QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

FACTORS OF THE ASYMMETRIC NON-UNIFORM DIF DETECTION RATE WHEN USING THE ALTERNATIVE MANTEL-HAENSZEL PROCEDURE

BY

MOHAMMAD D. MOLLAZEHI

A Thesis Submitted to

the College of Arts and Sciences

in Partial Fulfillment of the Requirements for the Degree of

Masters of Science in Master of Science in Applied Statistics

January 2020

COMMITTEE PAGE

The members of the Committee approve the Thesis of
Mohammad Mollazehi defended on 12/10/2019.

AbdelSalam

Dr. Abdel-Salam Gomaa
Thesis/Dissertation Supervisor

Dr. Esam Mahdi
Committee Member

Approved:

Prof. Ibrahim AlKaabi, Dean, College of Arts and Sciences

ABSTRACT

MOLLAZEHI, MOHAMMAD, D., Masters: January : 2020, Applied Statistics

Title: Factors of the Asymmetric Non-Uniform DIF Detection Rate When Using the Alternative Mantel-Haenszel Procedure

Supervisor of Thesis: Abdel-Salam G., Abdel-Salam.

Test-item bias has become an increasingly challenging investigation in statistics and education. A popular method, the Mantel-Haenszel (MH) Test, is used for detecting non-uniform differential item functioning (DIF) but requires constructing several performance tiers to maintain robustness. Within the last two decades, the Alternative Mantel-Haeszel (AMH) Test (1994) was developed as a proxy procedure requiring only two scoring tiers. However, there is little information on how important factors like comparison group sizes, difficulty of questions, or question discrimination affect its ability to detect bias. In this statistical study, we investigate how item difficulty and discrimination as well as the ratio between the focal and reference groups examined impact the likelihood of the AMH detecting DIF.

This research begins with a simulation phase, in which test scores are generated under three conditions: three commonly-used difficulty levels (easy, medium, and hard), two discrimination levels (referred to as 'low' and 'high'), and three group comparison ratios (1:1, 2:1, and 5:1). From the simulation, the detection rates of the AMH Test are compared to those of another common test, like the Breslow-Day (BD) Test. The results are then used as input to fit post-hoc statistical models to determine, which of the three factors affect AMH detection behavior. The

study concludes with an application involving college-level test data comparing students across gender, nationality, and meta-major.

Keywords: Differential item functioning, non-uniform DIF, discrimination, item difficulty, Breslow-Day, Mantel-Haenszel

DEDICATION

I would like to dedicate this thesis to my family and friends for the utmost support they have provided throughout. I would also like to thank Aisha for her unwavering inspiration, without which, this work would never have been completed. In addition, my shout out goes to all the students or anyone who seeks and gains any sort of knowledge from my work.

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to All Mighty Allah. I cannot express enough how much my family and my friends have supported me to reach this point in my life. They stayed every step of the way and helped me to shape my future. When it comes to shaping my future how can I forget my supervisor Dr. Abdel-Salam Gomaa who has always believed in me and picked me up ever since I decided to join statistics and paved the way for my graduation. Appreciation goes to the Manager of Student Success at Qatar University, Dr. Khalifa Hazaa for providing me the data. Last but not least I would like to express my gratitude to all the great professors in Statistics program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1. Overview

Differential Item Functioning (DIF), also known as item bias, was first introduced in the 80's, and is still considered a relatively new concept. It has been one of the most controversial and most studied subjects in the theory of measurements. As an assessment tool, DIF has been used extensively in quantitative psychology, educational measurement, business management, and insurance, as well as in the healthcare sector (Holland & Wainer, 2012).

The aim of DIF analysis is detecting response differences of items in questionnaires, rating scales, or tests across various subgroups such as gender and nationality while controlling for ability level. DIF can be examined through four general procedures, namely the Mantel-Haenszel (MH), Breslow-Day (BD), Logistic Regression (LR) and Item Response Theory (IRT) (Zhang, 2015).

The MH procedure is a common approach in the detection of DIF. MH procedure is used to measure the estimate or strength of association. This test was brought up as a technique of detecting DIF by Holland and Thayer (1988). The approach has been extensively used in educational assessments due to its ease of implementation in testing programs. However, the approach is usually used in detecting uniform DIF for dichotomous items.

The BD test was formulated by Breslow and Day (1980). This is a test for homogeneity of odds ratios, that is used in the evaluation of changes in the degree of difference between two datasets under analysis in two different periods. The approach nonetheless has played an important role in helping researchers to detect DIF.

The Logistic Regression (LR) procedure for DIF was introduced by Swaminathan and Rogers (1990) and can be used in detecting both uniform and non-uniform DIFs, and

can also include exogenous variables in models. Exogenous variables refer to variables such as age besides overall scores that are controlled for in the analysis.

The IRT technique has received more attention as it can model differences in item difficulty and discrimination parameters. It is a collection of models that analyze the individual items with an aim of providing information about their properties. The approach is known to be reliable and has been widely used to detect DIF (Langer et al., 2008). The challenges in variations of items between groups indicate uniform DIF, whereas variations in item discrimination parameters show non-uniform DIF.

A common feature between the LR and IRT procedures is that they are model-based. Both of them can identify uniform and non-uniform DIF for both dichotomous and polytomous items. Identification of the DIF item requires a proper application and understanding of the related SAS procedures.

As mentioned earlier, there are two types of DIF: uniform and non-uniform. Uniform DIF comes about in cases where there is no overlap between the ability level and the group membership. This implies that the likelihood of answering the item accurately is higher for one group compared to the other uniformly over all stages of ability. It is the simplest form of DIF in which the degree of conditional dependency tends to be relatively identical across the latent trait continuum (θ). In this case, the item of concern reliably accords one group an advantage across all levels of ability (θ). In the case of an IRT structure, this can be seen when both item characteristics curves (ICCs) are similarly discriminating and yet show variations in the difficulty units (Walker, 2011).

On the contrary, non-uniform DIF comes about when there is a connection between the level of ability and the membership of the group, implying that the variation in the likelihood of a correct response for the two clusters varies at all levels

of capability. Therefore, in this case, rather than according an advantage to the control group across the ability range, there occurs a variation in the conditional dependency and the variations in direction at various locations in the (θ) range. The non-uniform DIF appear when there is a contact among the level of ability and the membership of the group, which is represented by non-parallel ICC (Walker, Beretvas, & Ackerman, 2001).

Mellenbergh (1982) and Hambleton and Rogers (1989) have documented the occurrence of non-uniform DIF, and according to the findings of Mellenberghn, a Word Analogy Test provided to Tanzanian and Kenyan students had a number of items indicating non-uniform DIF. Hambleton and Rogers, on their part, noted that various items in the 1982 New Mexico High School Proficiency Examination indicated non-uniform DIF during the comparison of Anglo and Native Americans.

In large-scale assessments, it is important for test analysts to identify items that create bias as a function of the characteristics of the examinees (Jensen, 1980; Scheuneman & Bleistein, 1989). Several DIF methodologies have been developed and incorporated to identify both uniform and non-uniform test items. The MH procedure has been indicated to be an effective and popular technique for detecting uniform DIF, but it has also been shown to perform poorly in identifying items with non-uniform DIF (Hambleton & Rogers, 1989).

Several studies have shown that various conditions such as major differences in sample sizes, item performance, and ability distributions between groups of examinees can have tremendous impacts on how often non-uniform DIF is correctly detected with the MH procedure.

There are several considerations during the DIF detection process. The first consideration is the size of the sample, especially regarding the reference and the

focal groups. Before undertaking any analysis, there is a need to have information regarding the number of people in every group such as the ratio of males and females. The sample size is important in this case as it helps in determining if the number of subjects per group is acceptable for there to be sufficient statistical power for identification of DIF. In certain cases, there may occur unequal distribution of the group sizes and so in such a scenario, one would have to find a way of adjusting to eliminate the differences in the size of the reference and the focal groups (Awuor, 2008).

Another factor regarding the size of the sample involves the statistical procedure to be used in the detection of DIF. Other than the considerations of the sample size of the reference and the focal groups, there are other features of the sample that have to be taken into consideration to align with the conventions of each statistical test that is used in the detection of DIF. For example, when using IRT approaches, larger samples may be needed compared to when using the MH approach.

Another aspect that needs to be considered can be determination of the number of items that are being used for the detection of DIF. There is no standard for determining the number of items that should be used but as one moves from one study to another, it may be suitable to test all the items for DIF, whereas in other cases, this may not be essential (Zumbo, 1999).

1.2. Objectives and Research Questions

This research is an extension of the study done by Mazor, Clauser, and Hambleton (1994) in which we will use a predictive model to explain the rate at which a non-uniform DIF item is detected when partitioning subjects by high and low ability levels. In this respect, the study aims to achieve a number of objectives in a bid to understand DIF and approaches that are used to detect DIF. There are three

research questions in this study:

1. How does sample size ratio, item difficulty, and item discrimination affect the detection of non-uniform DIF using the AMH procedure?
2. What factors affect the rate of detection of non-uniform DIF using the AMH procedure?
3. What are the conditions under which AMH procedure works best in detecting non-uniform DIF compared to the BD procedure?

There were several predictors used in these models: item difficulty, item discrimination, and sample size ratios. As an extension to Mazor et al. (1994) work, the results also considered the case where the ability level distributions for the reference and focal groups were equal and unequal. We identified the significant factors that contributed to the detection rates and the conditions in which the highest non-uniform detection rates occurred.

Chapter 2: Literature Review

2.1. Test Bias

Several techniques have come up towards the statistical assessment of DIF. A test is said to show signs of DIF when subjects from different groups possess considerably higher or lower success probabilities on items even after controlling for the overall level of ability. Assessment of the DIF is one of many steps towards the assessment of test bias. Test bias refers to the differential validity of test scores for groups such as, education, sex, and age, to name a few. A systematic error occurs during the process of measurement that affects the scores differentially for a specific group.

Hope, Adamson, McManus, Chis, and Elder (2018) conducted a study in which they used DIF to assess the potential bias in high stakes postgraduate knowledge-based assessment. In their assessment, the researchers noted that despite trying to remove effects that may affect the performance of the candidates, they noted differential attainment of the candidates. This is an important component in research, as it helps establish the constructive validity of tests. Construct validity is the extent to which a test measures what it claims to be measuring.

Many of the techniques for the assessment of DIF were created in educational settings where the items are scored dichotomously as correct or incorrect. MH-based techniques were previously applied in the assessment of DIF. By the early 1990s, it was understood that the LR-based techniques were more powerful compared to the MH-based approaches. This power was accompanied by increased type I-error rates in the LR-based approaches.

The study of test item bias commenced in the late 1960s and continually developed in the subsequent decades. Gómez-Benito, Sireci, Padilla, Hidalgo, and

Benítez (2018) provide more insights into the development of bias in their study on DIF. According to the researchers, this development can be attributed to the deep social psychological and educational effects of the teachings at the time. Later, there was an introduction of the 1974 standards, where the authors attributed the revision as drawing from concerns such as discrimination against certain members of the society, like minority groups and women. Thus, the determining factor in stimulating the study of item and test bias has been social justice in the form of interest in equal treatment of ethnic and socio-economical groups.

Instrument assessment bias has developed into something more solely viewed as a technical issue in psychometric analysis and has thus become a subject of debate in the educational, social and legal areas (Reynolds & Suzuki, 2012). For example, the methods of identifying DIF were developed from the Golden Rule case by Davis-Becker and Buckendahl (2017), that involved screening out items on employment tests, which may be biased against certain subgroups of examinees.

Moreover, Benítez, Padilla, Hidalgo Montesinos, and Sireci (2016) show the existence of DIF through using mixed methods to interpret DIF. The researchers posited that currently, researchers have renewed their attention to equity and fairness in assessment, which comprises a broader conceptualization of the evidence of validity required in the justification of the use of a test for a certain purpose. In modern research, it is becoming more and more difficult to establish a clear distinction between DIF and item bias as the statistical DIF techniques are more sophisticated and the new contexts for DIF studies go beyond the traditional monolingual comparative groups developed by the demographic variables.

2.2. The MH Procedure

The MH technique is a chi-squared contingency table-based method that detect differences between reference and focal groups on all items of the test, one-by-one. The total test scores define the ability continuum, which is divided into k intervals that serves as the basis for matching the members of the two groups. A comparison of both groups at each interval of k is made through a 2×2 contingency table. The group membership (reference or focal) is represented by the rows of the table, whereas the columns represent the correct or incorrect responses. The reference group refers to the group that an item is suspected of favoring whereas a focal group refers to the group where the item is suspected to be functioning differently (Holland & Thayer, 1986).

Even though the MH procedure is one of the most utilized procedures in the determination of DIF due to its simplicity and practicality, it has its own disadvantages. One of its drawbacks is that using this procedure may yield misleading results in the detection of non-uniform DIF or in cases where more complex models are used. This, however, does not affect the usefulness of this procedure (Holland & Thayer, 1986).

Many researchers have used this procedure to examine the relationships between items, and this mainly because it is easy and does not require iterative calculation. Therefore, statisticians have favored this procedure over other procedure.

The data in the form of 2×2×k tables is as shown in Table 1, and is then used in the development of a chi-square test of null hypothesis against the specific alternative hypothesis.

Table 1: Data for the jth matched set of members of R and F

	Score on Studied item		
Groups	Correct (1)	Incorrect (0)	Total
Reference (R)	n_{11i}	n_{12i}	n_{1+i}
Focal (F)	n_{21i}	n_{22i}	n_{2+i}
Total	n_{+1i}	n_{+2i}	n_{++i}

Precisely, MH procedure allows researchers to obtain estimates of the odd ratio ($\hat{\alpha}_{MH}$) across the strata. Odd ratios quantify the strength of the relationship between items. According to Agresti & Kateri (2011), the odd ratio can be estimated by:

$$\hat{\alpha}_{MH} = \frac{\sum_{i=1}^{k}\left(\frac{n_{11i}\, n_{22i}}{n_{++i}}\right)}{\sum_{i=1}^{k}\left(\frac{n_{12i}\, n_{21i}}{n_{++i}}\right)} \tag{1}$$

and the asymptotic MH Chi-square test statistic with one degree of freedom summarized as below:

$$\chi^2_{MH} = \frac{\left[\sum_{i=1}^{k}\left(n_{11i} - E(n_{11i})\right)\right]^2}{\sum_{i=1}^{k} Var(n_{11i})} \tag{2}$$

n_{11i} is the first cell count in each partial table, k is the number of subgroups that are defined on the base of stratification variable. It is actually the consistent estimate of the odds ratio. This therefore means even if there is small data or zeros in

some cells then, one will be able to find the actual number of the $\hat{\alpha}_{MH}$ (Mannocci, 2009). $\hat{\alpha}_{MH}$ is equivalent to odds ratio when there are no confounders, i.e. when $k = 1$. From, the formula of odds ratio, MH estimate will be the weighted average of the subgroup-specific odds ratios, as long as the values of n_{12i} or n_{21i} are not equivalent to zero.

Odds ratios have been applied by a number of researchers who include: (Woolf, 1955; Birch, 1964; Goodman, 1969; Gart, 1970; and Mannocci, 2009). The previously mentioned researchers made assumption that odds ratios were constant in subgroup's 2×2. MH approach is good for detecting DIF since it provides useful comparisons of item performance for different groups. This approach compares subjects of the similar ability level instead of making comparisons on the total group performance on an item.

The MH approach nonetheless does not have the ability to control simultaneously multiple confounding variables. An attempt to control multiple confounding variables will interfere with the size of the strata, consequently, affecting the final results in the long run (Mannocci, 2009).

Since the MH approach is a nonparametric method, it does not build assumptions on a particular form of the item response function (IRF), together with the underlying latent trait distribution. Li (2015) shed more light on the MH formula by deriving the asymptotic power of the test for the DIF. In this study, the author uses the formula to define the performance of the power when the number of items is large, so that the measured latent trait can be regarded as the matching variable in the MH approach. The power relates to the size of the sample, the effect size of DIF, the IRF, and the distribution latent trait in the reference and focal groups. The formula gives an approximation of the power of the MH approach, and consequently offers a guideline

for the DIF detection in practice.

McDonald (2009) provides more insights on when to use the MH test and how the test works. According to the author, the Cochran Mantel–Haenszel (CMH) test is used when there is data from 2×2 tables that has been repeated at different times and locations. The test is to be used for repeated tests of independence, which in most cases occur when one has multiple 2×2 tables of independence for analysis. The test has three categorical variables, in which two variables of the 2×2 test of independence, and the third categorical variable that classifies the repeats such as different location, time or studies. The null hypothesis for the test is that the odds ratios in every repetition are equal to one. Generally, the probabilities are equals when the odds ratio is equal to one and the probabilities differ from each other when the odds ratio is different from one. The probability of rejecting a false null hypothesis is called power. The null hypothesis is, therefore, rejected for extremely large values from 1 under the Chi-Square distribution. This range of value that leads to the researcher rejecting the null hypothesis is known as the region of rejection. The rejection region is obtained in such a manner that the probability is α that is will contain the test statistic at the time when the null hypothesis is true and hence resulting in a Type I error. Normally, the value of α is taken to be small such as 0.01, 0.05, or 0.10 and is known as the level of significance of the test.

Kondratek and Grudniewska (2014) undertake a comparison of the MH procedure with the IRT approach for DIF detection and the effect size estimation. The authors compare the two approaches used in detecting DIF of dichotomously scored items. The results of the study proved that in detecting uniform DIF, the MH approach has more statistical power comparing to likelihood ratio test. Moreover, in case of non-uniform DIF detection, the MH approach has less power than the LR test.

Besides, Magis, Beland, Raiche, and Magis (2018) provide an overview of the different DIF detection techniques. Among the techniques covered in the study are the MH technique, the IRT approach and the BD approach. They provide a summary of different studies that involve the above-mentioned approaches and go further to provide the different factors affecting these approaches, among which are size and the number of items.

Furthermore, Zwick (2012) provides an overview of how different factors, including the sample size, the nature and stringency of the statistical rules used in flagging the items, and the efficacy of criterion refinement, affect DIF analysis. The findings of the study showed that the Educational Testing Service (ETS) C-rule often displays low DIF detection rates even in cases where samples are large. ETS is a service on which the first system of DIF classification for dichotomously scored items was developed. It makes use of the MH delta difference statistic $MH\ D\_DIF = -2.35\ ln(\hat{\alpha}_{MH})$, which is a transformation of the MH constant odds ratio $\hat{\alpha}_{MH}$. This statistic forms the basis of classification of items into three categories, namely category A, that has items with negligible DIF, category B that has items with slight to moderate values of DIF, and category C that has items with moderate to large values of DIF. The study result shows that with improved flagging rules in place, the lowest sample size requirements might be realised. Furthermore, the updated rules for combining data across administrations could help DIF analysis to be carried out in a wider range of situations. The study mentioned that the refinement of the matching criterion enhances the detection rates when DIF is mainly in one direction but can depress detection rates upon balancing of DIF. The study then goes on to recommend refinement in the event that nothing is known about the likely pattern of DIF.

Mazor et al. (1994) tried to determine whether a simple adjustment of the

MH procedure can improve detection rates for items showing non-uniform DIF. The researchers developed an ingenious alternative to the Mantel-Haenszel (AMH) procedure. They suggested that by partitioning examinees based on total test score, particularly separating the examinees into either a 'high' or a 'low' scoring group, one could use the MH procedure to identify non-uniform DIF. The results of their study led the researchers to conclude that the procedure helps increase detection rates without increasing the Type I error rate.

Since its development, studies investigating the AMH procedure have been rare. Fidalgo and Mellenbergh (1995) compared the performance of the AMH procedure to that of the standard MH and iterative logit procedures. They studied how sample size and effect size DIF affects the Type I error rate, power rate, and robustness of the three procedures. They found that the AMH procedure had a higher power rate than the other two procedures, but later cautioned that the AMH procedure was not as robust. This study showed some significant relationships between sample size and DIF effect on the AMH procedure performance, but several limitations surrounding this study warrant further investigation. Their study considered two sample sizes (200 and 1,000). However, conditioned the compared groups to having equal sizes. Additionally, they investigated non-uniform DIF as a composite measure of both difficulty and discrimination. Finally, the researchers investigated symmetric non-uniform DIF, thus leaving readers to question how these factors affect asymmetric non-uniform DIF.

2.3. The BD Procedure

This test was interpreted by Breslow and Day in 1980. They used BD test to do comparison among the exposure groups in a cancer research. Having used BD test in many researches and the cancer research, acknowledges that BD test is flexible

and, as a result, can handle a variety of data configuration. They also found out that BD test can handle wide range of problems that can be approached from conceptual foundation (Breslow & Day, 1980). Flexibility of the BD test is important since it promotes better understanding of the data consequently promoting good identification of the DIF. It is also used to focus on key parameters that are important to understand various groups for proper interventions to be undertaken.

As mentioned earlier, BD is a test for homogeneity of odds ratio. One of the requirements for the BD test to be valid is that the sample size has to be relatively large in each stratum and that atleast 80% of the expected cell counts have to be greater than 5. The BD approach has a more strict sample size requirement compared to the total overall requirement for the MH test table in the sense that the sample size for each stratum, rather than just the sample size, has to be relatively large. This is what limit its usefulness since it will not give a significant result for a small sample. A good test should give a result for any type of test irrespective of whether the sample is small or large. However, the main disadvantage of the approach is that even when it is valid, it may remain weak against certain alternatives (Bagheri, Ayatollahi, & Jafari, 2011).

To effectively deal with the weakness of this approach the entire inference or forecast problem can be put into the setting of an LR model. This can be achieved through investigating whether or not the association with the strata is necessary. A likelihood ratio test through ANOVA function can be used for this. BD test however does not apply when one or two explanatory variables have more than two levels. In such a case, alternative methods or solutions should be sought (Bagheri et al., 2011).

The BD method nonethless is good for analysing non-uniform DIF since it can assess trends in odds ratio heterogeneity. Though the test has numerous weaknesses, it

has helped significantly in detecting DIF. However, compared to other alternatives, the test is less accurate and, as a result, combining it with other tests will give more accurate results. This is important to ensure that the test achieves the results that were intended i.e. detecting DIF. In order to determine the non-uniformity of the DIF, the following formula is used:

$$BD = \sum_{i=1}^{k} \frac{\left(n_{11i} - E(n_{11i}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11i}|\hat{\alpha}_{MH})} \quad (3)$$

Where $\hat{\alpha}_{MH}$ is the MH estimator of common odds ratio, $E(n_{11i}|\hat{\alpha}_{MH})$ is the expected value of n_{11i} and $Var(n_{11i}|\hat{\alpha}_{MH})$ is the variance of n_{11i} under the H_0 of homogeneity of odds ratios. The BD is a test statistic distributed as a Chi-square with $k-1$ degree of freedom (Breslow & Day, 1980). When the deviations between n_{11i} and $E(n_{11i}|\hat{\alpha}_{MH})$ increases, the non-uniform DIF increases, and vice versa. For the BD test to achieve good results, large stratum sizes are required. Otherwise the BD test statistic might give biased results.

The BD test, however, has some disadvantages that makes it unpopular. The global test gives a global statistic, and, as a result, cannot assess the specific alternatives to determine increase or decrease in odds ratios across the ability continuum. Also, it cannnot approximate nominal chi square distribution if there are only a few observations per stratum even if null hypothesis of the homogeneity holds. Notably, BD test has more power if the sample size is larger and less power if the sample size is small (Penfield, 2003).

Furthermore, Magis, Beland, Raiche, and Magis (2018) provide an overview of the different DIF detection techniques. Aguerri, Galibert, Attorresi, and Marañón

(2009) used the BD test to detect non-uniform DIF when the average ability of one group is considerably larger than another group. DIF uses BD test to effectively determine factors that affect rate of detection of non-uniform DIF. The results from the BD test were compared to other methods of detecting non uniform DIF. In particular, the results of the test were compared with LR and the standard MH procedure. It also found that BD was much better compared to the logistic regression (LR) and the MH procedure. It also tested parameters which are the main focus of this study. The parameters that were tested include: sample size and item parameters to effectively determine factors that affect the use of BD test to detect non uniform DIF. It also found that when the item with the largest discrimination and difficulty parameters for equally sized groups was omitted from the goodness-of-fit to the binomial distribution, the test returned Type I error that was similar to the nominal one.

Penfield (2001) performs DIF analysis with BD method using single reference group and multiple focal groups. The study does this by examining several undesirable qualities. In particular, the researcher found that the Type I error rate exceeded nominal level if the individual tests were effectively adjusted. The study examines the drawbacks in the detection of the DIF. The study does this by making comparisons of the performance of different methods.

2.4. Item Response Theory (IRT)

IRT is also known as a latent trait or modern mental test theory. This model is used in the determination of the possible individual latent character through use of the observed total sores on an instrument (Embretson & Reise, 2013). IRT tries to elaborate on the relationship that exists between the latent traits and their manifestations. Latent traits refer to the unobservable characteristics or attributes that

different individuals exhibit when they respond or perform to different environments or conditions. Items and latent constructs including stress, knowledge and attitudes among others, of a particular measure are assumed to be organized in an invisible continuum. Consequently, Item Latent Theory is oriented on the determination of the position of the individual in that series. The correctness of the model is determined through the use of probability of the performance and results vary along the Item response and ICC.

2.4.1. One Parameter IRT Model

In this model, the IRT focuses on one parameter that is used to describe the latent trait of the individual (Chalmers, 2012). The attribute may be denoted as ability (θ). The ability is tested against another parameter of the same item, which is represented as the difficulty (b). One parameter IRT can be represented mathematically as in the equation below:

$$P(X_{ij} = 1|\theta_j, b_i) = \frac{e^{1.7a(\theta_j - b_i)}}{1 + e^{1.7a(\theta_j - b_i)}} \quad (4)$$

Where θ_j refer to the ability level of jth participant and b_i refer to the difficulty level of ith item. In this model, the value of (a) is fixed to the same value across items. Thus, there is no subscript on the (a) parameter (DeMars, 2010) .

In one parameter IRT model, the amounts of information given at a specific ability level equal the inverse of its variance. This implies that the more considerable the amount of information given, the higher the precision level of the test will be. The ability level of an individual is estimated through the comparison of the various probable performance in the response patterns. When the examinees are exposed to

tests, the results tend to give trends that would be predictive to the possible outcomes, since they tend to reach the median probability level, which lowers the possible error levels (DeMars, 2010).

2.4.2. Two Parameter IRT Model

This model is applied to the individuals in the cases where the probability of them successfully answering the test is above 50%. The two-parameter IRT model operates under two main parameters, namely difficulty (b) and discrimination (a). The discrimination parameter value is permitted to vary between various items that are being used in the equation. The lower asymptote's value (c) is fixed to zero. Thus, the model can be represented as follows mathematically:

$$P(X_{ij} = 1|\theta_j, \mathrm{b}_i, a_i) = \frac{e^{1.7a_i(\theta_j - \mathrm{b}_i)}}{1 + e^{1.7a_i(\theta_j - \mathrm{b}_i)}} \quad (5)$$

where a_i refers to the discrimination level of ith item (DeMars, 2010).

The model is served in an ICC, where various selected items intersect in different slope gradients. The discrimination of an object is proportional to the gradient of the slope. This implies that the higher the discrimination level, the steeper the slope will be due to the model's capability to detect even minute differences in the respondent's ability (Rizopoulos, 2006).

2.4.3. Three Parameter IRT Model

Three parameter model, similarly to the one and two parameter models, is used in prediction of the probability of individual responses to be correct. However, this model has an additional third parameter, namely the guessing parameter (c). The model can be represented as follows mathematically:

$$P(X_{ij} = 1|\theta_j, b_i, a_i, c_i) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta_j - b_i)}}{1 + e^{1.7a_i(\theta_j - b_i)}} \quad (6)$$

where c_i refers to the guessing parameter of ith item (DeMars, 2010).

The (c) parameter is responsible for restricting the probability of endorsing correctness of the response as the ability of the respondent's approached infinity. When a respondent answers items by guessing, the information amounts intended to be provided by the item tend to decrease; hence, the respondent's ability is less than the associated difficulty (Chalmers, 2012).

2.4.4. Four Parameter IRT Model

In this model, the underestimate of the respondent is reduced, which is more probable than in the previous models. Upper asymptote parameter (d) creates a room for the high-ability respondents to give a wrong response in an easy task without being underestimated. This model is represented in the following equation:

$$P(X_{ij} = 1|\theta_j, b_i, a_i, c_i, d_i) = c_i + (d_i - c_i)\frac{e^{1.7a_i(\theta_j - b_i)}}{1 + e^{1.7a_i(\theta_j - b_i)}} \quad (7)$$

Where d_i refers to the upper asymptote (guessing) parameter of ith item (Loken & Rulison, 2010).

2.5. DIF Detection Using Bayesian Inference

Bayesian inference is an essential technique in the field of mathematics and statistics where Bayes' theorem is used to validate the hypothesis of a particular situation using probability. The importance of Bayesian inference has been progressively increasing in IRT and other statistical techniques in DIF analysis.

Zwick et al. (1999), compared the empirical Bayes with classical MH

technique for DIF analysis. The main aim of their study was to predict the actual status of samples used using probability. The findings proved that results yielded by Empirical Bayes (EB) were more stable that MH DIF analysis, especially in samples of small quantities.

In an article titled "Using Loss Functions for DIF Detection: An Empirical Bayes Approach", Zwick (2000) used posterior distribution of the DIF experiment to estimate the expected loss of three samples that varied in quantity. Their findings showed an analysis that small quantities presented huge loss function compared to samples of larger numbers.

Moreover, Angel et al. (2007), compared empirical Bayes with standard MH statistics for Detecting DIF under small sample conditions. The main aim of his study was to determine whether samples that used the EB approach performed better than samples that used MH statistics. The results proved that loss functions have no advantage in terms of power, errors realized when MH analysis is used compared to when other analytical methods are used, such as the EB approach.

Soares et al. (2009) shed more light on detecting DIF using Bayesian Models. The main aim of their research was to determine DIF detection. They proved that the DIF analysis, in this case, has variables unlike the previous studies done in the field. This improved the quality of this research since the results obtained are not dependent on a single variable. The findings of this approach affirm that the integrated model that subjects DIF analysis to separate variables is efficient and can be applied in real-life situations because people vary.

2.6. Notations, Derivations and Definitions

2.6.1. Mantel-Haenszel Test

As for the data in 2×2 contigency table (Table 1) and MH procedure, both

margins are fixed. Taking that into consideration, there is only one cell which can vary independently. The focus is on n_{11i} cell without the loss-of-generality. The row margins and column margins, n_{1+i} and n_{+1i}, respectively, are conditioned under the null hypothesis $H_0: X \perp Y$. Thus, the conditional distribution of n_{11i} is given by the row and column margins, which yield the hypergeometric distribution. According to Agresti and Kateri (2011), the formulas related to hypergeometric are as follows:

$$P[n_{11i}] = \frac{\binom{n_{1+i}}{n_{11i}}\binom{n_{++i} - n_{1+i}}{n_{+1i} - n_{11i}}}{\binom{n_{++i}}{n_{+1i}}} \tag{8}$$

Since MH statistics conditioned both margins, the hypergeometric mean is given by:

$$E(n_{11i}) = \frac{(n_{1+i})(n_{+1i})}{n_{++i}} \tag{9}$$

and the hypergeometric variance is:

$$Var(n_{11i}) = \frac{(n_{1+i})(n_{2+i})(n_{+1i})(n_{+2i})}{n_{++i}^2(n_{++i} - 1)} \tag{10}$$

According to Casella and Berger (2002), the mean and variance of the hypergeometric random variable n_{11i} can be derived using the following definitions (See Appendix A):

$$Eg(X) = \sum\nolimits_{x \in X} g(x)\, P(X = x) \tag{11}$$

$$Vg(X) = Eg(X)^2 - \big(Eg(X)\big)^2 \tag{12}$$

As stated earlier, the MH test statistic is as follow:

$$\chi^2_{MH} = \left[\sum_{i=1}^{k} \big(n_{11i} - E(n_{11i})\big)\right]^2 \Big/ \sum_{i=1}^{k} Var(n_{11i})$$

As reported by Rayner and Best (2017), in particular, MH statistic is approximately chi-square as long as at least one of the two conditions are satisfied:

i. If the number of strata, k, is small, then n_{++i} should be large.

ii. If the strata sample sizes n_{++i} are small, then number of k^{th} strata should be large.

This is proven using central limit theorem where the sum of normal will be approximately normal. For this case, if all n_{++i} are large, then each n_{11i} will be approximately normal. Using central limit theorem, this can be shown by:

$$\left(\sum_{i=1}^{k} n_{11i}\right) \sim N\left(\sum_{i=1}^{k} E(n_{11i}), \sum_{i=1}^{k} Var(n_{11i})\right)$$

which is equivalent to:

$$\left(\sum_{i=1}^{k} n_{11i}\right) \sim N\left(E\left(\sum_{i=1}^{k} n_{11i}\right), Var\left(\sum_{i=1}^{k} n_{11i}\right)\right)$$

Thus, Z will be as follows:

$$Z = \frac{\theta - \hat{\theta}}{SE(\hat{\theta})} = \frac{\sum_{i=1}^{k} n_{11i} - E\left(\sum_{i=1}^{k} n_{11i}\right)}{\sqrt{Var\left(\sum_{i=1}^{k} n_{11i}\right)}} \tag{13}$$

Which is equivalent to:

$$Z = \frac{\sum_{i=1}^{k} n_{11i} - \sum_{i=1}^{k} E(n_{11i})}{\sqrt{\sum_{i=1}^{k} Var(n_{11i})}} = \frac{\sum_{i=1}^{k} \left(n_{11i} - E(n_{11i})\right)}{\sqrt{\sum_{i=1}^{k} Var(n_{11i})}} \tag{14}$$

and

$$Z^2 = \frac{\left(\sum_{i=1}^{k} \left(n_{11i} - E(n_{11i})\right)\right)^2}{\sum_{i=1}^{k} Var(n_{11i})} = \chi^2_{MH}$$

Since the square of a standard normal distribution follows a Chi square distribution asymptotically with 1 degree of freedom, we can conclude that the MH test statistic also follows a Chi square distribution asymptotically with 1 degree of freedom under the null hypothesis.

2.6.2. Breslow-Day Test

Similar to MH test statistic, it can be shown that BD test follow Chi square distribution with $k - 1$ degrees of freedom. Consider the following Z formula:

$$Z = \frac{\theta - \hat{\theta}}{SE(\hat{\theta})} = \frac{n_{11} - E(n_{11}|\hat{\alpha}_{MH})}{\sqrt{Var(n_{11}|\hat{\alpha}_{MH})}} \tag{15}$$

and

$$Z^2 = \frac{\left(n_{11} - E(n_{11}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11}|\hat{\alpha}_{MH})} \tag{16}$$

Now consider that $Z_1, \dots, Z_k$ are independent standard normal random variables, then the sum of their squares will be as follows:

$$\sum_{i=1}^{k} Z_i^2 = \frac{\left(n_{111} - E(n_{111}|\hat{\alpha}_{MH})\right)^2}{Var(n_{111}|\hat{\alpha}_{MH})} + \cdots + \frac{\left(n_{11i} - E(n_{11i}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11i}|\hat{\alpha}_{MH})}$$

$$= \sum_{i=1}^{k} \frac{\left(n_{11i} - E(n_{11i}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11i}|\hat{\alpha}_{MH})}$$

In general, it is known that $\sum_{i=1}^{k} Z_i^2$ follows a Chi-square distribution with k degrees of freedom. However, the n_{11} are not completely independent. Thus, if $k-1$ of the n_{11} are known, then the kth is necessarily already determined. The above expression can be reformulated as follows:

$$\sum_{i=1}^{k} \frac{\left(n_{11i} - E(n_{11i}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11i}|\hat{\alpha}_{MH})} = \sum_{i=1}^{k-1} \frac{\left(n_{11i} - E(n_{11i}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11i}|\hat{\alpha}_{MH})} + \frac{\left(n_{11k} - E(n_{11k}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11k}|\hat{\alpha}_{MH})}$$

$$= \sum_{i=1}^{k-1} \frac{\left(n_{11i} - E(n_{11i}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11i}|\hat{\alpha}_{MH})} + \frac{\left(\sum_{i=1}^{k-1}(n_{11i} - E(n_{11i}|\hat{\alpha}_{MH}))\right)^2}{Var(n_{11k}|\hat{\alpha}_{MH})}$$

$$= \sum_{i=1}^{k-1} \frac{\left(n_{11i} - E(n_{11i}|\hat{\alpha}_{MH})\right)^2}{Var(n_{11i}|\hat{\alpha}_{MH})} + \frac{\left(\sum_{i=1}^{k-1} n_{11k} - E(\sum_{i=1}^{k-1}(n_{11k}|\hat{\alpha}_{MH}))\right)^2}{Var(n_{11k}|\hat{\alpha}_{MH})}$$

where we have in last equation $\sum_{i=1}^{k} n_{11k} - E\left(\sum_{i=1}^{k-1}(n_{11k}|\hat{\alpha}_{MH})\right) = 0$. By rewriting the last equation, we will have BD test statistics.

We can conclude that the BD test statistic follows also a Chi square distribution asymptotically with $k-1$ degree of freedom under the null hypothesis (Breslow & Day, 1980).

2.6.3. Three parameter IRT Model

In the current study, the focus is on the three parameter IRT model, since it is used in simulation and application studies to simulate study and non-study items, and to estimate the parameters of items. As mentioned earlier, the three parameter IRT model is as follows:

$$P\left(X_{ij} = 1|\theta_j, \mathrm{b}_i, \mathrm{a}_i, \mathrm{c}_i\right) = \mathrm{c}_i + (1 - \mathrm{c}_i)\frac{e^{1.7a_i(\theta_j - \mathrm{b}_i)}}{1 + e^{1.7a_i(\theta_j - \mathrm{b}_i)}}$$

In the three parameter IRT model, the constant 1.7 was used, so that the scale would approximate the normal metric. The 1.7 is a scaling parameter; it is not necessary, but omitting it would change the scale of the a-parameter. The e in the function is a mathematical constant, the exponential function. Its value is approximately 2.72. Its counterpart is the natural log function; the natural log of $\mathrm{e} = 1$ (DeMars, 2010). The three parameter IRT model is composed of three different settings owing to its name. Each has its specific role in the equation, as discussed below.

2.6.3.1. Discrimination Parameter

The discrimination parameter, denoted as a_i in the three parameter IRT

equation, is also known as the slope parameter. It shows how well the items discriminate against stages of the latent traits (Rizopoulos, 2006). The most discriminative qualities are always situated at the center of the curve. This necessarily, implies that the discrimination level ranges from negative to positive infinity values. Figure 1 shows ICCs derived by varying the (a) parameters in a three parameter IRT model.

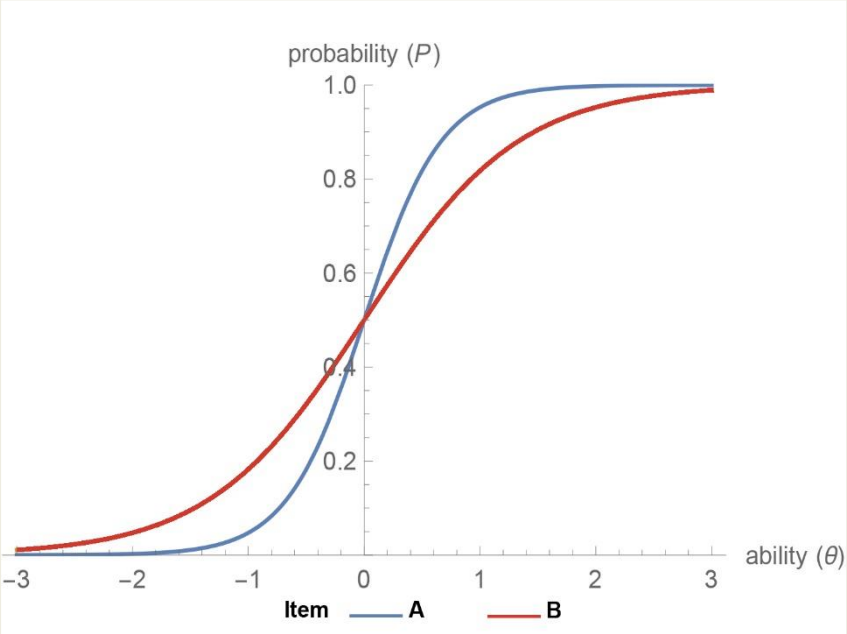

Figure 1: ICCs derived by varying the (a) parameters in a three parameter IRT model.

2.6.3.2. Difficulty Parameter

This parameter is used to predict how hard it is to achieve a 50% correct response at a given ability (Maydeu-Olivares, Cai, & Hernández, 2011). The item on the left have lower difficulty, hence would require fewer skills to answer it correctly. Figure 2 shows ICCs derived by varying the (b) parameters in a three parameter IRT model.

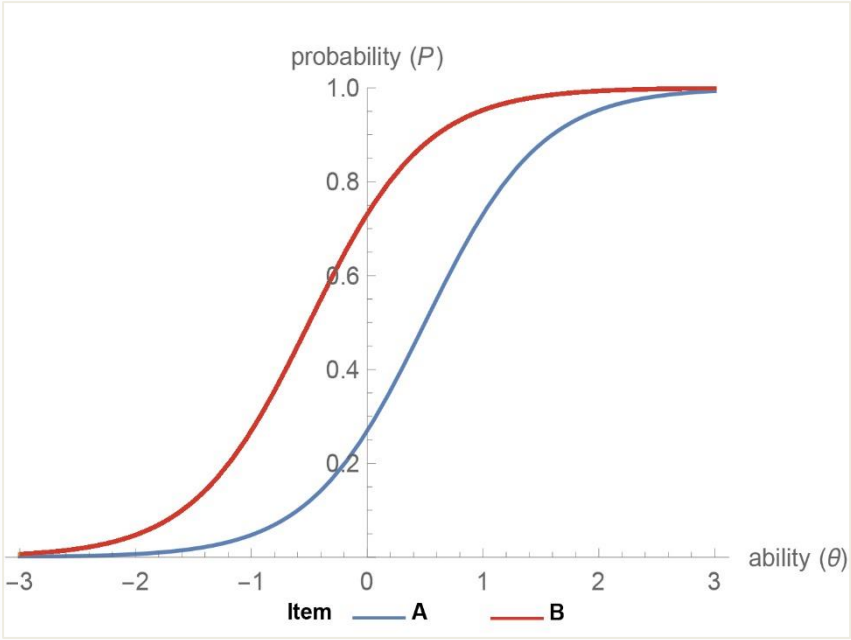

Figure 2: ICCs derived by varying the (b) parameters in a three parameter IRT model.

2.6.3.3. Guessing Parameter

Guessing parameter is commonly applied in the education, and it is also referred to as a pseudo-chance level parameter. The setting is used to predict the level at which the respondent can guess the answer to an item correctly (Maydeu-Olivares et al., 2011). However, the parameter is hardly used in health and personality assessment, since there is no wrong or correct answer in these fields. Figure 3 shows ICCs derived by varying the (c) parameters in a three parameter IRT model.

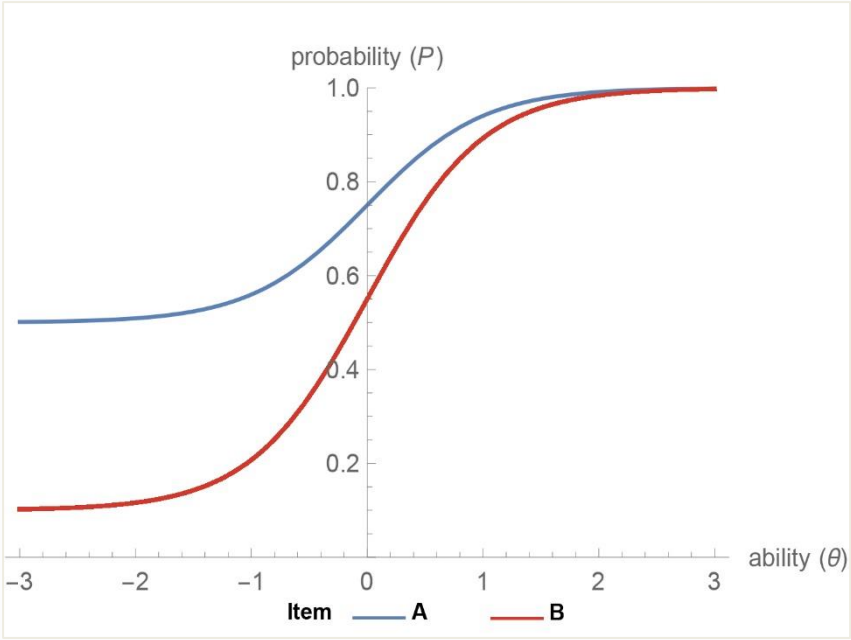

Figure 3: ICCs derived by varying the (c) parameters in a three parameter IRT model.

Chapter 3: Methodology

This study combines methods described in Mazor et al. (1994) and Penfield (2003). There are three processes: simulating the data, assessing the non-uniform DIF items, and creating the regression models. Each step will be explained in the following subsections.

3.1. Simulating the Data

Each simulation contained the item scores for a 75-item examination, in which the last item is created to contain non-uniform DIF representing the studied item. The number of items is considered since it reflected that of similar large-scale assessments. For the non-studied items, difficulty levels will be generated from a normal distribution with a mean of zero and standard deviation of one. The discrimination levels will be generated from a lognormal distribution with mean of 0 and standard deviation of 0.35. These considerations are based on a similar study by Penfield (2003). The guessing parameter for the non-studied items was 0.2. This setting will be used for performance comparisons as well.

Parameters for the studied item are manipulated to create non-uniform DIF based on six combinations between item discrimination and difficulty. Two levels for item discrimination will be tested: low ($a_r = 0.46$ and $a_f = 0.80$) and high ($a_r = 0.70$ and $a_f = 1.97$). Three levels for item difficulty will be considered : easy ($b = -1.50$), medium ($b = 0$), and difficult ($b = 1.50$). These parameters are taken from (Mazor et al., 1994). The guessing parameter, like the non-studied items, is set at 0.2. Three sample size ratios are considered in this study, namely 1:1 ($n_r = n_f = 1000$), 2:1 ($n_r = 1000$, $n_f = 500$), and 5:1 ($n_r = 1000$, $n_f = 200$). These ratios are similar to those from Penfield (2003).

A total of 10,000 simulations will be performed for each of the eighteen combinations of item discrimination, item difficulty, and reference-to-focal group ratio. These simulations are then stratified across two ability level distribution conditions: the first case, where reference and focal groups came from normal distributions with a mean of 0 and standard deviation of 1, and the second case, in which the ability levels for the reference group remained unchanged with the focal group having a mean that is one standard deviation lower than the reference group. This method is similar to both studies by Penfield (2003) and Mazor et al. (1994).

3.2. Assessing the Non-uniform DIF

After creating the simulated item scores, the AMH and BD tests will be performed to determine whether non-uniform DIF is present in the studied item. A significant level of five percent will be considered for both procedures for two main reasons; first, most studies in the domains of quantitative research use five percent level of significance and second, it was proven that the type one error rates were consistent at or below the nominal level of five percent in both cases, whether the group ability distributions were equal or unequal (Penfield, 2003). In the current study, a type one error occurs when DIF is detected while in fact there is no DIF. Mathematically, Type one error can be calculated as follows:

$$\alpha = P(Reject\ H_0: no\ DIF | H_0: no\ DIF\ is\ true)$$

3.3. Creating the Regression Models

Four logistic regression models will be created to assess the likelihood of detecting non-uniform DIF from the studied item. Two models each will assess the DIF using the AMH and BD procedures, where each distinguishes between equal and unequal ability distributions. Each logistic regression model uses item discrimination, item difficulty, and group ratio as explanatory variables to predict

the likelihood of detecting non-uniform DIF. The logit form of the logistic regression model carries the following form:

$$\begin{aligned} logit(p) = \beta_0 &+ \beta_1(group\ ratio) + \beta_2(discrimination) + \beta_3(difficulty) \\ &+ \beta_4(group\ ratio \times discrimination) \\ &+ \beta_5(group\ ratio \times difficulty) \\ &+ \beta_6(discrimination \times difficulty) \\ &+ \beta_7(group\ ratio \times discrimination \times difficulty) \end{aligned}$$

where $logit(p) = \log(p/1-p)$ represents the log of the odds of detecting non-uniform DIF.

Model regression estimates are produced from the logistic model. The Akaike Information Criterion, C-statistic, and correct classification rate will be reported to provide diagnostic statistics of the model.

3.4. Application Study

Several studies have showed that non-uniform DIF can also be utilized in real data sets from both psychological and educational assessments (Teresi & Fleishman, 2007; Woods & Grimm, 2011). Thus, real data will be used in the study to help in systematically investigating the consistencies of DIF detection and the effect size measures in the procedures of detection chosen. For purposes of analysis, the items subjected to the simulations are first organized to form data sets to make it easier to run the simulations and to be able to trace any instances of errors that may arise.

In the current study, the data was obtained from the Department of Student Experience and the Department of Mathematics, Statistics and Physics at Qatar

University. From the student's statistics I exam papers, four multiple choice questions were selected for the purpose of DIF analysis. Moreover, other information about students such as GPA, nationality, major, and gender was obtained in order to classify them into different groups.

Chapter 4: Analysis and Results

4.1. Simulation Study

The detection rate of DIF, which is the percentage of simulations in which the AMH and the BD procedures successfully detected non-uniform DIF in the studied item are shown in table 2 and 3. Both tables reflect percentages as a function of sample size ratios (RF), item discrimination (AL), and item difficulty (BL), but the two tables contrast by the consideration of equal and unequal ability distributions.

Table 2: Non-uniform DIF Detection Rates from Equal Ability Distributions

R:F Ratio	Discrimination	Difficulty	DIF Detection Rate	
			AMH	BD
1:1	Low	Easy	97.87%	47.28%
2:1	Low	Easy	91.34%	34.70%
5:1	Low	Easy	70.27%	22.28%
1:1	Low	Medium	49.95%	43.68%
2:1	Low	Medium	37.40%	29.76%
5:1	Low	Medium	25.86%	18.27%
1:1	Low	Hard	89.65%	18.53%
2:1	Low	Hard	75.16%	14.05%
5:1	Low	Hard	48.88%	9.74%
1:1	High	Easy	100.00%	99.98%
2:1	High	Easy	100.00%	99.16%
5:1	High	Easy	100.00%	90.51%
1:1	High	Medium	99.88%	99.94%
2:1	High	Medium	98.32%	99.22%
5:1	High	Medium	84.07%	87.58%
1:1	High	Hard	99.95%	44.43%
2:1	High	Hard	99.35%	29.69%
5:1	High	Hard	89.87%	19.44%

For reference and focal groups with equal ability distributions, detection rates ranged from 25.86% to 100.00% for the AMH procedure and 9.74% to 99.98% for the BD procedure. Variability of detection rates were higher for the BD procedure than

for the AMH procedure. The three highest detection rates for the AMH procedure occurred with items containing high ALs and easy BLs, while the three lowest detection rates occurrred with items containing low ALs and medium BLs. For the BD procedure, a smaller number of combinations with high detection rates were observed. For the BD procedure, the three highest detection rates were observed where items contained high ALs and easy or medium BLs. The three lowest detection rates were present for items with low ALs, 2:1 or 5:1 RFs, and medium or hard BLs. Regardless of the RFs, both procedures appeared to be most vunerable in detecting non-uniform DIF when the AL was low and the BL was medium.

Table 3: Non-uniform DIF Detection Rates from Unequal Ability Distributions

R:F Ratio	Discrimination	Difficulty	DIF Detection Rate	
			AMH	BD
1:1	Low	Easy	93.03%	47.47%
2:1	Low	Easy	85.37%	31.48%
5:1	Low	Easy	68.29%	21.95%
1:1	Low	Medium	48.90%	36.26%
2:1	Low	Medium	34.91%	24.72%
5:1	Low	Medium	22.95%	14.54%
1:1	Low	Hard	89.40%	11.79%
2:1	Low	Hard	73.06%	9.86%
5:1	Low	Hard	47.14%	7.68%
1:1	High	Easy	100.00%	99.97%
2:1	High	Easy	100.00%	99.23%
5:1	High	Easy	99.87%	89.06%
1:1	High	Medium	99.58%	99.39%
2:1	High	Medium	95.15%	95.63%
5:1	High	Medium	74.62%	73.98%
1:1	High	Hard	99.98%	16.16%
2:1	High	Hard	99.20%	11.87%
5:1	High	Hard	89.73%	7.80%

Similar behavior was observed with the two procedures when the ability distributions were unequal. Detection rates ranged between 22.95% to 100.00% for the AMH procedure and 7.68% to 99.97% for the BD procedure, with more variability in detection rates observed for the BD procedure. The three highest detection rates for the AMH procedure were found in items containing high ALs and easy BLs, but it is interesting to note that some high detection rates were observed even when the BL increased. The lowest detection rates for the AMH procedure were found for the case where the AL was low and the BL was medium, regardless of the RF ratio. For the BD procedure, the highest detection rates were observed for items containing high ALs and easy or medium BLs. The lowest detection rates occurred with items with hard BLs and large RFs (7.58% and 7.80%). The results from both tables suggest that considering equal or unequal ability distributions have little effect on non-uniform DIF detection rates. Rather, the size of the RF and item characteristics are more important factors.

To further understand the behavior and dependence that the AL, BL, and RF have on non-uniform DIF detection rates, four full-effect logistic regression models were created to predict the likelihood of detecting item non-uniform DIF: detecting non-uniform DIF for the AMH procedure with equal ability distributions (Model 4a), detecting non-uniform DIF for the BD procedure with equal ability distributions (Model 4b), detecting non-uniform DIF for the AMH procedure with unequal ability distributions (Model 4c), and detecting non-uniform DIF for the BD procedure with unequal ability distributions (Model 4d). Table 4 contains the estimates and standard errors from the four models, as well as the AIC, c-statistic, and CCR to represent the model diagnostics.

It is important to note that Models 4a and 4c were the focus of this study. The purpose of creating Models 4b and 4d was to compare the significance found in these models to those of Models 4a and 4c, thus identifying concordant and discordant predictor behavior when detecting non-uniform DIF using the AMH versus the BD proedures.

Tables 9 to 12 of the Appendix B contain model-building, main-effect models to show the inclusion of each predictor in the model and to determine whether a predictor's inclusion affect the significance or change in its effect.

Table 3: AMH & BD Full-Effects Logistic Models

Estimates	Equal				Unequal			
	AMH (4a)		BD (4b)		AMH (4c)		BD (4d)	
	β	se_β	β	se_β	β	se_β	β	se_β
Intercept (β_0)	2.89		-0.19		1.32		-0.10	
Reference-Focal Ratio (RF)								
RF = 2:1	-1.26***	0.06	-0.52***	0.03	-0.58***	0.09	-0.68***	0.08
RF = 5:1	-2.62***	0.06	-1.14***	0.03	-1.77***	0.33	-1.17**	0.38
Discrimination Level (AL)								
AL = high	15.32	89.84	8.63***	0.71	17.00	95.12	8.21***	0.58
Difficulty Level (BL)								
BL = medium	-5.86***	0.07	-0.15**	0.03	-3.40***	0.09	-0.46***	0.07
BL = hard	-0.87***	0.17	-1.34***	0.14	1.08***	0.09	-1.91***	0.07
Two-Way Interactions								
RF = 2 × AL = high	1.26	145.10	-3.22***	0.72	-10.61	95.12	-2.57***	0.59
RF = 5 × AL = high	-11.16	89.84	-5.12***	0.72	-13.03	95.12	-4.84***	0.69
RF = 2 × BL = medium	1.17***	0.14	-0.08	0.07	0.12	0.13	0.13	0.09
RF = 2 × BL = hard	0.09	0.17	0.19	0.14	-0.63***	0.10	0.48***	0.09
RF = 5 × BL = medium	2.60***	0.10	-0.10	0.06	1.14***	0.34	-0.34	0.39
RF = 5 × BL = hard	0.30	0.17	0.40**	0.14	-0.70*	0.33	0.69	0.39
AL = high × BL = medium	-15.38	89.84	-0.95	0.82	-15.29	95.12	-2.55	0.59
AL = high × BL = hard	-9.78	89.84	-7.34***	0.72	1.08	134.50	-7.84	0.58
Three-Way Interactions								
RF = 2 × AL = high × BL = medium	-0.94	145.10	1.25	0.84	10.16	95.12	1.11	0.61
RF = 2 × AL = high × BL = hard	-1.91	145.10	2.91	0.74	0.07	134.50	2.41	0.60
RF = 5 × AL = high × BL = medium	11.56	89.84	0.90	0.83	11.95	95.12	2.00**	0.71
RF = 5 × AL = high × BL = hard	8.18	89.84	4.67***	0.73	-0.13	134.50	4.49***	0.70
AIC	69,363.72		82,660.50		86,121.56		91,351.22	
C	0.935		0.895		0.938		0.924	
CCR	85.88%		82.01%		86.61%		88.08%	

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

The AIC for Model 4a was 69,363.72. The c-statistic and CCR of 0.935 and 85.88% respectively suggests that this model possesses strong predictive power. The RF was found to have a very strong negative effect on the detection of non-uniform DIF. Non-uniform DIF items from RFs of 2:1 and 5:1 have lower likelihoods of being detected than non-uniform DIF items from an RF of 1:1, decreasing the log of the odds of detection by 1.26 and 2.62 logits respectively. The behavior of the RF predictor was found to be concordant with that of the RF predictor from Model 4b. The AL was found to be a poor predictor in estimating non-uniform DIF items, and strongly discordant to the behavior of the AL predictor in Model 4b. Significant negative effects were present with the BL predictor. These effects suggest that non-uniform DIF items with hard BLs have lower detection rates than items with easy BLs when using the AMH procedure ($\hat{\beta} = -0.87, p_v < 0.1\%$), but items with medium BLs have even lower detection rates ($\hat{\beta} = -5.86, p_v < 0.1\%$). Little significance was observed with two-way and three-way interactions of these predictors. Significant positive effects present only for the interaction between the RF and BL, particularly for items with medium BLs involving 2:1 RF ($\hat{\beta} = 1.17, p_v < 0.1\%$) or 5:1 RFs ($\hat{\beta} = 2.60, p_v < 0.1\%$). These significant interactions were discordant to the behaviors and significant interactions found in Model 4b with the RF and and AL predictors (negative effects with the 2:1 and 5:1 RF levels with the high AL), RF and BL predictors (positive effect with the 5:1 RF with the hard BL level), the AL and BL predictors (negative effect with the high AL and hard BL), and all three predictors (positive effect with the 5:1 RF, high AL, and hard BL).

Model 4c had an AIC of 86,121.56. The c-statistic and CCR for Model 4c were 0.938 and 86.61% respectively, which was just slightly higher than those calculated in Model 4d. The RF predictor was found to have a strong negative effect

on non-uniform DIF detection. Items created from 2:1 or 5:1 RFs decreased DIF detection rates by -0.58 and -1.77 logits. It is interesting to recognize that the behaviors and significances observed with the RF predictor were similar in Models 4c and 4d, while the standard errors were different. One plausible reason could be due to the differences in the ability distributions. These results were similar to those in Models 4c and 4d where the effects were negative and the standard errors were equivalent. Discrimination was not found to be statistically significant in Model 4c, which was opposite of the effects observed in Model 4d. Strong significant effects were present with the BL on detection rates in Models 4c and 4d. Items with medium BLs decreased detection rates by 3.40 logits, but hard items tend to increase detection rates by 1.08 logits. This was somewhat discordant with what was observed in Model 4d in which medium and hard items exhibited a significant negative effect on the likelihood of detecting non-uniform DIF.

It is interesting to note that there were similarities found in Models 4a and 4c which suggest that some factors contribute to non-uniform DIF detection rates regardless of differences in ability level distributions. It appears that as the ratio of the reference and focal groups' sizes increas, the likelhood of detecting non-uniform DIF with the AMH decreases. Results showed that items created from groups with a 2:1 ratio significantly decreased DIF detection between -0.58 logits and -1.26 logits, while items created from groups with a 5:1 ratio significantly decreased DIF detection between 1.77 to 2.62 logits. Medium items also significantly decrease DIF detection with the AMH procedure between 3.40 and 5.86 logits. However, the interaction of the RF and BL factors had a significant positive effect on non-uniform DIF detection as it increases the likelihood between 1.14 and 2.60 logits.

4.2. Real Application

4.2.1. Sample Characteristics

The sample size for this study were 257 students with average GPA of 2.83. According to Table 5, based on gender, 16.3% of them with GPA average of 2.70 were males; and 83.7% with GPA average of 2.85 were females. Qatari students represent 42.0% of sample with GPA average of 2.73 while non-Qataris represent 58.0% with GPA average of 2.90. Based on their majors, students were classified in two main categories: STEM and non-Stem major. STEM major students represent 32.7% of the total sample while non-STEM students represent 67.3%. The average GPA of STEM and non-STEM major students were 2.91 and 2.79 respectively. Finally, for DIF analysis, students were categorized in two main ability classes according to their GPA. Students with a GPA less than 2.5 were classified as students with low ability levels and students with a GPA 2.5 and more were categorized as high ability level students. The GPA was used because students that earn a 2.5 GPA, in most universities are considered students with good standing and are eligible to enroll in majors and minors (Scheffler, 1992).

Table 4: Demographics of Application Data

Groups	Percentages	GPA
By Gender		
Male	16.3	2.70
Female	83.7	2.85
By Nationality		
Qatari	42.0	2.73
non-Qatari	58.0	2.90
By Majors		
STEM	32.7	2.91
non-STEM	67.3	2.79

Tables 6 cross-classifies students according to their gender, nationality, major and their ability level. 38.1% of male students ability level was low, while 61.9% of them was high. For female students, the percentage of low and high ability was 24.7% and 75.3%, respectively. Furthermore, 32.4% of Qatari students' ability level was low, while 67.6% of them was high. For non-Qatari students, these percentages were 22.8% and 77.2%, consecutively. Finally, 28.6% of STEM students have low ability level and 71.4% of them have high ability level and for non-STEM students, 26% of them have low ability level and 74% of them have high ability level.

Table 5: Ability Level by Gender, Nationality, and Major

Groups		Ability Level	
		Low	High
By Gender	Male	38.1%	61.9%
	Female	24.7%	75.3%
By Nationality	Qatari	32.4%	67.6%
	Non-Qatari	22.8%	77.2%
By Major	STEM	28.6%	71.4%
	Non-STEM	26.0%	74.0%

Figure 4 shows the percentages of correct answers rate of study items according to gender. 73.8% of male and 57.2% of female students answered Item 1 correctly. For Item 2, 71.4% of male and 70.2% of female students answered it correctly. The correct answer rate for Item 3 was 50.0% for males and 61.9% for females. Finally, 76.2% of male and 70.2% of female students answered Item 4 correctly.

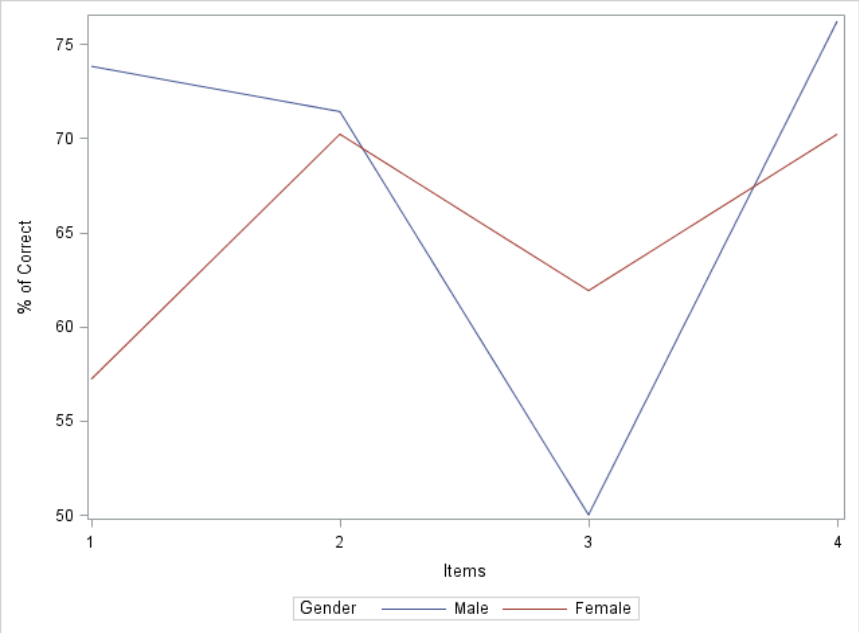

Figure 4: Percentages of Correct Answers Rate of Study Items by Gender

Figure 5 shows the percentages of correct answers rate of study items according to nationality. 54.6% of Qatari and 63.8% of non-Qatari students answered Item 1 correctly. 69.4% of Qatari and 71.1% of non-Qatari students answered Item 2 correctly. The correct answer rate for Item 3 was 56.5% for Qataris and 62.4% for non-Qataris. Finally, in terms of Item 4, 63.9% of Qatari and 76.5% of non-Qatari students answered it correctly.

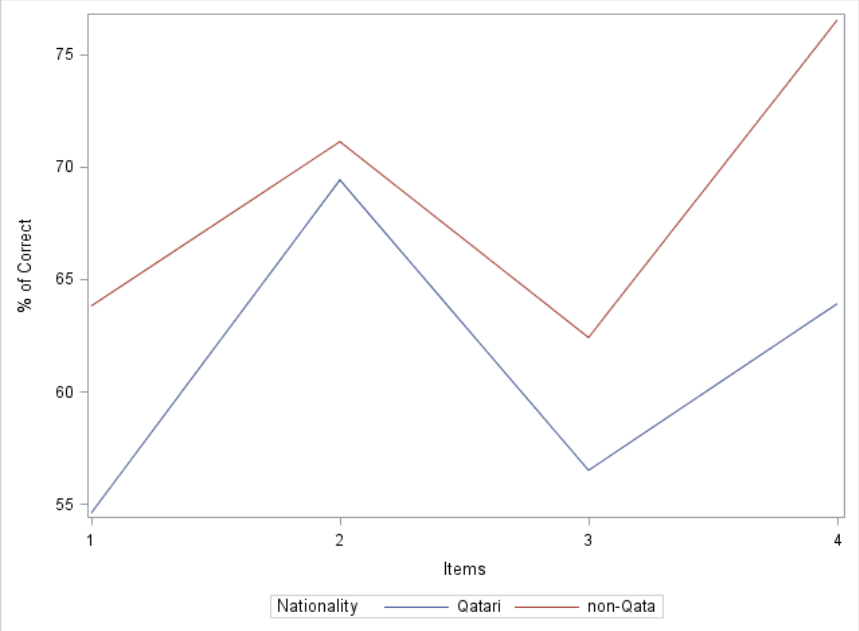

Figure 5: Percentages of Correct Answers Rate of Study Items by Gender

Figure 6 shows the percentages of correct answers rate of study items according to students' major. 58.3% of STEM students and 60.7% of non-STEM students answered Item 1 correctly. For Item 2, 84.5% of STEM students and 63.6% of non-STEM students answered it correctly. The correct answer rate for Item 3 was 61.9% for STEM and 59.0% for non-STEM students. Finally, 76.2% of STEM and 68.8% of non-STEM student answered Item 4 correctly.

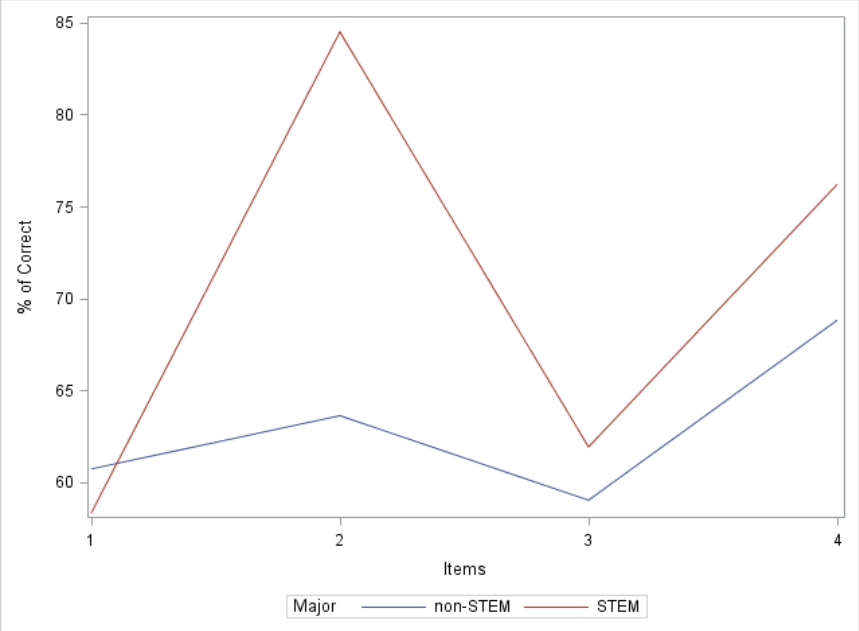

Figure 6: Percentages of Correct Answers Rate of Study Items by Major

4.2.2. Assumptions

In order to apply IRT, several conditions and assumptions must be satisfied. The very basic assumption that should be taken into consideration is that the performance of the students could be predicted based on their abilities. Moreover, the relationship between the probability of answering an item correctly and the ability is directly proportional.

Another required assumption in order to use the IRT procedure is the unidimensionality, which means that there is only one factor that might affect the ability of students. In this study, the GPA of students was the only factor that has been used to measure such an ability. In addition, the criteria of grading students' test from several classes was the same among all instructors.

One last assumption for IRT is the local independence. Local independence means that after taking students' abilities into consideration, there is no association between responses to different items. Pairwise independence between the items can be

proven by using Chi-square test of independence. The following table illustrates the pairwise independence.

Table 6: Testing the Local Independence Assumption

Item Pairs	Chi Square (P-value) *
1 and 2	1.604 (0.205)
1 and 3	0.726 (0.394)
1 and 4	2.255 (0.133)
2 and 3	2.388 (0.122)
2 and 4	2.386 (0.122)
3 and 4	0.883 (0.347)

* P-values are based on 2-sided test

According to Table 6, there is no significant relationship between different pairs of study items. In other words, the study items are independent from each other.

4.2.3. Items' Parameter Estimations

Table 7 shows parameter estimation values for items *guessing*, *difficulty*, and *discrimination* parameters. The guessing parameters are ranged between 0.02 and 0.54. It suggests that respondents with low ability may answer the questions correctly. In this case, Item 1 and 4 have the lowest and Item 2 has the highest guessing parameter.

Moreover, in terms of difficulty level, its ranged between -1.28 and 1.16. Item 4 is the easeist item among others, while Item 3 is the most difficult item compared to other items. Similarly, the discrimination parameters are ranged between 0.37 and 9.12. In general, if an item has the highest discrimination value, it means that an item may discriminate examinees more accurately and clearly. According to Table 7, Item 1 has the lowest discrimination value, whereas Item 2 has the highest value.

Table 7: Items Parameter Estimates

Items	Parameter	Estimate	SE
	Guessing (c)	0.02	0.38
1	Difficulty (b)	-1.03	2.16
	Discrimination (a)	0.37	0.31
	Guessing (c)	0.54	0.14
2	Difficulty (b)	0.39	0.99
	Discrimination (a)	9.12	33.25
	Guessing (c)	0.52	0.12
3	Difficulty (b)	1.16	0.57
	Discrimination (a)	2.49	7.12
	Guessing (c)	0.02	0.37
4	Difficulty (b)	-1.28	1.25
	Discrimination (a)	0.77	0.65

Figure 7 shows the ICCs for four study items. This figure confirms the results of table 7. Its clearly displays that items 2 and 3 have higher guessing parameter than item 1 and 4. Moreover, the slop of green curve, which represents Item 2 is higher than other items. This indicates that the discrimination level of item 2 is higher than other items. In terms of difficulty parameter, the curve of Item 4 reachs highest probabilities faster compared to other items. This indicates that this item is the easiest among all items.

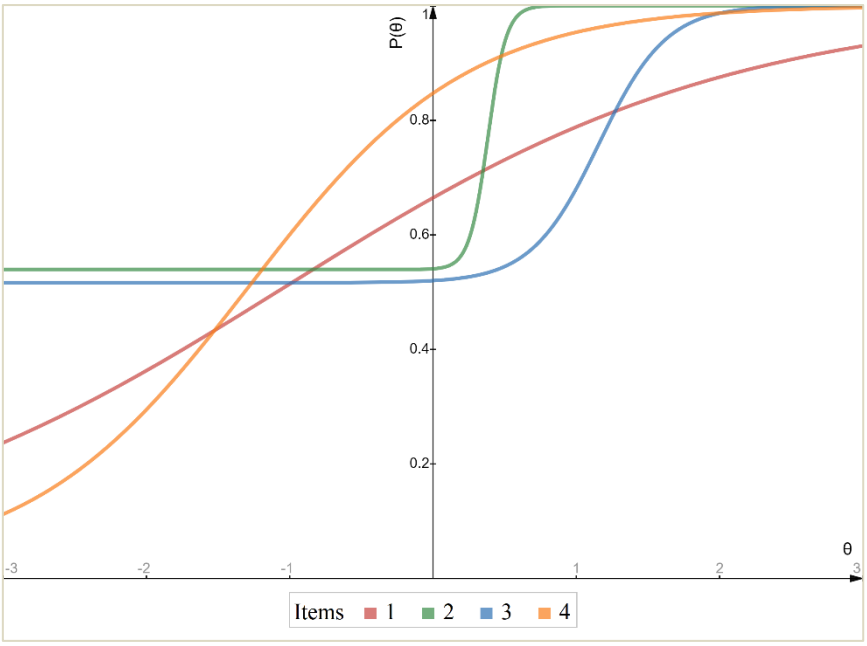

Figure 7: ICCs for Study Items

4.2.4. Detecting DIF

In order to test the null hypothesis of no DIF detection (an Item is not biased), AMH and BD procedures are used. Table 8 shows whether DIF is detected in different secanrios or not. In terms of Item 1 and using gender as group membership variable, DIF is detected using AMH procedure. The Chi square values for AMH was 4.64. It clearly shows that this values is significantly greater than the critical value of Chi square with one degree of freedom, which is equal to 3.84. Moreover, In terms of Item 2 and using major as group membership, DIF is detected using AMH procedure. The Chi square value for AMH is 13.01, which is considerably greater than the critical value of Chi square with one degree of freedom. Similarly, In terms of Item 4 and using nationality as group membership, DIF is detected using AMH procedure. The Chi square value for AMH is 4.19, which is significantly greater than the critical value of Chi square with one degree of freedom.

Table 8: DIF Detection for Different Scenarios

Items	Methods	Gender	Nationality	Major
		Chi Square (P-value) *		
1	BD	0.00 (1.00)	0.53 (0.47)	2.29 (0.13)
	AMH	4.64 (0.03)	1.81 (0.18)	0.11 (0.74)
2	BD	0.94 (0.33)	1.27 (0.26)	1.24 (0.27)
	AMH	0.32 (0.57)	0.01 (0.92)	13.01 (0.00)
3	BD	0.31 (0.58)	0.44 (0.51)	0.14 (0.71)
	AMH	1.68 (0.19)	0.69 (0.41)	0.24 (0.62)
4	BD	1.99 (0.16)	1.40 (0.24)	0.74 (0.39)
	AMH	0.92 (0.34)	4.19 (0.04)	1.62 (0.20)

* P-values are based on 2-sided test

The following figures show the ICCs of three cases that DIF detected successfully. According to figure 8, the male students with low ability level have more chances to answer Item 1 correctly, compared to female students. Moreover, since both curves are not overlapping, which means that the probability of answering Item 1 is uniformly higher for one group compared to the other over all stages of ability level.

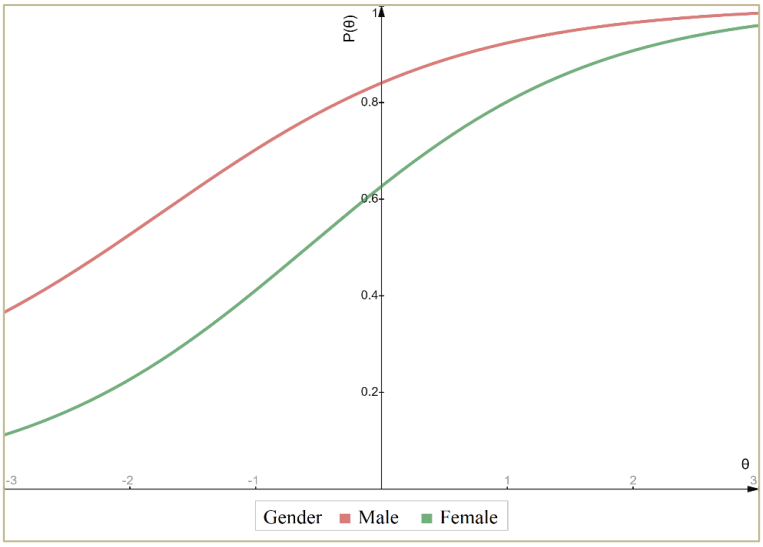

Figure 8: ICCs for Males and Females Respect to Item 1

In terms of Item 2, figure 9 shows that the probability of answering Item 2 correctly is non-uniformly higher for one group compared to the other, over all stages of ability level. This is because both curves are overlapping. In this case, the probability of low ability STEM students to answer Item 2 accurately is higher compared to Non-STEM students.

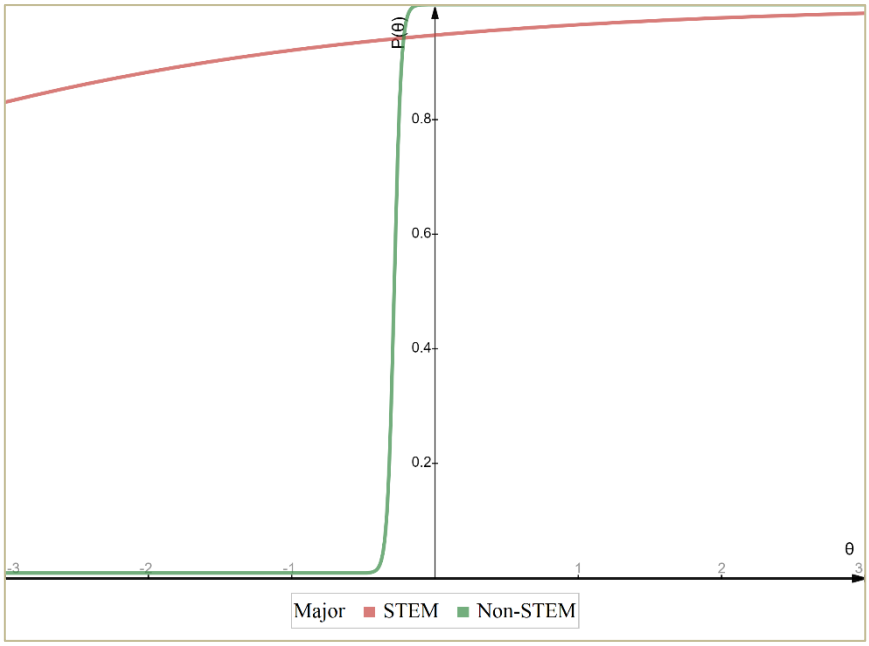

Figure 9: ICCs for STEM and Non-STEM Students Respect to Item 2

Similar to Item 2, in case of Item 4, figure 10 shows that the probability of answering Item 4 correctly is non-uniformly higher for one group compared to the other, over all stages of ability level, as both curves are overlapping. In this case, the probability of low ability non-Qatari students to answer Item 4 accurately is higher compared to Qatari students.

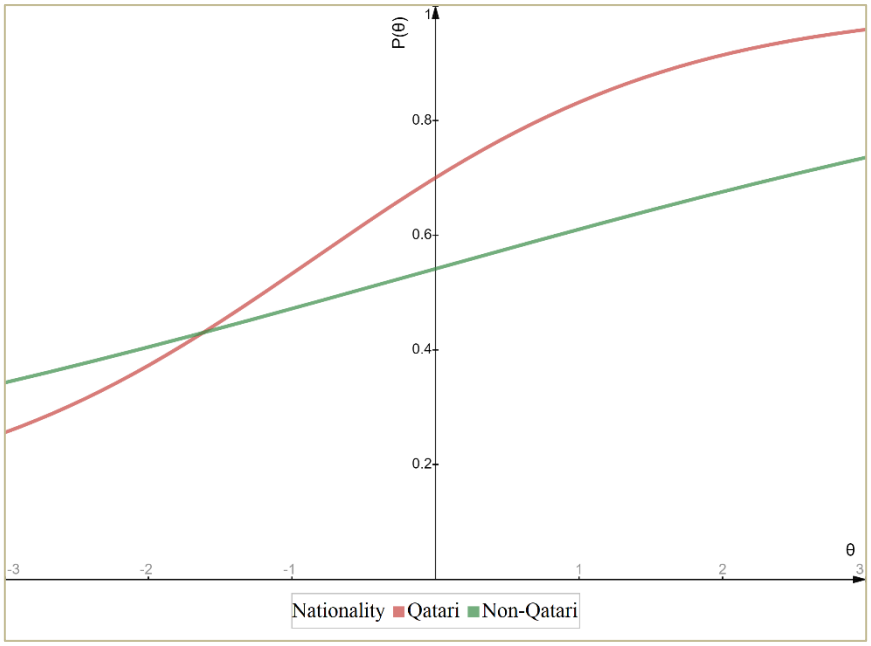

Figure 10: ICCs for Qataris and Non-Qataris Respect to Item 4

Chapter 5: Discussion and Conclusions

The purpose of the study was to determine whether changes in sample size ratios of the reference and focal groups, item discrimination, and item difficulty affect how often the AMH and BD procedures detect non-uniform DIF. It also attempted to find out , which factors significantly affect the detection rate in the AMH procedure, and to determine whether particular combinations of these factors yield higher detection rates with the AMH procedure compared to the BD procedure.

With equal ability distributions, the results suggest that the detection rate of non-uniform DIF using the AMH procedure is most affected by items with high discrimination, then by items with easy item difficulty, and then by items answered by reference and focal group of equal sizes. This was based on comparing the six largest detection rates found in Table 1. Using the AMH procedure, items with high discrimination tend to have 84% chance or more of being detected, 70% or more for items with easy difficulty, and 50% or more for items with equal reference and focal group ratios. Items associated with all three characteristics have the highest chance of being detected by this procedure. To the reader, the order of these factors may be trivial since item characteristics have been shown to play an important part in the detecting non-uniform DIF using several other DIF procedures.

Differences in reference and focal group ratios have a significant negative effect when modeling non-uniform DIF detection rates, with the significance becoming greater as the magnitude of the difference increases. Item difficulty also has a significantly negative effect on non-uniform DIF detection rates with the AMH procedure. Items with medium or hard difficulty exhibit lower chances of being detected, with medium-level items posessing the lowest chance of detection. The negative effect for medium items slightly offsets higher group ratios.

For the unequal ability distribution case, the results are slightly similar. Items with high discrimination, followed by items with easy item difficulty, and then by items answered by reference and focal group of equal sizes give the highest detection rates for nonuni-form DIF with the AMH procedure. Items with high discrimination tend to have about 75% chance or more of being detected, 68% or more for items with easy difficulty, and about a 49% chance or more for items with equal reference and focal group ratios. These percentages are similar to those of the equivalent case, and so it is believed that non-uniform DIF detection is insensitive to unequal ability levels between the reference and focal groups.

Group ratio and item difficulty still possess significant effects on non-uniform DIF detection rates. Group ratio particular has a dominant negative effect, and items involving higher ratios have a stronger negative effect. With regards to item difficulty, medium items have a negative impact on uniform DIF detection rates when compared to easy items, but hard items have a positive impact. This behavior is in contrast to what was observed with the equivalent ability distribution case and is also discordant to the behavior found that one would observe using the BD procedure, in which item difficulty has a completely negative effect.

The results suggest that the effects of item difficulty and group ratios are similar to those that one would observe using the BD procedure, but the effects observed for item discrimination would differ. One possible reason could be that the test statistic formulas involved with the two methods are sample-size dependent. The AMH procedure is a special situation of the MH statistic, and like other IRT-based methods, are affected by item difficulty. With regards to sample size, it is expressed how decreases in group sizes have significant effects on the MH's ability to detect DIF-affected items.

Results from the simulation indicate that the AMH procedure detects non-uniform DIF which is the best for items with easy difficulty levels or high discrimination. Test analysts should expect at least a 2.36:1 odds of successfully detecting non-uniform DIF when either property is present in an item.

In comparison to the BD procedure, the AMH procedure appears to have a higher chance of detecting non-uniform DIF items except when items contain high discrimination and medium difficulty. The AMH procedure yields strongest potential to detect non-uniform DIF items for ones with easy difficulty or high discrimination, and the weakest detection rates for high-discriminating, medium-difficulty items.

Finally, results from the application study showed that out of four study items, AMH procedure detected DIF in three items. Item 1 was student gender-bias, Item 2 was university major-bias, and Item 4 was student nationality-bias. Moreover, the AMH procedure did a better job compared BD procedure in DIF detection. In all three cases, AMH detected the DIF while BD procedure failed to detect DIF. In compatibility with the simulation study, results from the application study showed that compared to the BD process, AMH worked better in DIF detection in both uniform and non-uniform cases.

There are interesting limitations worth noting in regards to this research. First, recall that the application study investigated 257 collegiate students from a major university. This is relatively small compared to the 1,200 to 2,000 observations used in the simulation study. Researchers may extend on this research by using much small sample sizes between the reference and focal groups, and examine how detectability is affected by such change. Another limitation involves the constraints on discrimination. In the simulation study, discrimination was limited between 0.46 and 1.97 between the reference and focal groups. However, rather large

discrimination values were observed in the application data. One interesting exploration for readers would be to examine discrimination at a large level of variability, and determine how it affects AMH and BD DIF detection. It is believed that larger levels of item discrimination would make DIF detection more evident. Furthermore, this research has used the GPA of students to catorise them into low and high ability level, which is considered as limitation. This is due to the fact that there are several other factors that can signify or measure students' ability such as language the and motivation to learn. A final limitation involved the types of DIF tests performed. While the current study tested the effects of several factors on non-uniform DIF detection rate using MH and BD procedures, additional tests can be considered. Comparing MH and BD procedures to other methods such as IRT and LR methods may be of interest to some education analysts.

REFRENCESS

Agresti, A., & Kateri, M. (2011). *Categorical data analysis*: Springer.

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

Angel, M., Fidalgo, K. H., Dave, B., & Jose, M. (2007). Empirical Bayes Versus Standard Mantel-Haenszel Statistics for Detecting Differential Item Functioning Under Small Sample Conditions, The Journal of Experimental Education, 75:4, 293-314.

Awuor, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection: an IRT-based monte carlo study with SIBTEST and Mantel-Haenszel procedures.* Virginia Tech,

Bagheri, Z., Ayatollahi, S. M. T., & Jafari, P. (2011). Comparison of three tests of homogeneity of odds ratios in multicenter trials with unequal sample sizes within and among centers. *BMC medical research methodology, 11*(1), 58.

Benítez, I., Padilla, J.-L., Hidalgo Montesinos, M. D., & Sireci, S. G. (2016). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education, 29*(1), 1-16.

Birch, M. (1964). The detection of partial association, I: the 2× 2 case. *Journal of the Royal Statistical Society: Series B (Methodological), 26*(2), 313-324.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research. Vol. 1. The analysis of case-control studies* (Vol. 1): Distributed for IARC by WHO, Geneva, Switzerland.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2): Duxbury Pacific Grove, CA.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Davis-Becker, S., & Buckendahl, C. W. (2017). *Testing in the Professions: Credentialing Policies and Practice*: Taylor & Francis.

DeMars, C. (2010). *Item response theory*: Oxford University Press.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*: Psychology Press.

Fidalgo, A., & Mellenbergh, G. (1995). *Evaluación del procedimiento Mantel-Haenszel frente al método logit iterativo en la detección del funcionamiento diferencial de los ítems uniforme y no uniforme.* Paper presented at the Comunicación presentada al IV Simposio de Metodología de las Ciencias del Comportamiento. La Manga del Mar Menor.

Gart, J. J. (1970). Point and interval estimation of the common odds ratio in the combination of 2× 2 tables with fixed marginals. *Biometrika, 57*(3), 471-475.

Gómez-Benito, J., Sireci, S., Padilla, J.-L., Hidalgo, M. D., & Benítez, I. (2018). Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*(1), 104-109.

Goodman, L. A. (1969). How to ransack social mobility tables and other kinds of cross-classification tables. *American Journal of Sociology, 75*(1), 1-40.

Scheffler, E. A. (1992). *American universities and colleges*. New York: W. de Gruyter.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel- Haenszel procedure. *ETS Research Report Series, 1986*(2), i-24.

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

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*: Routledge.

Hope, D., Adamson, K., McManus, I., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC medical education, 18*(1), 64.

Jensen, A. R. (1980). Bias in mental testing.

Kondratek, B., & Grudniewska, M. (2014). Comparison of Mantel-Haenszel with IRT procedures for DIF detection and effect size estimation for dichotomous items. *Edukacja Quarterly, 130*(5).

Langer, M. M., Hill, C. D., Thissen, D., Burwinkle, T. M., Varni, J. W., & DeWalt, D. A. (2008). Item response theory detected differential item functioning between healthy and ill children in quality-of-life measures. *Journal of Clinical Epidemiology, 61*(3), 268-276.

Li, Z. (2015). A power formula for the Mantel–Haenszel test for differential item functioning. *Applied psychological measurement, 39*(5), 373-388.

Loken, E., & Rulison, K. L. (2010). Estimation of a four- parameter item response theory model. *British Journal of Mathematical and Statistical Psychology, 63*(3), 509-525.

Magis, D., Beland, S., Raiche, G., & Magis, M. D. (2018). Package 'difR'.

Mannocci, A. (2009). The Mantel-Haenszel procedure. 50 years of the statistical method for confounders control. *Italian Journal of Public Health, 6*(4).

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(3), 333-356.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*(2), 284-291.

McDonald, J. H. (2009). *Handbook of biological statistics* (Vol. 2): sparky house publishing Baltimore, MD.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of educational statistics, 7*(2), 105-118.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*(3), 235-259.

Penfield, R. D. (2003). Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of non-uniform DIF. *Alberta Journal of Educational Research, 49*(3).

Rayner, J., & Best, D. J. (2017). Unconditional analogues of Cochran–Mantel–Haenszel tests. *Australian & New Zealand Journal of Statistics, 59*(4), 485-494.

Reynolds, C. R., & Suzuki, L. A. (2012). Bias in psychological assessment: An empirical review and recommendations. *Handbook of Psychology, Second Edition, 10*.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2*(3), 255-275.

Soares, T. M., Gonçalves, F. B., & Gamerman, D. (2009). An Integrated Bayesian Model for DIF Analysis. *Journal of Educational and Behavioral Statistics*, *34*(3), 348–377.

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

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*(1), 33-42.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment, 29*(4), 364-376.

Walker, C. M., Beretvas, S. N., & Ackerman, T. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. *Applied Measurement in Education, 14*(1), 3-16.

Woods, C. M., & Grimm, K. J. (2011). Testing for non-uniform differential item functioning with multiple indicator multiple cause models. *Applied psychological measurement, 35*(5), 339-361.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann Hum Genet, 19*(4), 251-253.

Zhang, Y. (2015). *Multiple ways to detect differential item functioning in SAS.* Paper presented at the Proceedings of SAS Global Forum 2015 Conference.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1), i-30.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using Loss Functions for DIF Detection: An Empirical Bayes Approach. *Journal of Educational and Behavioral Statistics, 25*(2), 225–247.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel- Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28.

APPENDICES

Appendix A: Mean and Variance of Hypergeometric Random Variable

Mean and variance of the hypergeometric random variable n_{11i} can be derived as follow:

$$Eg(X) = \sum_{x \in X} g(x)\, P(X = x)$$

$$E(n_{11i}) = \sum_{n_{11i}=0}^{n_{1+i}} (n_{11i}) \frac{\binom{n_{1+i}}{n_{11i}}\binom{n_{++i}-n_{1+i}}{n_{+1i}-n_{11i}}}{\binom{n_{++i}}{n_{+1i}}} = \sum_{n_{11i}=1}^{n_{1+i}} \frac{(n_{1+i})\binom{n_{1+i}-1}{n_{11i}-1}\binom{n_{++i}-n_{1+i}}{n_{+1i}-n_{11i}}}{\frac{n_{++i}}{n_{+1i}}\binom{n_{++i}-1}{n_{+1i}-1}}$$

$$= \frac{(n_{1+i})(n_{+1i})}{n_{++i}} \sum_{n_{11i}=1}^{n_{1+i}} \frac{\binom{n_{1+i}-1}{n_{11i}-1}\binom{n_{++i}-n_{1+i}}{n_{+1i}-n_{11i}}}{\binom{n_{++i}-1}{n_{+1i}-1}}$$

Let $\mathrm{j} = n_{11i} - 1$

$$E(n_{11i}) = \frac{(n_{1+i})(n_{+1i})}{n_{++i}} \sum_{j=0}^{n_{1+i}-1} \frac{\binom{n_{1+i}-1}{j}\binom{n_{++i}-n_{1+i}}{n_{+1i}-j-1}}{\binom{n_{++i}-1}{n_{+1i}-1}} = \frac{(n_{1+i})(n_{+1i})}{n_{++i}}$$

$$Var(n_{11i}) = E(n_{11i}^2) - \big(E(n_{11i})\big)^2$$

$$E(n_{11i}^2) = \sum_{n_{11i}=0}^{n_{1+i}} (n_{11i})^2 \frac{\binom{n_{1+i}}{n_{11i}}\binom{n_{++i}-n_{1+i}}{n_{+1i}-n_{11i}}}{\binom{n_{++i}}{n_{+1i}}}$$

$$= (n_{1+i}) \sum_{n_{11i}=1}^{n_{1+i}} \frac{(n_{1+i})\binom{n_{1+i}-1}{n_{11i}-1}\binom{n_{++i}-n_{1+i}}{n_{+1i}-n_{11i}}}{\frac{n_{++i}}{n_{+1i}}\binom{n_{++i}-1}{n_{+1i}-1}}$$

$$= \frac{(n_{1+i})(n_{+1i})}{n_{++i}} \sum_{n_{11i}=1}^{n_{1+i}} \frac{(n_{1+i}) \binom{n_{1+i}-1}{n_{11i}-1} \binom{n_{++i}-n_{1+i}}{n_{+1i}-n_{11i}}}{\binom{n_{++i}-1}{n_{+1i}-1}}$$

Let j $= n_{11i} - 1$

$$E(n_{11i}^2) = \frac{(n_{1+i})(n_{+1i})}{n_{++i}} \sum_{j=0}^{n_{1+i}-1} \frac{(n_{11i}) \binom{n_{1+i}-1}{j} \binom{(n_{++i}-1)-(n_{1+i}-1)}{(n_{+1i}-1)-j}}{\binom{n_{++i}-1}{n_{+1i}-1}}$$

$$= \frac{(n_{1+i})(n_{+1i})}{n_{++i}} E(Y+1) = \frac{(n_{1+i})(n_{+1i})}{n_{++i}} \left(\frac{(n_{1+i}-1)(n_{+1i}-1)}{n_{++i}-1} + 1 \right)$$

$$= \frac{(n_{1+i})(n_{+1i})}{n_{++i}} \left(\frac{(n_{1+i})(n_{+1i}) - (n_{1+i}) - (n_{+1i}) + (n_{++i})}{(n_{++i}-1)} \right)$$

$$Var(n_{11i}) = \frac{(n_{1+i})(n_{+1i})}{n_{++i}} \left(\frac{(n_{1+i})(n_{+1i}) - (n_{1+i}) - (n_{+1i}) + (n_{++i})}{(n_{++i}-1)} \right)$$

$$- \left(\frac{(n_{1+i})(n_{+1i})}{n_{++i}} \right)^2$$

$$= \frac{(n_{+1i})(n_{1+i})(n_{++i}-n_{1+i})(n_{++i}-n_{+1i})}{n_{++i}^2(n_{++i}-1)} = \frac{(n_{1+i})(n_{2+i})(n_{+1i})(n_{+2i})}{n_{++i}^2(n_{++i}-1)}$$

Appendix B: Main Effects Logistic Models

Table 9: AMH Main Effects Logistic Models: Equal Ability Distributions

Estimates	Model 1		Model 2		Model 3	
	β	se_β	β	se_β	β	se_β
Intercept (α)	0.02		0.11		2.37	
Reference-Focal Ratio						
RF_2	0.51***	0.01	0.50***	0.01	-0.68***	0.03
RF_5	-0.41***	0.02	-0.46***	0.02	-1.91***	0.03
Discrimination Level						
AL_H			-0.16***	0.01	2.13***	0.04
Difficulty Level						
BL_M					-6.49***	0.04
BL_T					-0.73***	0.02
AIC	152,078.76		151,921.62		72,911.56	
c	0.599		0.606		0.934	
CCR	57.69%		55.19%		85.88%	

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

Table 10: BD Main Effects Logistic Models: Equal Ability Distributions

Estimates	Model 1		Model 2		Model 3	
	β	se_β	β	se_β	β	se_β
Intercept (α)	0.02		- 0.16		2.37	
Reference-Focal Ratio						
RF_2	-0.83***	0.02	-0.88***	0.02	-0.35***	0.02
RF_5	-1.81***	0.02	-1.50***	0.02	-1.05***	0.02
Discrimination Level						
AL_H			2.91***	0.02	3.16***	0.02
Difficulty Level						
BL_M					0.01	0.02
BL_T					-2.54***	0.03
AIC	142,446.36		103,656.00		90,064.18	
c	0.681		0.854		0.889	
CCR	64.85%		79.59%		80.26%	

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

Table 11: AMH Main Effects Logistic Models: Unequal Ability Distributions

Estimates	Model 1		Model 2		Model 3	
	β	se_β	β	se_β	β	se_β
Intercept (α)	1.18***		0.53***		0.82***	
Reference-Focal Ratio						
RF_2	-0.46***	0.01	-0.52***	0.01	-0.76***	0.02
RF_5	-1.07***	0.01	-1.23***	0.01	-1.79***	0.02
Discrimination Level						
AL_H			1.65***	0.01	2.42***	0.02
Difficulty Level						
BL_M					-2.04***	0.02
BL_T					1.86***	0.02
AIC	224,423.86		200,806.62		145,088.81	
c	0.617		0.741		0.888	
CCR	65.46%		70.62%		80.26%	

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

Table 12: BD Main Effects Logistic Models: Unequal Ability Distributions

Estimates	Model 1		Model 2		Model 3	
	β	se_β	β	se_β	β	se_β
Intercept (α)	0.13***		-0.65***		0.50***	
Reference-Focal Ratio						
RF_2	-0.27***	0.01	-0.32***	0.01	-0.51***	0.02
RF_5	-0.75***	0.01	-0.87***	0.01	-1.40***	0.02
Discrimination Level						
AL_H			1.59***	0.01	2.58***	0.02
Difficulty Level						
BL_M					-0.70***	0.01
BL_T					-4.34***	0.02
AIC	243,495.60		218,229.33		145,181.07	
c	0.582		0.724		0.903	
CCR	53.27%		68.17%		83.63%	

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

Appendix C: SAS Codes for Simulation Study

1. DIF Detection for Equal Ability Distributions

```
%macro doit(nref=,nfoc=,af=,ar=,b=,niters=,case=,al=,bl=,rt=);
data outcomes;
run;
%do times = 1 %to &niters;
data test;
        array a {75} a1-a75;
        array d {75} d1-d75;
        array q {75} q1-q75;
        do j = 1 to 74;
                a{j} = round(exp(0.35*rannor(456789 + &times)),.01);
                d{j} = round(rannor(234567 + &times),.01);
        end;
        d{75} = &b;
        do i = 1 to (&nref + &nfoc);
                ability = round(rannor(123456 + &times),.01);
                if i <= &nref then group = 'f';
                else group = 'r';
                do k = 1 to 75;
                        if k < 75 then do;
                                prob = 0.2 + (0.8)/(1+exp(-1*a{k}*(ability-d{k})));
                                u = ranuni(567890 + &times);
                                if prob >= u then q{k} = 1;
                                        else q{k} = 0;
```

```
                end;
                else do;
                        if group = 'f' then prob = 0.2 + (0.8)/(1+exp(-
1*&af*(ability-d{k})));
                        else prob = 0.2 + (0.8)/(1+exp(-1*&ar*(ability-d{k})));
                        u = ranuni(345678 + &times);
                        if prob >= u then q{k} = 1;
                                else q{k} = 0;
                end;
        end;
        ts = sum(of q1-q75);
        * needed to link mean to observations later;
        _type_ = 0;
        output;
    end;
    drop a1-a75 i j k u d1-d75 q1-q74 prob;
run;
proc means data = test mean noprint;
    var ts;
    output out = ms (drop = _freq_) mean = mu;
run;
proc sort data = test;
    by _type_;
run;
```

```
proc sort data = ms;
        by _type_;
run;
data testa;
        merge test ms;
        by _type_;
        * hscoring = 0 represents low performance group;
        * hscoring = 1 represents high performance group;
        if ts < mu then hscoring = 0;
                else hscoring = 1;
        drop mu _type_;
run;
data testlow testhigh;
        set testa;
        if hscoring = 0 then output testlow;
        else output testhigh;
run;
proc rank data = testa out = testa2 groups = 5;
        var ts;
        ranks strt_mh;
run;
proc rank data = testlow out = testlow2 groups = 5;
        var ts;
        ranks strt_mh;
run;
```

```
proc rank data = testhigh out = testhigh2 groups = 5;
        var ts;
        ranks strt_mh;
run;
proc freq data = testa2 noprint;
        tables strt_mh*group*q75 / cmh2;
        output out = it cmh2;
run;
proc freq data = testlow2 noprint;
        tables strt_mh*group*q75 / cmh2;
        output out = itlow cmh2;
run;
proc freq data = testhigh2 noprint;
        tables strt_mh*group*q75 / cmh2;
        output out = ithigh cmh2;
run;
data myout;
        merge  it
                        itlow (rename=(p_cmhcor = pl_cmhcor p_bdchi = pl_bdchi))
                        ithigh (rename=(p_cmhcor = ph_cmhcor p_bdchi = ph_bdchi));
        keep    _cmhcor_ p_cmhcor p_bdchi pl_cmhcor ph_cmhcor;
run;
data outcomes;
        set outcomes myout;
run;
```

```
data it.outcomes_case&case.;
        set outcomes;
        ratio = &nref./&nfoc.;
        disc_level = &al.;
        diff = &bl.;
        if _cmhcor_ ^= .;
        if p_cmhcor < .05 then detected_amh = 'Y';
                else detected_amh = 'N';
        if p_bdchi < .05 then detected_bd = 'Y';
                else detected_bd = 'N';
        if ph_cmhcor < .05 or pl_cmhcor < .05 then detected_amh2 = 'Y';
                else detected_amh2 = 'N';
        keep p_cmhcor p_bdchi pl_cmhcor ph_cmhcor detected_amh detected_bd
detected_amh2 ratio disc_level diff;
run;
proc freq data = it.outcomes_case&case.;
        tables ratio*disc_level*diff*detected_amh / list;
        title "AMH DIF Item Detection Results: Ratio = &rt.";
run;
proc freq data = it.outcomes_case&case.;
        tables ratio*disc_level*diff*detected_amh2 / list;
        title "AMH DIF Item Detection LOW-HIGH Results: Ratio = &rt.";
run;
```

```
proc freq data = it.outcomes_case&case.;
        tables ratio*disc_level*diff*detected_bd / list;
        title "BD DIF Item Detection Results: Ratio = &rt.";
run;
%doit(nref=1000,nfoc=1000,af=0.80,ar=0.46,b=-
1.50,niters=10000,case=L1E,al='L',bl='E',rt=1:1);
%doit(nref=1000,nfoc=500,af=0.80,ar=0.46,b=-
1.50,niters=10000,case=L2E,al='L',bl='E',rt=2:1);
%doit(nref=1000,nfoc=200,af=0.80,ar=0.46,b=-
1.50,niters=10000,case=L5E,al='L',bl='E',rt=5:1);
%doit(nref=1000,nfoc=1000,af=0.80,ar=0.46,b=0.01,niters=10000,case=L1M,al='L',b
l='M',rt=1:1);
%doit(nref=1000,nfoc=500,af=0.80,ar=0.46,b=0.01,niters=10000,case=L2M,al='L',bl
='M',rt=2:1);
%doit(nref=1000,nfoc=200,af=0.80,ar=0.46,b=0.01,niters=10000,case=L5M,al='L',bl
='M',rt=5:1);
%doit(nref=1000,nfoc=1000,af=0.80,ar=0.46,b=1.50,niters=10000,case=L1T,al='L',bl
='T',rt=1:1);
%doit(nref=1000,nfoc=500,af=0.80,ar=0.46,b=1.50,niters=10000,case=L2T,al='L',bl
='T',rt=2:1);
%doit(nref=1000,nfoc=200,af=0.80,ar=0.46,b=1.50,niters=10000,case=L5T,al='L',bl
='T',rt=5:1);
%doit(nref=1000,nfoc=1000,af=1.97,ar=0.70,b=-
1.50,niters=10000,case=H1E,al='H',bl='E',rt=1:1);
```

```
%doit(nref=1000,nfoc=500,af=1.97,ar=0.70,b=-
1.50,niters=10000,case=H2E,al='H',bl='E',rt=2:1);
%doit(nref=1000,nfoc=200,af=1.97,ar=0.70,b=-
1.50,niters=10000,case=H5E,al='H',bl='E',rt=5:1);
%doit(nref=1000,nfoc=1000,af=1.97,ar=0.70,b=0.01,niters=10000,case=H1M,al='H',
bl='M',rt=1:1);
%doit(nref=1000,nfoc=500,af=1.97,ar=0.70,b=0.01,niters=10000,case=H2M,al='H',bl
='M',rt=2:1);
%doit(nref=1000,nfoc=200,af=1.97,ar=0.70,b=0.01,niters=10000,case=H5M,al='H',bl
='M',rt=5:1);
%doit(nref=1000,nfoc=1000,af=1.97,ar=0.70,b=1.50,niters=10000,case=H1T,al='H',b
l='T',rt=1:1);
%doit(nref=1000,nfoc=500,af=1.97,ar=0.70,b=1.50,niters=10000,case=H2T,al='H',bl
='T',rt=2:1);
%doit(nref=1000,nfoc=200,af=1.97,ar=0.70,b=1.50,niters=10000,case=H5T,al='H',bl
='T',rt=5:1);
```

2. DIF Detection for Unequal Ability Distributions

```
%macro doit(nref=,nfoc=,af=,ar=,b=,niters=,case=,al=,bl=,rt=);
data outcomes;
run;
%do times = 1 %to &niters;
data test;
        array a {75} a1-a75;
        array d {75} d1-d75;
        array q {75} q1-q75;
```

```
do j = 1 to 74;
        a{j} = round(exp(0.35*rannor(567890 + &times)),.01);
        d{j} = round(rannor(345678 + &times),.01);
end;
d{75} = &b;
do i = 1 to (&nref + &nfoc);
        *ability = round(rannor(123456 + &times),.01);
        if i <= &nref then do;
                group = 'f';
                ability = round(rannor(234567 + &times)-0.5,.01);
        end;
        else do;
                group = 'r';
                ability = round(rannor(234567 + &times),.01);
        end;
        do k = 1 to 75;
                if k < 75 then do;
                        prob = 0.2 + (0.8)/(1+exp(-1*a{k}*(ability-d{k})));
                        u = ranuni(678901 + &times);
                        if prob >= u then q{k} = 1;
                                else q{k} = 0;
                end;
                else do;
                        if group = 'f' then prob = 0.2 + (0.8)/(1+exp(-
1*&af*(ability-d{k})));
```

```
                                else prob = 0.2 + (0.8)/(1+exp(-1*&ar*(ability-d{k})));
                                u = ranuni(456789 + &times);
                                if prob >= u then q{k} = 1;
                                        else q{k} = 0;
                        end;
                end;
                ts = sum(of q1-q75);
                _type_ = 0;
                output;
        end;
        drop a1-a75 i j k u d1-d75 q1-q74 prob;
run;
proc means data = test mean noprint;
        var ts;
        output out = ms (drop = _freq_) mean = mu;
run;
proc sort data = test;
        by _type_;
run;
proc sort data = ms;
        by _type_;
run;
```

```
data testa;
        merge test ms;
        by _type_;
        * hscoring = 0 represents low performance group;
        * hscoring = 1 represents high performance group;
        if ts < mu then hscoring = 0;
                else hscoring = 1;
        drop mu _type_;
run;
data testlow testhigh;
        set testa;
        if hscoring = 0 then output testlow;
        else output testhigh;
run;
proc rank data = testa out = testa2 groups = 5;
        var ts;
        ranks strt_mh;
run;
proc rank data = testlow out = testlow2 groups = 5;
        var ts;
        ranks strt_mh;
run;
```

```
proc rank data = testhigh out = testhigh2 groups = 5;
        var ts;
        ranks strt_mh;
run;
proc freq data = testa2 noprint;
        tables strt_mh*group*q75 / cmh2;
        output out = it cmh2;
run;
proc freq data = testlow2 noprint;
        tables strt_mh*group*q75 / cmh2;
        output out = itlow cmh2;
run;
proc freq data = testhigh2 noprint;
        tables strt_mh*group*q75 / cmh2;
        output out = ithigh cmh2;
run;
data myout;
        merge  it
                        itlow (rename=(p_cmhcor = pl_cmhcor p_bdchi = pl_bdchi))
                        ithigh (rename=(p_cmhcor = ph_cmhcor p_bdchi = ph_bdchi));
        keep    _cmhcor_ p_cmhcor p_bdchi pl_cmhcor ph_cmhcor;
run;
data outcomes;
        set outcomes myout;
run;
```

```
data it.outcomes_case&case.;
        set outcomes;
        ratio = &nref./&nfoc.;
        disc_level = &al.;
        diff = &bl.;
        if _cmhcor_ ^= .;
        if p_cmhcor < .05 then detected_amh = 'Y';
                else detected_amh = 'N';
        if p_bdchi < .05 then detected_bd = 'Y';
                else detected_bd = 'N';
        if ph_cmhcor < .05 or pl_cmhcor < .05 then detected_amh2 = 'Y';
                else detected_amh2 = 'N';
        keep p_cmhcor p_bdchi pl_cmhcor ph_cmhcor detected_amh detected_bd
detected_amh2 ratio disc_level diff;
run;
proc freq data = it.outcomes_case&case.;
        tables ratio*disc_level*diff*detected_amh / list;
        title "AMH UNEQUAL DIF Item Detection Results: Ratio = &rt.";
run;
proc freq data = it.outcomes_case&case.;
        tables ratio*disc_level*diff*detected_amh2 / list;
        title "AMH UNEQUAL DIF Item Detection LOW-HIGH Results: Ratio =
&rt.";
run;
```

```
proc freq data = it.outcomes_case&case.;
        tables ratio*disc_level*diff*detected_bd / list;
        title "BD UNEQUAL DIF Item Detection Results: Ratio = &rt.";
run;
%doit(nref=1000,nfoc=1000,af=0.80,ar=0.46,b=-
1.50,niters=10000,case=L1E,al='L',bl='E',rt=1:1);
%doit(nref=1000,nfoc=500,af=0.80,ar=0.46,b=-
1.50,niters=10000,case=L2E,al='L',bl='E',rt=2:1);
%doit(nref=1000,nfoc=200,af=0.80,ar=0.46,b=-
1.50,niters=10000,case=L5E,al='L',bl='E',rt=5:1);
%doit(nref=1000,nfoc=1000,af=0.80,ar=0.46,b=0.01,niters=10000,case=L1M,al='L',b
l='M',rt=1:1);
%doit(nref=1000,nfoc=500,af=0.80,ar=0.46,b=0.01,niters=10000,case=L2M,al='L',bl
='M',rt=2:1);
%doit(nref=1000,nfoc=200,af=0.80,ar=0.46,b=0.01,niters=10000,case=L5M,al='L',bl
='M',rt=5:1);
%doit(nref=1000,nfoc=1000,af=0.80,ar=0.46,b=1.50,niters=10000,case=L1T,al='L',bl
='T',rt=1:1);
%doit(nref=1000,nfoc=500,af=0.80,ar=0.46,b=1.50,niters=10000,case=L2T,al='L',bl
='T',rt=2:1);
%doit(nref=1000,nfoc=200,af=0.80,ar=0.46,b=1.50,niters=10000,case=L5T,al='L',bl
='T',rt=5:1);
%doit(nref=1000,nfoc=1000,af=1.97,ar=0.70,b=-
1.50,niters=10000,case=H1E,al='H',bl='E',rt=1:1);
%doit(nref=1000,nfoc=500,af=1.97,ar=0.70,b=-
```

```
1.50,niters=10000,case=H2E,al='H',bl='E',rt=2:1);
%doit(nref=1000,nfoc=200,af=1.97,ar=0.70,b=-
1.50,niters=10000,case=H5E,al='H',bl='E',rt=5:1);
%doit(nref=1000,nfoc=1000,af=1.97,ar=0.70,b=0.01,niters=10000,case=H1M,al='H',
bl='M',rt=1:1);
%doit(nref=1000,nfoc=500,af=1.97,ar=0.70,b=0.01,niters=10000,case=H2M,al='H',bl
='M',rt=2:1);
%doit(nref=1000,nfoc=200,af=1.97,ar=0.70,b=0.01,niters=10000,case=H5M,al='H',bl
='M',rt=5:1);
%doit(nref=1000,nfoc=1000,af=1.97,ar=0.70,b=1.50,niters=10000,case=H1T,al='H',b
l='T',rt=1:1);
%doit(nref=1000,nfoc=500,af=1.97,ar=0.70,b=1.50,niters=10000,case=H2T,al='H',bl
='T',rt=2:1);
%doit(nref=1000,nfoc=200,af=1.97,ar=0.70,b=1.50,niters=10000,case=H5T,al='H',bl
='T',rt=5:1);
```

1. Logistic Models for Equal Ability Cases

```
data combined;
        set it2.outcomes_caseh1e it2.outcomes_caseh1m it2.outcomes_caseh1t
                it2.outcomes_caseh2e it2.outcomes_caseh2m it2.outcomes_caseh2t
                it2.outcomes_caseh5e it2.outcomes_caseh5m it2.outcomes_caseh5t
                it2.outcomes_casel1e it2.outcomes_casel1m it2.outcomes_casel1t
                it2.outcomes_casel2e it2.outcomes_casel2m it2.outcomes_casel2t
                it2.outcomes_casel5e it2.outcomes_casel5m it2.outcomes_casel5t;
        ratio2 = put(ratio,2.);
run;
```

```
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_amh = ratio2;
        output out = tempout predprobs = i;
        title 'AMH EQUAL: Model 1';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_amh = ratio2 disc_level;
        output out = tempout predprobs = i;
        title 'AMH EQUAL: Model 2';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
```

```
proc freq data = tempout;
	tables is_correct;
run;
proc logistic data = combined descending;
	class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
	model detected_amh = ratio2 disc_level diff;
	output out = tempout predprobs = i;
	title 'AMH EQUAL: Model 3';
run;
data tempout;
	set tempout;
	if _from_ = _into_ then is_correct = 'Yes';
	else is_correct = 'No ';
run;
proc freq data = tempout;
	tables is_correct;
run;
proc logistic data = combined descending;
	class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
	model detected_amh = ratio2|disc_level|diff @3;
	output out = tempout predprobs = i;
	title 'AMH EQUAL: Model 4 (Full Model)';
run;
```

```
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_bd = ratio2;
        output out = tempout predprobs = i;
        title 'BD EQUAL: Model 1';
run
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
```

```
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_bd = ratio2 disc_level;
        output out = tempout predprobs = i;
        title 'BD EQUAL: Model 2';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_bd = ratio2 disc_level diff;
        output out = tempout predprobs = i;
        title 'BD EQUAL: Model 3';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
```

```
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_bd = ratio2|disc_level|diff @3;
        output out = tempout predprobs = i;
        title 'BD EQUAL: Model 4 (Full Model)';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
```

2. Logistic Models for Unequal Ability Cases

```
data combined;
        set it2.outcomes_caseh1e it2.outcomes_caseh1m it2.outcomes_caseh1t
                it2.outcomes_caseh2e it2.outcomes_caseh2m it2.outcomes_caseh2t
                it2.outcomes_caseh5e it2.outcomes_caseh5m it2.outcomes_caseh5t
                it2.outcomes_casel1e it2.outcomes_casel1m it2.outcomes_casel1t
                it2.outcomes_casel2e it2.outcomes_casel2m it2.outcomes_casel2t
                it2.outcomes_casel5e it2.outcomes_casel5m it2.outcomes_casel5t;
```

```
        ratio2 = put(ratio,2.);
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_amh = ratio2;
        output out = tempout predprobs = i;
        title 'AMH UNEQUAL: Model 1';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_amh = ratio2 disc_level;
        output out = tempout predprobs = i;
        title 'AMH UNEQUAL: Model 2';
run;
```

```
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_amh = ratio2 disc_level diff;
        output out = tempout predprobs = i;
        title 'AMH UNEQUAL: Model 3';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
```

```
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_amh = ratio2|disc_level|diff @3;
        output out = tempout predprobs = i;
        title 'AMH UNEQUAL: Model 4 (Full Model)';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_bd = ratio2;
        output out = tempout predprobs = i;
        title 'BD UNEQUAL: Model 1';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
```

```
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_bd = ratio2 disc_level;
        output out = tempout predprobs = i;
        title 'BD UNEQUAL: Model 2';
run;
data tempout;
        set tempout;
        if _from_ = _into_ then is_correct = 'Yes';
        else is_correct = 'No ';
run;
proc freq data = tempout;
        tables is_correct;
run;
proc logistic data = combined descending;
        class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
        model detected_bd = ratio2 disc_level diff;
        output out = tempout predprobs = i;
        title 'BD UNEQUAL: Model 3';
run;
```

```
data tempout;
	set tempout;
	if _from_ = _into_ then is_correct = 'Yes';
	else is_correct = 'No ';
run;
proc freq data = tempout;
	tables is_correct;
run;
proc logistic data = combined descending;
	class ratio2 (ref = '1') disc_level (ref = 'L') diff (ref = 'E') / param = ref;
	model detected_bd = ratio2|disc_level|diff @3;
	output out = tempout predprobs = i;
	title 'BD UNEQUAL: Model 4 (Full Model)';
run;
data tempout;
	set tempout;
	if _from_ = _into_ then is_correct = 'Yes';
	else is_correct = 'No ';
run;
proc freq data = tempout;
	tables is_correct;
run;
```

Appendix D: SAS Codes for Application Study

```
data data;
input Gender $ Items $ Correct;
datalines;
Male    1       73.8
Female1         57.2
Male    2       71.4
Female2         70.2
Male    3       50.0
Female3         61.9
Male    4       76.2
Female4         70.2;
proc sgplot data=data;
series x=Items y=Correct / group=Gender;
yaxis label="% of Correct";
run;
data data;
input Nationality   $ Items $ Correct;
datalines;
Qatari      1      54.6
```

```
non-Qatari      1       63.8
Qatari  2       69.4
non-Qatari      2       71.1
Qatari  3       56.5
non-Qatari      3       62.4
Qatari  4       63.9
non-Qatari      4       76.5;
proc sgplot data=data;
series x=Items y=Correct / group=Nationality;
yaxis label="% of Correct";
run;
data data;
input Major $ Items $ Correct;
datalines;
non-STEM        1       60.7
STEM 1          58.3
non-STEM        2       63.6
STEM 2          84.5
non-STEM        3       59.0
STEM 3          61.9
non-STEM        4       68.8
STEM 4          76.2;
proc sgplot data=data;
series x=Items y=Correct / group=Major;
yaxis label="% of Correct";
```

```
run;

proc freq data=WORK.data;
    table Gender;
    table Nationality;
    table MajororDepartment;
run;
proc sort data = WORK.data;
by Gender;
run;
proc means data = WORK.data mean;
var GPA;
by Gender;
run;
proc sort data = WORK.data;
by Nationality;
run;
proc means data = WORK.data mean;
var GPA;
by Nationality;
run;
proc sort data = WORK.data;
```

```
by MajororDepartment;
run;



proc means data = WORK.data mean;
var GPA;
by MajororDepartment;
run;
proc freq data=WORK.data;
    table Gender*Ability;
    table Nationality*Ability;
    table MajororDepartment*Ability;
run;
proc irt data=WORK.data itemfit out=fscore3PL plots=ICC;
var Q1-Q4;
model Q1-Q4;
run;
proc freq data=data;
    tables Ability*Gender*Q1/ cmh;
run;
proc freq data=data;
    tables Ability*Nationality*Q1/ cmh;
run;
proc freq data=data;
```

```
    tables Ability*MajororDepartment*Q1/ cmh;
run;
proc freq data=data;
    tables Ability*Gender*Q2/ cmh;
run;
proc freq data=data;
    tables Ability*Nationality*Q2/ cmh;
run;
proc freq data=data;
    tables Ability*MajororDepartment*Q2/ cmh;
run;
proc freq data=data;
    tables Ability*Gender*Q3/ cmh;
run;
proc freq data=data;
    tables Ability*Nationality*Q3/ cmh;
run;
proc freq data=data;
    tables Ability*MajororDepartment*Q3/ cmh;
run;
proc freq data=data;
    tables Ability*Gender*Q4/ cmh;
run;
proc freq data=data;
    tables Ability*Nationality*Q4/ cmh;
```

```
run;
proc freq data=data;
    tables Ability*MajororDepartment*Q4/ cmh;
run;
```

Appendix E: Study Items Included in Real Application

Table 13: Study Items Included in Real Application

Items	Description
1	The number of customers entering a bank per minute is a Poisson random variable with a mean of * customers per minute. What is the probability that * customers enter the bank in a minute?
2	The probability that a certain machine will produce a defective item is *. If a random sample of *items is taken, what is the probability that exactly * items are defective?
3	A sample of size * lambs is selected from a production line. The sample mean was * hours and the standard deviation was *. We are interested in testing whether mean life of all lambs exceeds * hours. The value of the test statistic is:
4	If $P(A) = *$, $P(B) = *$, and $P(A \text{ and } B) = *$, then $P(A\|B)$ is: