
5%	1:1	BB	<0.01%	4.24%	**95.76%**	<0.01%
5%	1:1	CC	<0.01%	<0.01%	2.52%	**97.48%**
5%	2:1	AA_1	**94.75%**	5.25%	<0.01%	<0.01%
5%	2:1	BB	<0.01%	6.45%	**93.55%**	<0.01%
5%	2:1	CC	<0.01%	<0.01%	4.59%	**95.41%**
5%	5:1	AA_1	**96.55%**	3.45%	<0.01%	<0.01%
5%	5:1	BB	<0.01%	23.75%	**76.25%**	<0.01%
5%	5:1	CC	<0.01%	<0.01%	25.58%	**74.42%**
5%	10:1	AA_1	**97.60%**	2.40%	<0.01%	<0.01%
5%	10:1	BB	<0.01%	**57.43%**	42.57%	<0.01%
5%	10:1	CC	<0.01%	<0.01%	**71.46%**	28.54%

Chapter 5

Discussion

The focus of this research study was the investigation of preserving uniform DIF after a disclosure avoidance method, namely data swapping, was applied to protect the data. Three research questions were addressed as a result of this study, and predictive models and several statistical methods were used to answer these questions. This study estimated the rate at which uniform DIF is preserved in dichotomous and polytomous items after applying data swapping to the data. It then determined whether association was found between the likelihood of preserving uniform DIF, the data swapping rate, the item scoring, the reference-to-focal group ratio, and the severity of the DIF originally detected. Then finally, it determined whether a generalized linear model could explain the association between the likelihood of preserving uniform DIF, the data swapping rate, the item scoring, the reference-to-focal group ratio, and the severity of the DIF originally detected. The purpose of this chapter is to summarize the findings, address the significance of the research study, and

discuss the study's limitations which address several directions for further research.

It is concluded that when a dichotomous item contains B-level DIF, its level of DIF is preserved between 70 to 97 of every 100 times after data swapping. When the dichotomous item contains C-level DIF, its preservation happens between 79 to 99 of every 100 data swapping attempts. At the baseline swapping rate of 3% and a 1:1 reference-to-focal group ratio, the preservation rates for both DIF levels were approximately equal to 96 of every 100 data swapping attempts. However, increases in the swapping rate or group ratios decrease these preservation rates. With polytomous items, BB-level DIF is preserved between 71 to 99 of every 100 data swapping attempts and CC-level DIF is preserved between 28 to 99 of every 100 attempts. Similar to the dichotomous case, increases in the swapping rate and group ratios decrease these rates. At the baseline, both levels of DIF are preserved 99 of every 100 data swapping attempts. Despite the lower limit of the preservation rate range for the polytomous case, preservation rates were more stable than those of the dichotomous case. When omitting the largest swapping rate and group ratio treatment, preservation rates for CC-level DIF were found to be between 74 to 99 of very 100 data swapping attempts.

Results from the Cochran-Mantel Haenszel tests and Somer's D correlation coefficients suggest that pre-swapped DIF levels in dichotomous and polytomous items significantly correlate with post-swapped DIF levels for all combinations of swapping rate and reference-to-focal group ratio. Associations between pre-swapped and post-swapped DIF levels were found to be stronger for polytomous items than for dichotomous items, smaller reference-to-focal ratios than larger ones, and smaller swapping rates than larger swapping rates. One

possible explanation is that data swapping does not add enough noise to alter the distribution and variance of scores for polytomous items than for dichotomous items.

It is concluded that a generalized linear model, specifically the cumulative logit model, can model the likelihood of preserving dichotomous and polytomous DIF conditional on swapping rate and reference-to-group ratio. When using a main-effects model to measure DIF preservation likelihoods in dichotomous items, one should expect to see significant negative effects in the logit when the items contain DIF before swapping or when data are underprotected. Significant positive effects in the logit occur when the data is overprotected or when the data contains unequal reference and focal group sample sizes. When using a full-effects model, one should expect to observe the same effects found in the main-effects model, significant negative two-way effects between pre-swap DIF level and the 1% swapping rate, pre-swap DIF and the 10:1 group ratio, and group ratio and the 1% swapping rate, and significant positive two-way effects between pre-swap DIF and the 5% swapping rate, group ratios and the C-level pre-swap DIF, and group ratios and the 5% swapping rate. Additionally, significant negative three-way effects occur at B- and C-level pre-swap DIF levels and the 5% swapping rate with relatively large group ratios. For polytomous items, one should observe significant negative effects when data is underprotected, significant positive effects when the data is overprotected, and significant positive effects when data contains unequal reference and focal group sample sizes. Under the full-effects model, one should expect to observe a significant positive effect when underprotecting data or when using data containing unequal reference and focal group sample sizes. Significantly negative two-way effects are

expected between pre-swap DIF and the 1% swapping rate, but positive effects at the 5% swapping rate. Positive two-way effects are also expected between group ratios at BB- and CC-level pre-swap DIF levels.

It is interesting to note that the GLMs suggest concordant pre- and post-swap DIF levels in dichotomous items when DIF is present before data swapping. For dichotomous items containing no DIF, data swapping increases the presence of DIF but not enough to reach the B- and C-DIF levels. However, this is the case when the data contains smaller reference-to-focal group ratios. Concordant behavior also occurred in the polytomous model for nearly all combinations of the pre-swapped item DIF, swapping rate, and reference-to-focal group ratio. But at high swapping rates and group ratios, there is potential for polytomous items containing moderate and high levels of DIF to contain negligible and moderate DIF levels after data swapping respectively.

In determining the optimal swapping rate to use, several suggestions are recommended to the data proprietor. First, one should consider the number of dichotomous and polytomous items containing DIF in their data. Assessments containing items with little to no DIF should experience no significant harm with overprotection, but proprietors should use caution when overprotecting data containing any DIF items. Swapping rates should be more conservative for data containing polytomous items with DIF than dichotmous items, and the severity of DIF in such items makes it more important to use a swapping rate less than three percent. Finally, data proprietors should also take into account the magnitude of the group sizes that are responsible for the item DIF, and use the results from Table 4.1 to make an empirical

decision on the best swapping rate to use that minimizes the amount of item preservation lost.

Although little research exists in investigating the relationship between disclosure avoidance protection and factors of assessment data, the results of this study validate two points addressed in prior research that investigated these factors individually. First, the findings in this study support published literature that states that forms of disclosure avoidance protection, including data swapping, help preserve statistical properties in data (Willenborg and de Waal, 2001; Shlomo et al., 2010; Fienberg and McIntyyre, 2004; Dalenius and Reiss, 1982). In this research, the statistical property to preserve was uniform DIF. The findings from this study also validate that disclosure avoidance methods add noise to data, and noise results in information loss in data (Nayak et al., 2011; Zayatz, 2007). This means that the behavior found in data before protection can be significantly different afterwards.

There is a strong significance for this research with regard to educational assessments. The need to disseminate data is an important task, and so is protecting the privacy of the respondents that were used to make it. Data swapping is one of several powerful disclosure avoidance methodologies, but more research is needed to investigate how well it can preserve the statistical properties within the data. Several studies have evaluated the efficiency of data swapping using metadata produced by non-government and government agencies, but none of these studies has investigated its efficiency using assessment data. This, therefore, questions the generalizability of this knowledge into the educational setting and identifies the significance of this research. The idea and results from this research study represent

the first steps in gaining insight of how disclosure avoidance techniques protect privacy in assessment data by integrating the two issues together. In doing so, this study also provides a significant contribution by comparing changes in DIF preservation rates when the swapping rate is compromised by a data proprietor's choice or by legislation (see Table 4.1).

The findings of this study have several limitations worth addressing. One of the strongest limitations is the study's attention to uniform DIF. Although it is a common and popular type of DIF to test, it is not the only one. With standardized assessments, it is as important to test for non-uniform DIF since items can exhibit a significant difference in group performances as a function of the ability levels.

A second limitation of this study involves the particular DIF methodology used to detect DIF. As explained in the literature review, other DIF approaches such as logistic regression, the SIBTEST, and the likelihood ratio test have been widely used and accepted as effective methods to identify biased items based on group performances. Considering all of these methods in one study would create much confusion in understanding the holistic nature of DIF preservation likelihoods.

It is important to also emphasize that this study considered focal and reference groups containing no difference in ability levels. Although constant difficulty was added to items to contain B- and C-levels of DIF, this difficulty was added to address that these items biased against focal groups containing just as enough ability to answer these items as the reference groups. Adding such a constant to the items while considering the use of differences in the reference and focal groups would create a challenge in simulating items with all levels of

DIF.

Lastly, the findings of this study is limited to the use of unidimensional models. Item scores were created using item response theory models. Particularly, the the three-parameter and generalized partial credit model were used to create the dichotmous and polytomous scores respectively. Both models considered the assumption of a unidimensional latent trait and local independence of the item scores. Although these properties are pivotal for the models to work, one cannot rule out the possiblity for assessments to contain items with a multidimensional trait space or item scores being dependent.

This study suggests that more work is needed to understand the relationship between disclosure avoidance methods and DIF. For example, one could consider modeling the likelihood of preserving nonuniform DIF and how it decreases or increases the likelihood. As discussed previously, items with nonuniform DIF are said to possess a trend in performance that is a function of the examinee ability levels.

Researchers could investigate whether data swapping strengthens or weakens this trend, resulting in possibly higher or lower DIF preservation likelihoods than the likelihoods found in this study. Another topic worth exploring is using other disclosure avoidance techniques to protect privacy. Data shuffling (Muralidhar and Sarathy, 2006) permutes the values of the confidential data within the data. Unlike data swapping which depends on marginal distributions, data shuffling was found to have lower risks of data disclosure (Muralidhar and Sarathy, 2006). Synthetic data (Reiter, 2002) and multiplicative noise (Nayak et al., 2011) are other techniques that are said to preserve the nature of estimators from non-perturbed

data produced from several sampling designs. Since these techniques contain unique steps and impact the data in different ways, one could discover that the DIF preservation likelihoods are significantly smaller or larger than those for data swapping.

Resarchers can also explore how to model preservation likelihoods for other DIF detection methods. The SIBTEST and logistic regression methods are among other popular robust methodologies for detecting DIF, and therefore one may be curious as to how changes in the methodology used plays a role in how DIF is preserved. One may find that modeling preservation likelihoods is more challenging when these methods are used.

Finally, an interesting question for item analysts is to investigate how the number or percentage of items containing DIF prior to data swapping affect DIF likelihood preservation. In this study, a forty-item assessment was used in which one dichotomous and one polytomous item contained DIF. Although it is common for large-scale assessments to contain either none or one DIF item (Camilli, 2006) since items are often pre-tested prior to production, one could investigate how the presence of multiple DIF items impacts the ability to preserve item DIF at any level.

References

Agresti, A. (2002). *Categorical Data Analysis* (Third Edition). John Wiley & Sons.

Aguerri, M. E., Galibert, M. S., Attorresi, H. F., and Maranon, P. P. (2009). Erroneous detection of nonuniform DIF using the Breslow-Day test in a short test. *Quality and Quantity*, *43*(1), 35-44.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.*

Anderson, R.D. and DeMars, C. (2002). Differential item functioning: investigating item bias. *Assessment Update*, *14*(3), 12-13.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*(3), 375-386.

Assembly of Behavioral and Social Sciences. Panel on Privacy and Confidentiality as Factors in Survey Response (1979). *Privacy and Confidentiality as Factors in Survey Response*, *v.2878*, National Academies.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*(2), 113-141.

Bolt, S. and Ysseldyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. *Applied Measurement in Education*, *19*(4), 329-355.

Breslow, D.E., and Day, N.E. (1980). *Statistical methods in cancer research*. International Agency for Research on Cancer.

Camilli, G. (2006). Test fairness. *Educational Measurement*, *4*, 221-256.

Camilli, G., and Shepard, L.A. (1994). *Methods for identifying biased test items*. Sage.

Clauser, B. E. and Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*. *17*(1), 31-44.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, *10*(4), 417-451.

Cody, R. P. (2010). *SAS functions by example*. SAS Institute.

Cohen, A. S., Gregg, N., and Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice*, *20*, 225-233.

Cox, M. L., Herner, J. G., Demczyk, M. J., and Nieberding, J. J. (2006). Provision of testing accommodations for students with disabilities on statewide assessments: Statistical links with participation and discipline rates. *Remedial and Special Educations*, *27*, 346-354.

Dalenius, T., and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, *6*(1), 73-85.

Dorans, N. J., and Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. *Construction versus Choice in Cognitive Measurement*, 135-165.

Doyle, P., Lane, J. I., Theeuwes, J. J. M., and Zayatz, L. (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier Science, Amsterdam, Netherlands.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., and Roehrig, S. F. (2001). Chapter 7: disclosure limitation methods and information loss for tabular data. In Doyle, P., Lane, J. I., Theeuwes, J. J. M., and Zayatz, L. (Eds.), *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, 135-166. Amsterdam, Netherlands: Elsevier Science.

Elbaum, B. (2007). Effects of an oral testing accommodations on the mathematics performance of secondary students with and without learning disabilities. *The Journal of Special Education*, *40*, 218-229.

Fidalgo, A.M., Ferreres, D., and Muniz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for type I and type II error rates. *The Journal of Experimental Education*, *73*(1), 23-39.

Fienberg, S. E., and McIntyre, J. (2004). Data swapping: variation on a theme by Dalenius and Reiss. In Domingo-Ferrer, J. and Torra, V. (Eds.), *Privacy in Statistical Databases 2004 LNCS 3050*, 14-29, Berlin, Germany: Springer-Verlag.

Finch, H., Barton, K., and Meyer, P. (2009). Differential item functioning analysis for accommodated versus nonaccommodated students. *Educational Assessment*, *14*, 38-56.

Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732-764.

Hambleton, R. K. and Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*, 313-334.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, *8*(1), 35-41.

Herrera, A. N., and Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the MantelHaenszel and logistic regression techniques. *Quality & Quantity*, *42*(6), 739-755.

Hidalgo, M. D. and Lopez-Pina, J.A. (2004). Differential item functioning detection and

effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903-915.

Holland, P.W. and Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test Validity*, 129-145.

Hosmer, D.W. and Lemeshow, S. *Applied Logistic Regression* (2nd Edition). New York, NY: John Wiley & Sons.

Hunderpool, H, Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and de Wolf, Peter-Paul (2012). *Statistical Disclosure Control*. John Wiley & Sons Ltd., West Sussex, United Kingdom.

Institute of Education Sciences (2013). *Statistical Standards Program: Confidentiality Procedures*. Retrieved January 4, 2013 from website http://nces.ed.gov/statprog/confproc.asp.

Kamata, A. and Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal*, *2*(2), 49-69.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1992). Cell suppression: Disclosure protection for sensitive tabular data. *Networks*, *22*(4), 397-417.

Ketterlin-Geller, L., Alonzo, J., Braun-Monegan, J., and Tindal, G. (2007). Recommendations for accommodations: Implications of (in)consistency. *Remedial and Special Education*, *28*(4), 194-206.

Kettler, R. J., Niebling, B. C., Mroch, A. A., Feldman, E. S., Newell, M. L., Elliott, S. N., Kratochwill, T.R., and Bolt, D.M. (2005). Effects of testing accommodations on math and reading scores: An experimental analysis of the performance of students with and without disabilities. *Assessment for Effective Intervention*, *31*(1), 3748.

Kim, Do-Hong, Schneider, C., and Siskind, T. (2009). Examining equivalence of accommodations on a statewide elementary-level science test. *Applied Measurement in Education*, *22*, 144-163.

Liu, I-M and Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics*, *52*, 1223-1234.

Lucero, J. and Zayatz, L. (2010). The microdata analysis system at the U.S. Census Bureau. In Domingo-Ferrer, J. and Magkos, E. (Eds.), *Privacy in Statistical Databases 2010 LNCS 6344*, 41-51, Berlin, Germany: Springer-Verlag.

Magis, D. and De Boeck, P. (2012). A robust outlier approach to prevent Type I error inflation in DIF. *Educational and Psychological Measurement*, *72*(2), 291-311.

Mantel, N. (1963). Chi-Squared tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690-700.

Massell, P. and Funk, J. (2007). Protecting the Confidentiality of Tables by Adding Noise to the Underlying Microdata. In *Proceedings of the 2007 Third International Conference on Establishment Surveys (ICES-III)*, Montreal, Canada.

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

Meyer, J. P., Huynh, H., and Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement*, *41*(4), 331-344.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Muralidhar, K., and Sarathy, R. (2006). Data shuffling-a new masking approach for numerical data. *Management Science*, *52*(5), 658-670.

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.

Nayak, T. K., Sinha, B., and Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection. *Journal of Official Statistics*, *27*(3), 527.

No Child Left Behind Act of 2001 (2002). Pub. L. No. 107-110, 115 Stat. 1425.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business & Economic Statistics*, *6*(4), 487-500.

Parshall, C. G., and Miller, T. R. (1995). Exact Versus Asymptotic Mantel-Haenszel DIF Statistics: A Comparison of Performance Under Small-Sample Conditions. *Journal of Educational Measurement*, *32*(3), 302-316.

Pearson Educational Measurement (2007). *2006 Technical Report of the Washington Assessment of Student Learning: Grade 10*. Retrieved June 13, 2013 from website http://www.k12.wa.us/assessment/pubdocs/2006HSWASLTechReport.pdf.

Penfield, R. D. (2003). Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. *Alberta Journal of Educational Research*, *49*, 231-243.

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

Penfield, R. D., and Lam, T. (2000). Assessing differential item functioning in performance assessment: review and recommendations. *Educational Measurement: Issues and Practice*, *19*(3), 5-15.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, *19*(1), 1-16.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Newbury Park, CA: Sage.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics-Stockholm*, *18*(4), 531-544.

Roussos, L. A., and Stout, W. F. (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.

Rubin, D. (1993). Discussion statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461-468.

Ruggles, R., and Ruggles, N. D. (1975). The role of microdata in the national economic and social accounts. *Review of Income and Wealth*, *21*(2), 203-216.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. KTL Scientific Publishers, MA.

Salend, A. (2008). Determining appropriate testing accommodations: complying with NCLB and IDEA. *Teaching Exceptional Children*, *40*(4),14-22.

Samarati, P. (2001). Protecting respondents identities in microdata release. *Knowledge and Data Engineering, IEEE Transactions on*, *13*(6), 1010-1027.

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

Shepard, L. A., Camilli, G., and Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*, 77-105.

Shlomo, N., Tudor, C., and Groom, P. (2010). Data swapping for protecting census tables. In Domingo-Ferrer, J. and Magkos, E. (Eds.), *Privacy in Statistical Databases 2010 LNCS 6344*,41-51, Berlin, Germany: Springer-Verlag.

Snijders, T. A. B., and Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications. Thousand Oaks, CA.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 799-811.

Spray, J. A. (1989). *Performance of three conditional DIF statistics in detecting differential item functioning on simulated tests (Research Rep. No. 89-7)*. Iowa City, IA: American College Testing.

Stokes, M. E., Davis, M. E. S. C. S., Koch, G. G., and Davis, C. S. (2003). *Categorical data analysis using the SAS system*. SAS Institute.

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

Thissen, D., Steinberg, L., and Wainer, H. (1988). Detection of differential item functioning using the parameters of item response theory models. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Wacks, R. (2010). *Privacy: A Very Short Introduction*. Oxford University Press. New York, NY.

Wang, Wen-Chung, and Su, Ya-Hui. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28*(6), 450-480.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York, NY: Springer-Verlag.

Williams, V. (1997). The "unbiased" anchor: bridging the gap between DIF and item bias. *Applied Measurement in Education*, *10*, 253-267.

Woods, C. M. (2008). Likelihood-ratio DIF testing: effects of nonnormality. *Applied Psychological Measurement*, *32*(7), 511-526.

Zayatz, L. (2007). Disclosure avoidance practices and research at the U.S. Census Bureau: an update. *Journal of Official Statistics*, *23*(2), 253-265.

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

Zumbo, B. (2007). Three generations of DIF analyses: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223-233.

Zwick, R., Thayer, D. T., and Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10*(4), 321-344.

Appendix A

Record Layout of Simulated Data

Table A.1: Record Layout of Simulated Data

Variable Name	**Variable Description**	**Variable Values**
Distsch	Unique school identification number	001-580
Setting	Test setting accommodation needed	0 = no (reference), 1= yes (focal)
Msum	Examinee Total Score	0-100
itXX	Score for dichotomous item XX	0 = no credit, 1 = full credit
pitXX	Score for polytomous item XX	0 = no credit, 1 - 3 = partial credit, 4 = full credit
flag_clang	examinee is unique by language flag	0 = flagged as unique, 1 = flagged as unique
flag_cethnic	examinee is unique by ethnicity flag	0 = not flagged as unique, 1 = flagged as unique
flag_cdisabil	examinee is unique by disability flag	0 = not flagged as unique, 1 = flagged as unique
flag_cgender	examinee is unique by gender flag	0 = not flagged as unique, 1 = flagged as unique
flag_clapmath	examinee is unique by special education flag	0 = not flagged as unique, 1 = flagged as unique
flag_ctasmath	examinee is unique by taking Title 1 class flag	0 = not flagged as unique, 1 = flagged as unique
flag_cgifted	examinee is unique as a gifted student flag	0 = not flagged as unique, 1 = flagged as unique

Appendix B

DIF Preservation Tables-Dichotomous

Table B.1: DIF Preservation-Dichotomous (Ratio = 10:1 & Swap Rate = 1%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	5,584	0	2,241	0	2,136	0	39
B-Level	49	0	1,498	401	7,031	6,543	1,422	87
C-Level	0	0	10	0	1,431	462	8,559	8,097

Table B.2: DIF Preservation-Dichotomous (Ratio = 10:1 & Swap Rate = 3%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	6,258	0	2,039	0	1,691	0	12
B-Level	34	0	1,529	1,064	6,997	5,848	1,440	85
C-Level	0	0	1	0	1,464	1,373	8,535	7,162

Table B.3: DIF Preservation-Dichotomous (Ratio = 10:1 & Swap Rate = 5%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	6,997	0	1,816	0	1,178	0	9
B-Level	47	13	1,577	2,045	6,918	4,815	1,458	45
C-Level	0	0	6	1	1,508	2,849	8,486	5,636

Table B.4: DIF Preservation-Dichotomous (Ratio = 5:1 & Swap Rate = 1%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	1,893	0	6,369	0	1,735	0	3
B-Level	0	0	982	211	8,217	7,963	801	43
C-Level	0	0	0	0	863	220	9,137	8,917

Table B.5: DIF Preservation-Dichotomous (Ratio = 5:1 & Swap Rate = 3%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	2,394	0	6,200	0	1,404	0	2
B-Level	1	0	1,007	629	8,256	7,585	736	42
C-Level	0	0	0	0	929	783	9,071	8,288

Table B.6: DIF Preservation-Dichotomous (Ratio = 5:1 & Swap Rate = 5%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	3,004	0	5,974	0	1,021	0	1
B-Level	0	0	1,027	1,273	8,187	6,887	786	27
C-Level	0	0	0	0	938	1,851	9,062	7,211

Table B.7: DIF Preservation-Dichotomous (Ratio = 2:1 & Swap Rate = 1%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	27	0	8,763	0	1,210	0	0
B-Level	0	0	576	114	9,165	9,041	259	10
C-Level	0	0	0	0	519	126	9,481	9,355

Table B.8: DIF Preservation-Dichotomous (Ratio = 2:1 & Swap Rate = 3%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	309	0	8,703	0	988	0	0
B-Level	0	0	556	297	9,168	8,859	276	12
C-Level	0	0	0	0	466	435	9,534	9,099

Table B.9: DIF Preservation-Dichotomous (Ratio = 2:1 & Swap Rate = 5%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	424	0	8,832	0	744	0	0
B-Level	0	0	544	748	9,176	8,426	280	2
C-Level	0	0	0	0	487	1,153	9,513	8,360

Table B.10: DIF Preservation-Dichotomous (Ratio = 1:1 & Swap Rate = 1%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	111	0	8,757	0	1,132	0	0
B-Level	0	0	464	79	9,397	9,313	139	5
C-Level	0	0	0	0	392	95	9,608	9,513

Table B.11: DIF Preservation-Dichotomous (Ratio = 1:1 & Swap Rate = 3%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	117	0	8,895	0	988	0	0
B-Level	0	0	442	248	9,441	9,191	117	2
C-Level	0	0	0	0	375	336	9,625	9,289

Table B.12: DIF Preservation-Dichotomous (Ratio = 1:1 & Swap Rate = 5%)

	A_1-Level		A_2-Level		B-Level		C-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
A_1-Level	10,000	167	0	9,121	0	712	0	0
B-Level	0	0	434	648	9,419	8,770	147	1
C-Level	0	0	0	0	376	989	9,624	8,635

Appendix C

DIF Preservation Tables- Polytomous

Table C.1: DIF Preservation-Polytomous (Ratio = 10:1 & Swap Rate = 1%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	8,877	8,745	1,123	132	0	0	0	0
BB-Level	0	0	323	357	9,676	9,319	1	0
CC-Level	0	0	0	0	62	235	9,938	9,703

Table C.2: DIF Preservation-Polytomous (Ratio = 10:1 & Swap Rate = 3%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	8,882	8,669	1,118	213	0	0	0	0
BB-Level	0	0	322	2,139	9,676	7,537	2	0
CC-Level	0	0	0	0	77	2,182	9,923	7,741

Table C.3: DIF Preservation-Polytomous (Ratio = 10:1 & Swap Rate = 5%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	8,883	8,670	1,117	213	0	0	0	0
BB-Level	0	0	330	5,551	9,666	4,115	4	0
CC-Level	0	0	0	0	88	7,083	9,912	2,829

Table C.4: DIF Preservation-Polytomous (Ratio = 5:1 & Swap Rate = 1%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	8,031	7,881	1,969	150	0	0	0	0
BB-Level	0	0	96	89	9,904	9,815	0	0
CC-Level	0	0	0	0	16	31	9,984	9,953

Table C.5: DIF Preservation-Polytomous (Ratio = 5:1 & Swap Rate = 3%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	7,873	7,632	2,127	241	0	0	0	0
BB-Level	0	0	95	587	9,905	9,318	0	0
CC-Level	0	0	0	0	17	338	9,983	9,645

Table C.6: DIF Preservation-Polytomous (Ratio = 5:1 & Swap Rate = 5%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	7,920	7,647	2,080	273	0	0	0	0
BB-Level	0	0	103	2,351	9,897	7,546	0	0
CC-Level	0	0	0	0	17	2,554	9,983	7,429

Table C.7: DIF Preservation-Polytomous (Ratio = 2:1 & Swap Rate = 1%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	6,301	6,135	3,699	166	0	0	0	0
BB-Level	0	0	25	20	9,975	9,955	0	0
CC-Level	0	0	0	0	4	4	9,996	9,992

Table C.8: DIF Preservation-Polytomous (Ratio = 2:1 & Swap Rate = 3%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	6,315	6,056	3,685	259	0	0	0	0
BB-Level	0	0	44	116	9,956	9,840	0	0
CC-Level	0	0	0	0	8	33	9,992	9,959

Table C.9: DIF Preservation-Polytomous (Ratio = 2:1 & Swap Rate = 5%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	6,362	6,028	3,638	334	0	0	0	0
BB-Level	0	0	29	643	9,971	9,328	0	0
CC-Level	0	0	0	0	2	459	9,998	9,539

Table C.10: DIF Preservation-Polytomous (Ratio = 1:1 & Swap Rate = 1%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	5,854	5,726	4,146	128	0	0	0	0
BB-Level	0	0	27	20	9,973	9,953	0	0
CC-Level	0	0	0	0	3	2	9,997	9,995

Table C.11: DIF Preservation-Polytomous (Ratio = 1:1 & Swap Rate = 3%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
AA_1-Level	5,840	5,591	4,160	249	0	0	0	0
BB-Level	0	0	21	85	9,979	9,894	0	0
CC-Level	0	0	0	0	2	32	9,998	9,966

Table C.12: DIF Preservation-Polytomous (Ratio = 1:1 & Swap Rate = 5%)

	AA_1-Level		AA_2-Level		BB-Level		CC-Level	
Created	BDS	ADS	BDS	ADS	BDS	ADS	BDS	ADS
No-DIF	5,798	5,535	4,202	263	0	0	0	0
BB-Level	0	0	23	423	9,977	9,554	0	0
CC-Level	0	0	0	0	3	252	9,997	9,745

Appendix D

DIF Level Preservation Likelihood Tables

Table D.1: Dichotomous DIF Level Preservation Likelihoods: A = A_1

		DIF Preservation Levels		
Swap Rate	**Ref:Foc Ratio**	A	B	C
1%	10:1	55.84%	93.06%	94.60%
1%	5:1	18.93%	96.91%	97.59%
1%	2:1	0.27%	98.65%	98.67%
1%	1:1	1.11%	99.11%	99.01%
3%	10:1	62.58%	83.58%	83.91%
3%	5:1	23.94%	91.87%	91.37%
3%	2:1	3.09%	96.63%	95.44%
3%	1:1	1.17%	97.35%	96.51%
5%	10:1	69.67%	69.60%	66.42%
5%	5:1	30.04%	84.12%	79.57%
5%	2:1	4.24%	91.83%	87.88%
5%	1:1	1.67%	93.11%	89.72%

Table D.2: Dichotomous DIF Level Preservation Likelihoods: A = $A_1 \bigcup A_2$

		DIF Preservation Levels		
Swap Rate	**Ref:Foc Ratio**	A	B	C
1%	10:1	78.25%	93.06%	94.60%
1%	5:1	82.62%	96.91%	97.59%
1%	2:1	87.90%	98.65%	98.67%
1%	1:1	88.68%	99.11%	99.01%
3%	10:1	82.97%	83.58%	83.91%
3%	5:1	85.94%	91.87%	91.37%
3%	2:1	90.18%	96.63%	95.44%
3%	1:1	90.12%	97.35%	96.51%
5%	10:1	88.83%	69.60%	66.42%
5%	5:1	89.78%	84.12%	79.57%
5%	2:1	92.56%	91.83%	87.88%
5%	1:1	92.88%	93.11%	89.72%

Table D.3: Polytomous DIF Level Preservation Likelihoods: AA = AA_1

		DIF Preservation Levels		
Swap Rate	**Ref:Foc Ratio**	AA	BB	CC
1%	10:1	98.51%	96.31%	97.64%
1%	5:1	98.13%	99.10%	99.69%
1%	2:1	97.37%	99.80%	99.96%
1%	1:1	97.81%	99.80%	99.98%
3%	10:1	97.60%	77.89%	78.01%
3%	5:1	96.94%	94.07%	96.61%
3%	2:1	95.90%	98.83%	99.67%
3%	1:1	95.74%	99.15%	99.68%
5%	10:1	97.60%	42.57%	28.54%
5%	5:1	96.55%	76.25%	74.42%
5%	2:1	94.75%	93.55%	95.41%
5%	1:1	95.46%	95.76%	97.48%

Table D.4: Polytomous DIF Level Preservation Likelihoods: AA = $AA_1 \bigcup AA_2$

		DIF Preservation Levels		
Swap Rate	**Ref:Foc Ratio**	AA	BB	CC
1%	10:1	100.00%	96.31%	97.64%
1%	5:1	100.00%	99.10%	99.69%
1%	2:1	100.00%	99.80%	99.96%
1%	1:1	100.00%	99.80%	99.98%
3%	10:1	100.00%	77.89%	78.01%
3%	5:1	100.00%	94.07%	96.61%
3%	2:1	100.00%	98.83%	99.67%
3%	1:1	100.00%	99.15%	99.68%
5%	10:1	100.00%	42.57%	28.54%
5%	5:1	100.00%	76.25%	74.42%
5%	2:1	100.00%	93.55%	95.41%
5%	1:1	100.00%	95.76%	97.48%

Appendix E

Dichotomous Effect Size Tables

Table E.1: Mean Effect Size-Dichotomous (Ratio = 10:1 & Swap Rate = 1%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.76	0.67	0.61	0.51
B-Level	N/A	1.56	1.71	1.87
C-Level	N/A	N/A	1.93	2.18

Table E.2: Mean Effect Size-Dichotomous (Ratio = 10:1 & Swap Rate = 3%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.75	0.66	0.61	0.52
B-Level	N/A	1.59	1.72	1.85
C-Level	N/A	N/A	1.97	2.21

Table E.3: Mean Effect Size-Dichotomous (Ratio = 10:1 & Swap Rate = 5%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.74	0.65	0.61	0.52
B-Level	1.56	1.62	1.73	1.84
C-Level	N/A	1.91	2.02	2.24

Table E.4: Mean Effect Size-Dichotomous (Ratio = 5:1 & Swap Rate = 1%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.79	0.70	0.62	0.52
B-Level	N/A	1.55	1.70	1.88
C-Level	N/A	N/A	1.91	2.13

Table E.5: Mean Effect Size-Dichotomous (Ratio = 5:1 & Swap Rate = 3%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.78	0.70	0.62	0.52
B-Level	N/A	1.56	1.71	1.87
C-Level	N/A	N/A	1.94	2.15

Table E.6: Mean Effect Size-Dichotomous (Ratio = 5:1 & Swap Rate = 5%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.77	0.69	0.62	0.55
B-Level	N/A	1.59	1.72	1.87
C-Level	N/A	N/A	1.97	2.17

Table E.7: Mean Effect Size-Dichotomous (Ratio = 2:1 & Swap Rate = 1%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.86	0.72	0.63	N/A
B-Level	N/A	1.54	1.69	1.88
C-Level	N/A	N/A	1.91	2.10

Table E.8: Mean Effect Size-Dichotomous (Ratio = 2:1 & Swap Rate = 3%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.82	0.71	0.63	N/A
B-Level	N/A	1.55	1.70	1.88
C-Level	N/A	N/A	1.92	2.11

Table E.9: Mean Effect Size-Dichotomous (Ratio = 2:1 & Swap Rate = 5%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.80	0.71	0.63	N/A
B-Level	N/A	1.57	1.70	1.88
C-Level	N/A	N/A	1.95	2.12

Table E.10: Mean Effect Size-Dichotomous (Ratio = 1:1 & Swap Rate = 1%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.83	0.71	0.63	N/A
B-Level	N/A	1.54	1.69	1.88
C-Level	N/A	N/A	1.90	2.09

Table E.11: Mean Effect Size-Dichotomous (Ratio = 1:1 & Swap Rate = 3%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.82	0.71	0.63	N/A
B-Level	N/A	1.55	1.69	1.88
C-Level	N/A	N/A	1.92	2.09

Table E.12: Mean Effect Size-Dichotomous (Ratio = 1:1 & Swap Rate = 5%)

	After Data Swapping			
Before Data Swapping	No-DIF	A-Level	B-Level	C-Level
No-DIF	0.82	0.71	0.63	N/A
B-Level	N/A	1.56	1.69	1.88
C-Level	N/A	N/A	1.94	2.11

Appendix F

Polytomous Effect Size Tables

Table F.1: Mean Effect Size-Polytomous (Ratio = 10:1 & Swap Rate = 1%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.98	0.96	N/A	N/A
BB-Level	N/A	1.02	1.22	N/A
CC-Level	N/A	N/A	1.28	1.32

Table F.2: Mean Effect Size-Polytomous (Ratio = 10:1 & Swap Rate = 3%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.96	N/A	N/A
BB-Level	N/A	1.18	1.21	N/A
CC-Level	N/A	N/A	1.27	1.31

Table F.3: Mean Effect Size-Polytomous (Ratio = 10:1 & Swap Rate = 5%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.96	N/A	N/A
BB-Level	N/A	1.17	1.20	N/A
CC-Level	N/A	N/A	1.26	1.30

Table F.4: Mean Effect Size-Polytomous (Ratio = 5:1 & Swap Rate = 1%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.98	0.97	N/A	N/A
BB-Level	N/A	1.18	1.22	N/A
CC-Level	N/A	N/A	1.28	1.32

Table F.5: Mean Effect Size-Polytomous (Ratio = 5:1 & Swap Rate = 3%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.97	N/A	N/A
BB-Level	N/A	1.18	1.21	N/A
CC-Level	N/A	N/A	1.28	1.31

Table F.6: Mean Effect Size-Polytomous (Ratio = 5:1 & Swap Rate = 5%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.97	N/A	N/A
BB-Level	N/A	1.18	1.20	N/A
CC-Level	N/A	N/A	1.28	1.30

Table F.7: Mean Effect Size-Polytomous (Ratio = 2:1 & Swap Rate = 1%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.98	N/A	N/A
BB-Level	N/A	1.18	1.21	N/A
CC-Level	N/A	N/A	1.28	1.32

Table F.8: Mean Effect Size-Polytomous (Ratio = 2:1 & Swap Rate = 3%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.98	N/A	N/A
BB-Level	N/A	1.18	1.21	N/A
CC-Level	N/A	N/A	1.28	1.32

Table F.9: Mean Effect Size-Polytomous (Ratio = 2:1 & Swap Rate = 5%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.98	N/A	N/A
BB-Level	N/A	1.18	1.20	N/A
CC-Level	N/A	N/A	1.28	1.30

Table F.10: Mean Effect Size-Polytomous (Ratio = 1:1 & Swap Rate = 1%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.98	N/A	N/A
BB-Level	N/A	1.18	1.21	N/A
CC-Level	N/A	N/A	1.28	1.32

Table F.11: Mean Effect Size-Polytomous (Ratio = 1:1 & Swap Rate = 3%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.98	N/A	N/A
BB-Level	N/A	1.18	1.21	N/A
CC-Level	N/A	N/A	1.28	1.31

Table F.12: Mean Effect Size-Polytomous (Ratio = 1:1 & Swap Rate = 5%)

	After Data Swapping			
Before Data Swapping	No-DIF	AA-Level	BB-Level	CC-Level
No-DIF	0.99	0.98	N/A	N/A
BB-Level	N/A	1.18	1.20	N/A
CC-Level	N/A	N/A	1.28	1.31