Comparison of Two Non-IRT Based Multi-Groups DIF Detection Methods' Performances on Type I Error, Power and Precision Rates

By

Ayse Esen

University of Kansas

Submitted to the graduate degree program in the Department of Psychology and Research in Education and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson Dr. William Skorupski

Dr. Jonathan Templin

Dr. Neal Kingston

Dr. Bruce Frey

Dr. Milena Stanislavova

Date Defended: March 29, 2017

The dissertation committee for Ayse Esen certifies that this is the approved version of the following dissertation:

Comparison of Two Non-IRT Based Multi-Groups DIF Detection Methods' Performances on Type I Error, Power and Precision Rates

Chairperson Dr. William Skorupski

Date Approved: March 29, 2017

Abstract

Detecting Differential Item Functioning (DIF) is an early step and very critical to investigate any possible bias between groups (e.g., males vs. females). Many early DIF studies only focused on two-group comparison. However, there are many cases where more than two groups exist: Cross-cultural studies are administered in many countries and any simultaneous comparison across countries is often interest to many researchers. Even for the same administered test in a country, ethnicity is another case for multiple groups. As a need, DIF detection studies among multiple groups have increased as well as the number of DIF methods for multiple groups. Even though there is still not enough study, researchers have compared existing multiple groups DIF detection methods by conducting simulation studies for only uniform DIF items. However, multiple groups DIF detection methods including the Generalized Mantel-Haenszel and the Logistic Regression have not been assessed in any simulation study to determine how well these methods control type I error, power and precision for nonuniform DIF items. This dissertation examined the performance of two non-IRT based multi-groups DIF detection methods on the type I error, power and precision rates for both uniform and nonuniform DIF items. Two methods used in the study are the Generalized Mantel-Haenszel (GMH) and the Generalized Logistic Regression (GLR). A simulation study was conducted in addition to a real data analysis. In the simulation study, total number of groups, groups experiencing DIF, the types and the magnitudes of DIF were manipulated factors. These manipulated factors were considered to get an insight for the real data analysis done in advance. Only dichotomously scored data that was generated by using 2PL IRT model was considered. Total number of items was 50 and first 5 items were DIF items. There were 58 total cases and 168 outcomes of interest. 1,000 iterations were used for each case to ensure the accuracy of results. For all cases, type I error was the ratio of falsely detected non-DIF items over all non-DIF items (45 items), power was the ratio of truly detected DIF items over total number of DIF items (5 items), and precision was the ratio of truly detected DIF items over all detected items (the items that were above the detection threshold).

Type I error, power and precision rates were calculated as the average of 1,000 iterations for cases. The research questions examined in this study were; (1) In investigating uniform DIF, does the magnitude of DIF affect the performance of the GMH and the GLR under different number of total groups and different groups experiencing DIF? What are the type I error, power, and precision rates of these two methods for these conditions?, (2) In investigating nonuniform DIF, does the type of nonuniform DIF affect the performance of the GMH and the GLR under different number of total groups and different groups experiencing DIF? What are the type I error, power, and precision rates of these two methods for these conditions?

The study showed that, for uniform DIF items, the GMH had slightly higher power and precision rates for two group cases. As the magnitude of uniform DIF increased, the power of two methods increased as well for two groups. For nonuniform DIF items with both *a* and *b* parameter change, the results of the GMH was still comparable with the GLR, however, for nonuniform DIF items with only *a* parameter change, the power and the precision rates of the GMH were very low compared to the GLR. In general, when only one group experienced DIF, methods had the lowest power and precision rates and, the highest power and precision rates when it was reference group experienced DIF. For 6 groups, the GHM had higher power rate for uniform DIF and had similar power rates with the GLR for nonuniform DIF with both *a* and *b* parameter change. For 12 groups, both methods had the lowest type I error, power and precision rates when it was medium magnitude of uniform DIF for all cases. Overall, the GLR had better precision rates and lower type I error rates for both 6 and 12 groups. The result indicated that even for nonuniform DIF, the GMH was still able to detect DIF items. However, the type I error rate for both methods were usually above the nominal level of 0.05 with the highest value of 0.2 that meant 9 items were falsely detected. The finding of the simulation study with respect to the type I error rate could explain the findings of real data analysis. All items were found to have DIF with both methods in the real data analysis. However, before concluding that the items is biased toward one group with the real data, further investigation is required by experts. When the power is the only concern, both methods should be used since each of them has its advantages for uniform and nonuniform DIF. However, when the precision is main concern, the GLR is better than the GMH for the majority of cases.

Acknowledgments

All praise and thanks exclusively belong to God, the Lord of worlds. I have many people to thank. First of all, I want to thank my advisor, Dr. William Skorupski, for his guidance, encouragement and support during my study. I have learned a lot from your classes about my dissertation topic. I want to thank my committee members, Drs. Templin, Frey, Kingston and Stanislavova. I specifically thank Dr. Jonathan Templin for his help and support. You were always very kind and friendly, and thank you for your great help with programming. I also specifically thank Dr. Milena Stanislavova for being my committee member after being my master advisor. I also thank Dr. David Magis for answering my questions about the DifR. I want to thank my friends, Anu Sharma, Jessica Loughran and Bo Ho for their friendship, help and support. My special thanks to Aslihan Demirkaya and Gorkem Ozkaya for being a great role model to me and for their guidance and help during my academic study. Aslihan, I admire you as a great academician, a great wife, a great mom and, as a great sister to me. Another special thanks to my closest friend, Esin Yilmaz. You are more than a friend to me. I do also thank so many other friends of me for their help and support during my study.

It would have never been possible to finish my study without the support and love from my family and my fiancé. Thank you to my mom for being with me and always loving us unconditionally. This dissertation is as much as yours as it is mine. You already had three college degrees, one master degree, and one medical degree, and now you have a PhD. You are a great mom! Thank you to my dad for being a strong and great father to us. Thank you to my lovely sisters for being such wonderful sisters and friends. Special thanks to my youngest sister! Yes, it was so difficult to be your guardian here in the USA without even being a parent but your existence made my life so fun and a lot more bearable. Lastly, thank you to my beloved fiancé,

Isa Gunaydin, for loving, supporting and motivating me all the time. There are certain times I could never forget during my doctoral life but it seems this is the end of this journey☺.

This dissertation is dedicated to my grandfather whom passed away during my stay in the USA.

Rest in peace!

Abbreviations

CTT	**:** Classical Test Theory
df	**:** Degrees of Freedom
DIF	**:** Differential Item Functioning
GLM	**:** Generalized Mantel-Haenszel
GLR	**:** Generalized Logistic Regression
ICC	**:** Item Characteristic Curve
IRT	**:** Item Response Theory
LR	**:** Logistic Regression
MH	**:** Mantel-Haenszel
OECD	**:** Organization for Economic Co-operation and Development
PISA	**:** Programme for International Student Assessment
PL	**:** Parameter Logistic
TIMMS	**:** Trends in International Mathematics and Science Study

Table of Contents

List of Figures

List of Tables

Chapter 1: INTRODUCTION

In this chapter, first, I will provide the definition of Differential Item Functioning (DIF) in the literature broadly and define a related term, item bias, in general. Second, I will explain the necessity of studying DIF with both real data and the simulated data. Next, I will mention about two-group DIF studies. Then, I will elaborate the importance of multi-groups DIF studies in the field of educational measurement. Lastly, I will define the purpose of this study and address the significance of the study.

Introduction

Detecting DIF has been one of the never-ending and most studied topics in educational measurement since it is an ongoing challenge to assure that a test is appropriate for the use with subsamples of a general population (Cardall & Coffman, 1964; Svetina & Rutkowski, 2014). To validate the test with respect to every recognizable subgroup in the population and, to develop a subgroup specific analytical guide would be very difficult (Cardall & Coffman, 1964). However, the case of the same item's functioning differently for different groups was found important to study and essential to investigate the reasons behind it with respect to test questions and the backgrounds of the different groups of examinees (Holland & Thayer, 1988; Magis & De Boeck, 2011). When an item is written, it is expected that the examinees with same ability will have equal probability of answering the same question correct regardless of group membership such as gender or any ethnicity background (e.g. white Americans and African Americans). In this way, the result will truly reflect the ability of the examinees on the construct being measured by the test with an unsystematic error. In contrast, any interaction found between group membership and the item will indicate that the item is biased toward one group and examinees are at a disadvantaged position due to the group they belong. As a result, the scores of examinees will not be a true indicator of their ability with respect to what test measures and it will threat the validity of the assessment regarding test fairness. Initially, the term “item bias” was used instead of DIF to indicate the group differences regarding the items’ functioning differently between groups (Flier, Mellenbergh, Adèr, & Wijn, 1984; Ironson & Subkoviak, 1979; Scheuneman, 1979). In this study, two terms, DIF and item bias, are used separately. Their definitions and nuances in their meaning are explained in the next chapter in detail. For now, DIF is simply described as the statistical properties difference of an item after controlling the differences in the abilities of the groups (Angoff, 1993). It is not having equal probability of correct answer even if people from different manifest groups have the same level of ability when ability scores are on a common scale (De Ayala, 2009). Item bias is further investigation of DIF and it has both a social and statistical meaning in it (Angoff, 1993). DIF is the earliest statistical step that should be conducted before claiming any biasness of an item. When a real data analysis is done, any items that are labeled as having DIF are possible biased items. However, these detected DIF items should be carefully explored before claiming any biasness (Camilli & Shepard, 1994) and the identification and even the removal of them are a necessity for valid conclusions (Magis & De Boeck, 2011). Even though a real data analysis was done in this study and all items were found to be DIF items, it was beyond the scope of this dissertation to investigate the reasons and sources of DIF for these items.

Studying DIF and having assessments with DIF free items are vital to ensure the fairness of the assessment toward subgroups in the population and test fairness in an important component of the validity of test scores’ use and interpretation (Messick, 1990). Violation of test fairness indicates that items within an assessment cannot accurately measure the construct. One way to ensure test fairness is to have assessment with DIF free items. Nevertheless, besides analyzing real data with this purpose by using the statistical methods developed, it is also important to investigate the performance of these DIF methods on identifying DIF items accurately when they exist and not labeling any items as having DIF when items do not have DIF. For this reason, simulation studies are conducted by considering different factors such as different levels of uniform DIF, nonuniform DIF, sample size of groups, impact (true mean differences between groups) and so on to know how well methods perform (these terms will be explained in the next chapter).

Most commonly two groups are compared in DIF studies (Camilli & Shepard, 1994; Holland & Thayer, 1988; Lord, 1980; Swaminathan & Rogers, 1990; Thissen, Steinberg, & Wainer, 1988; Zumbo, 1999). Since gender naturally has two groups, it has been intensively studied in the literature as well to see if there is any gender difference on the construct being measured by the assessment (Kalaycioğlu & Berberoğlu, 2011; Smith & Reise, 1998; Walstad & Robson, 1997). Hence, there is tremendous number of two-group DIF studies in the literature. Therefore, many methods have been developed for two-group comparison for DIF, and their performances have been compared a lot (See next section). Two groups that are compared in DIF studies are called *reference group* and *focal group*. The bigger group or the main group in the population is the reference group; whereas, the smaller group or the minor group in the population is the focal group, that is of primary interest (Holland & Thayer, 1988). However, there are many cases where more than two groups exist. Simply, more than one ethnic group may exist in the population or even researchers may be interested in school differences or state differences for the same administered assessments by considering each of them as a separate group (Magis, Raîche, Béland, & Gérard, 2011). At the international level, large-scale assessments (cross-cultural studies) are administered at many different countries. For example, assessments such as Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) have been administered in more than fifty countries in 2015.

Just like two-group cases, it is very important to establish fairness among multiple groups to ensure the meaningful comparison among these groups. Recently, international level large-scale assessments (like the ones mentioned above and many others) have captured special attention from scholars and researchers who want to compare and contrast the educational qualities of various countries. For example, the PISA results have been used as the benchmarks of quality of national educational system world widely (OECD, 2011). For this reason, it is critically important that these assessments are fair among countries (Klieme & Baumert, 2001; Poortinga, 1989) by ensuring that tests are equivalent both linguistically and culturally across participating countries. Any comparisons of these large scale assessments among different groups or countries are only possible when it is assured to have equivalent scale from the instrument and the scaling process (Hui & Triandis, 1985; Hulin, 1987). The attribute that is being compared among groups and the scale units that are used must be same to be able to make any comparison among groups (Poortinga, 1989; Van de Vijver & Tanzer, 1997). Issues related to translation of same test into multiple languages and the ignorance of cultural differences among countries during translation threat the construct validity and cause DIF (further item bias) among groups (Ercikan, 1998; 2002). To avoid any possible problems with respect to test items that may cause disadvantages between groups, items should be investigated before the test is administered. However, even after the test is administered, DIF studies should be conducted among multiple groups to make a valid comparison and prevent any unfair judgment with respect to test scores of examinees. With all these possible sources of DIF with multi-groups and the necessities of detecting DIF among multi-groups DIF cases, the number of DIF studies with multiple groups has increased during the last few decades (Ellis & Kimmel, 1992; Kim, Cohen, & Park, 1995; Magis & De Boeck, 2011; Magis et al., 2011; Oshima, Wright, & White, 2015; Penfield, 2001; Sari & Huggins, 2015). However, there is still very limited number of studies to simultaneously assess DIF across multiple groups.

Statement of Problem

As Magis and De Boeck (2011) discussed in their paper, there is a real need for multiple group DIF methods studies in the field of measurement. Even though multi-groups DIF has been taken into consideration in the relevant literature, many early and current studies have either done pairwise comparison across all groups or used the composite of groups as one group and done comparison between two groups instead of simultaneous comparisons among groups (Ellis & Kimmel, 1992; Sari & Huggins, 2015). Nevertheless, doing a pairwise comparison provokes the Type I error inflation and Bonferroni correction is needed for further investigations between groups (Kim et al., 1995; Penfield, 2001). Decomposing all focal groups into one group controls for Type I error inflation, but it may cause the misidentification of items as DIF or non-DIF especially when one focal group is a lot smaller than other focal groups (Magis et al., 2011). Oshima et al. (2015) also stated that when minority groups are lumped, it could obscure existing DIF when one of the minor groups is in advantage when others are in disadvantage.

Another issue is that, especially for these cross-cultural studies there may not be a specific reference group to be compared with the focal group of interest. In contrast, cross-cultural studies like PISA and TIMMS, all groups (examinees from each country) are almost same size so that no groups are minor or major. In addition, it is not clear which groups are of interest with respect to the DIF study since the translated version of items is used. The translation of items may affect the difficulty of vocabulary for the item and it may be interpreted differently between groups even though the item is aimed to measure the same construct (Poortinga & Van de Vijver, 1987). For this reason, instead of doing pairwise comparison between any two groups, a simultaneous comparison among groups is more reasonable and beneficial.

To deal with these shortcomings, many existing DIF detection methods that have been commonly used for two groups have been extended or further developed for simultaneous detection of DIF in multiple groups (Fidalgo & Scalon, 2009; Kim et al., 1995; Magis et al., 2011; Penfield, 2001; Woods, Cai, & Wang, 2013) in addition to developing new methods (Magis & De Boeck, 2011). However, as Magis and De Boeck (2011) indicated that there are not enough study conducted yet; these multi-groups DIF methods have to be compared on their performance via simulation studies by considering the same factors mentioned for two group simulation studies in addition to doing real data analysis. Penfield (2001) compared three Mantel-Haenszel procedures including the Generalized Mantel-Haenszel Statistic (GMH) to assess DIF among multiple groups in his simulation study. He considered many factors in his study including different levels of focal groups experiencing DIF. However, only two levels of magnitude of uniform DIF were considered. Finch (2015) compared the performance of four multi-group methods by considering only uniform DIF. However, nonuniform DIF is also common in real data and it has been considered a lot in two group DIF studies (Güler & Penfield, 2009; Mazor, Clauser, & Hambleton, 1994; Rogers & Swaminathan, 1993; Woods, 2009). For this reason, this factor should be more considered in multi-groups DIF simulation studies. Precision is another outcome that is never mentioned in any DIF studies even though type I error and power have been studied a lot with respect to the methods' performance. Precision is about the accuracy of truly detecting DIF items and it should be investigated as well. Instead of investigating power and type I error separately and individually, precision helps determining exactness at once.

The Purpose of the Study

The present study compares the performances of two non-IRT based multi-groups DIF detection methods on type I, power and precision rates via a simulation study guided by a real data analysis. The purpose of real data analysis is to investigate the factors to be manipulated in the simulation study. The purpose of simulation study is to contribute to the current limited literature cited above by comparing the GMH and the Generalized Logistic Regression (GLR) performances on accurately identifying both uniform DIF and nonuniform items (power), not to label any items that do not have DIF (inverse of type I error) and accuracy of detecting true DIF items (precision) for dichotomously scored items. Even though early studies compared the performance of these two methods for two-group cases (Hidalgo & LÓPez-Pina, 2004; Narayanon & Swaminathan, 1996; Swaminathan & Rogers, 1990), there is no study comparing them for multiple groups with respect to nonuniform DIF. In his study, Finch (2015) only considered uniform DIF and suggested researchers to consider nonuniform DIF for future researches.

The Significance of the Study

The current study considered both uniform and nonuniform DIF items besides considering a wider range of DIF magnitudes as Penfield (2001) suggested and Finch (2015) did in his study for uniform DIF items. Penfield (2001) also recommended the use of the GMH for future studies. More importantly, the Mantel-Haenszel and the Logistic Regression are still the most commonly used DIF detection methods in the literature (Mannocci, 2012). Both are non-IRT based and the most important advantage of both methods is that they do not require any specific model fit for data and use summed scores as the criteria to detect DIF. Their assumptions are easily met with any data, the process is not complicated and the interpretations of statistical results found from analyses are easy. Hence, it is also important to compare their performance for multiple groups by considering a wide range of total numbers of groups, both types of DIF with a wide range of magnitude, and different focal groups experiencing DIF. For any large-scale assessment, it is very like that some focal groups may share cultural or linguistic background together.

Chapter 2: LITERATURE REVIEW

In this chapter, first, Classical Test Theory and Item Response Theory (IRT) will be summarized. Then some important terms related to DIF and types of DIF will be defined. Next, the overview of two-group DIF detection methods for dichotomously scored items will be provided and general steps for the analyses will be explained. Then current multi-groups DIF studies will be summarized with emphasis on the scarcity of research on DIF with multiple groups in the literature. Finally, two non-IRT based multi-groups DIF detection methods; the Generalized Mantel-Haenszel and the Generalized Logistic Regression will be given in details.

Classical Test Theory

Classical Test Theory (CTT) is based on the true score model and, the interest is the nature and the characteristics of the measure, not the individuals. It determines the amount of error within a test (Raykov & Marcoulides, 2011). Examinee's observed score (raw score or summed score) is a combination of his true score and some random error term related to measure (DeVellis, 2006). The equation for CTT is defined as

$$X_i = T_i + E_i \tag{1}$$

where X_i is the observed score of the examinee i, T_i is the expected score of the examinee when the same test is theoretically administered infinite times (what the examinee is expected to have in the ideal world) and, E_i is the error score or error of measurement for the same examinee (Crocker & Algina, 1986). Once the examinee's true score is estimated, the amount of error for the test examinee is given is calculated by taking the examinee's true score and subtracting it from the raw score (Crocker & Algina, 1986). The total score (summed score) of the examinee for dichotomously scored data is calculated as the sum of correct answers (sum of 1s). In many DIF studies, this summed score is used as the matching criteria.

There are two important components of CTT: item difficulty and item discrimination. Item difficulty is determined by the ratio of the number of examinees who answer the item correctly over total number of examinees (Kline, 2005). Item difficulty ranges between 0 and 1, and any item difficulty that is closer to 1 is considered easier item; whereas, the item difficulty that is closer to 0 is considered more difficult item. Item discrimination is the item- total score correlation and it is point bi-serial correlation for dichotomously scored items. Even if the same item is administered to different groups, item difficulty and item discrimination are sample dependent (Hambleton & Jones, 1993). Whether an item is hard or easy depends on the ability of the examinees being measured, and the ability of the examinees depends on whether the test items are hard or easy (Hambleton, Swaminathan, & Rogers, 1991).

There are three important assumptions to the CTT. The first one is that the mean of the error scores for any population of examinees have to be zero (Crocker & Algina, 1986). The second one is that the correlation between any population of examinee's true score and its error score have to be zero, that means these two are independent from each other (Crocker & Algina, 1986). The third one is that the correlation between two error scores for the two parallel tests administered to a population at different times has to be zero (Crocker & Algina, 1986). Even though these assumptions are easily met by any data, there is an inseparable interdependence between item characteristic and the examinee's total score and, there are reported on the different scales. These drawbacks made researchers search new models and Item Response Theory was a result of it (Hambleton & Jones, 1993).

Item Response Theory

Item Response Theory (IRT) is based on the mathematical models that place the items and examinees' ability on the common scale. This allows the comparison of trait (what is measured by the test) levels (Embretson & Reise, 2013). In contrast to CTT, the interest is each individual item, not the test itself. Items have their own characteristics regardless of the individual examinees in the sample. Item Characteristic Curves (ICC) is the mathematical graph allowing us to see the relationship between examinee's ability and the probability of answering the item correctly. Examinee's ability is represented by theta (θ) (Hambleton et al., 1991) within IRT and although it ranges from -∞ to +∞, its typical reported range is from -3 to +3 or -4 to +4 (Raykov & Marcoulides, 2011). The value of 0 represents the average ability with positive value representing a higher ability and negative value representing a poorer ability. Unlike to CTT, item difficulty and item discrimination are not sample dependent in IRT. Depending on the IRT models explained below, these components are used. Briefly, item difficulty is directly related to the ability level. Any difficult item requires a higher level of ability; whereas, any easy item requires a lower level of ability to be answered correctly. Item discrimination is about the item's distinguishing examinees with low and high ability. It assesses how much the item can differentiate among different ability levels (De Ayala, 2009).

Dichotomous IRT Models

There are three most commonly used IRT models when the items have only two scoring options (scored as incorrect or correct; "0" or "1"). These models are 1) the one-parameter logistic regression IRT model (1PL; Rasch (1960)); 2) the two-parameter logistic IRT model (2PL; Birnbaum (1968); and 3) the three-parameter logistic IRT model (3PL; Birnbaum (1968)).

All three models describe the nonlinear relationship between examinee ability, θ, and the probability of correct response (Hambleton et al., 1991) up to three parameters: difficulty, discrimination and guessing depending on the model.

1 Parameter Logistic Model

One parameter logistic (1 PL) model is the simplest IRT model used for dichotomously scored data and it is often considered to be "Rasch Model". It only gives information about the difficulty of the individual item. An easier item means that less ability is required to have a high probability of answering an item correct; whereas, a difficult item means high ability is required to have a high probability of answering an item correct. For difficult items, ICC shifts to the right and for easy items, it shifts to the left. Discrimination and guessing parameters of the individual items are not estimated in the model. The formula for 1 PL IRT model is given by Equation (2) where $P_{ij}(\theta)$ is the probability of person i answering the item j correctly, θ_i is the ability of the person, b_j is the difficulty of item j, e is the base of the natural logarithm ($e \approx 2.72$).

$$P_{ij}\,(\theta)= \frac{e^{(\theta_i - b_j)}}{1+ e^{(\theta_i - b_j)}} \tag{2}$$

2 Parameter Logistic Model

Unlike to 1 PL IRT model, two-parameter logistic (2 PL) model considers item discriminations for individual items. However, it still does not provide any information with respect to likelihood an item can be answered by guessing. The formula for 2 PL IRT model is given by Equation (3) where a_j is the discrimination parameter of item j. Item discrimination gives information about how well the item can distinguish examinees with low ability than examinees with high ability. When discrimination parameters for all items are set to be 1, it is the same equation for 1 PL IRT model (Equation 2). This model is very commonly used model for simulation studies (Finch, 2015; Svetina & Rutkowski, 2014) since guessing parameter is somehow an unwanted phenomenon with test items.

$$P_{ij}\,(\theta) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \tag{3}$$

3 Parameter Logistic Model

Three parameter logistic (3 PL) model is more advanced model than the models mentioned above since it provides information about guessing parameters of individual items in addition to providing information about difficulty and discrimination parameters of the items. It is the best suitable model among these three models when it is expected that some items to be answered correctly by guessing (de Ayala, 2009). The formula for 3 PL IRT model is given by Equation 4, where c_j is the guessing parameter of item *j*. When guessing parameters for all items are set to be 0, it is the same equation for 2 PL IRT model (Equation 3).

$$P_{ij}\,(\theta) = c_j + (1 - c_j)\ \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \tag{4}$$

There are strong assumptions of IRT such as unidimensionality of the item, local independence, monotonicity of ICC and the parameter invariance. DIF detection in IRT models is directly related to IRT assumptions. ICCs of different groups that are estimated separately are compared with respect to item parameters to investigate the presence of DIF (Thissen et al., 1988). These assumptions are not easily met by any data and large sample sizes are required for data fit.

Differential Item Functioning and Related Terms

Differential Item Functioning (DIF)

DIF is explained as psychometric differences in item performance after groups are matched with respect to the construct being measured by the test (Dorans & Holland, 1992). To detect DIF items, examinees from different groups are matched according to the ability levels. In this way, the effects of ability levels between groups are controlled so that any group differences will not be confounded by ability.

The focus of DIF research is to determine the characteristics of test items that may be different among subgroups of examinees and thus that might explain or be a cause of DIF (Schmitt, Holland, & Dorans, 1992). It is the investigation of items to see if they measure the same latent construct in the same way across identifiable groups. It is important to have DIF free items so that total score could predict the abilities of examinees from subgroups more reliably (Schmitt et al., 1992). It will ensure that the trait being measured is performing same way for multiple groups within the population.

Item Bias

Item bias exists if examinees of the same ability do not have the same probability of answering the item correctly (Holland & Thayer, 1988). Detecting DIF is an early statistical procedure that may indicate that the item may be biased. However, an item displaying DIF does not necessarily mean that the item is biased against one group; it could be because of the multidimensionality of the item. Detecting DIF is the measure of violations from the unidimensionality (Dorans & Holland, 1993), not the multidimensionality of an item itself. Any unintentional dimension is determined as being a cause of DIF only if it favors one of the groups even after the performance of the examinees on the matching variable (commonly total test score) are controlled (Camilli & Shepard, 1994). After an item is labeled as having DIF, before saying that the item is biased, substantive analysis should be conducted by cultural, linguistic experts or by test developers who could make more precise and appropriate decision on item revision or even item removing if needed (De Ayala, 2009; Svetina & Rutkowski, 2014). In other words, any biased items should be investigated to uncover the source of the unintended subgroup differences (Camilli & Shepard, 1994).

Measurement Invariance

Measurement variance indicates the lack of measurement invariance across groups even after accounting for the measurement trait(s) that is the construct being measured. This puts one group in advantage (the cause of DIF) and dangers the fairness and unbiasedness of the assessment since group membership becomes another undesirable variable. To be able to make meaningful comparison, it is required to have measurement equivalence between different groups of test takers (Wiberg, 2007) that is a validity issue with respect to the construct being measured by the assessment.

Item Impact

Item impact is the case, when true difference exists between the groups on the underlying construct being measured (Camilli & Shepard, 1994). In this case, as being different to item bias, it will be natural to have different probabilities of answering an item correctly depending on the true group difference on ability between groups and, any differences between performances of groups on the underlying construct measured will not necessarily imply DIF (the true ability differences between groups on the item). Item impact has been considered and manipulated in many simulation studies related to DIF detection to see the effect of true mean difference on ability between groups (Finch, 2015; Güler & Penfield, 2009; Oshima et al., 2015; Penfield, 2001). However, in this study, real mean of ability estimates of different groups from a real data is used to control the effect of impact.

Types of Differential Item Functioning

There are two type of DIF depending on the interaction between the ability level and the groups, either uniform DIF or nonuniform DIF. From IRT perspective, uniform DIF is present when only the *b* parameter (difficulty) differs across groups that means the difference are always same across the lines; one group is always in advantage. In other words, there is no interaction between ability level and group membership (Mellenbergh, 1982). Uniform DIF can be visualized by Figure 1 where reference group consistently has higher probability of correct answer than the focal group across all levels of ability (θ). The item is easier for reference group.

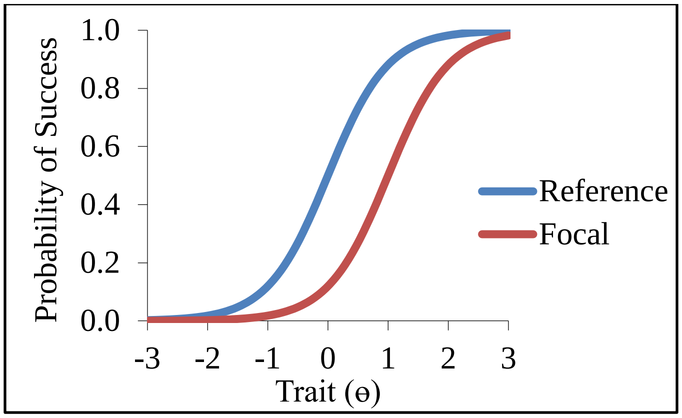

Figure 1: Uniform DIF

Nonuniform DIF exists when *a* parameter (discrimination) differs across groups, regardless of *b* parameter; the change is not same across the groups; there are nonparallel item ICCs of two groups. At certain ability levels, the item may be difficult for one group, whereas for other ability level same item becomes easier for the same group. In contrast to uniform DIF definition, there is an interaction between ability level and groups membership (Mellenbergh, 1982). Nonuniform DIF can be visualized by Figure 2 where focal group has a lower probability of correct answer than the reference group at some levels of the ability matching variable but higher probability of correct answer at other levels of the matching variable.

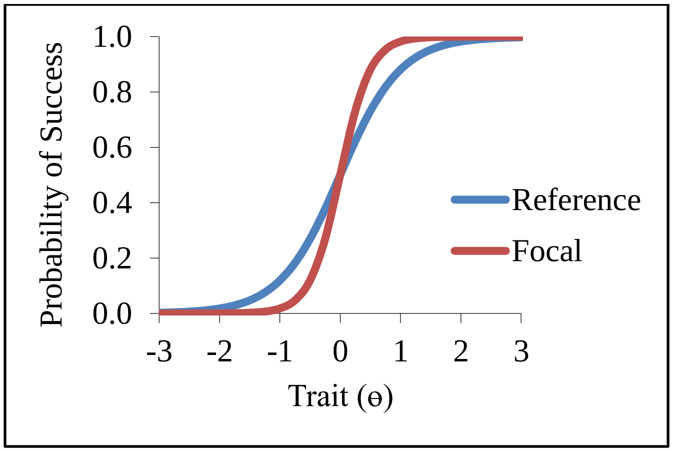

Figure 2: Nonuniform DIF

DIF Detection in General

DIF analysis has commonly three main steps. The first step is to assure the construct equivalence between groups, it is a prerequisite for considering item equivalence (Hui & Triandis, 1985). After it, statistical analyses are conducted at the item level to check the dimensionality of the items. If any dimensionality is found, it is checked to see if it is related to the group membership. Third and last step is the investigation of item biasness done by expertise or item reviewers and decision of either item revision or item removal are decided that finalizes the whole DIF analysis (Thissen, Steinberg, & Gerrard, 1986). Even though all three steps are essential for a whole DIF study, only item level DIF analysis will be considered in this study by conducting a simulation study in the light of a real data analysis.

DIF detection methods are mainly categorized into two groups with respect to the underlying assumptions; either IRT based or non-IRT based (Svetina & Rutkowski, 2014). IRT based techniques use the mathematical models for testing the item equivalence between groups. They require the fitting of item response models and items parameters are estimated from the fitting model. Then comparisons of item parameters are done among groups. Some disadvantages of these methods are that they require larger sample sizes and they are harder to interpret than non-IRT based methods. Most commonly used IRT based methods are Lord's χ^2 test (Lord, 1980), the likelihood-ratio test (Thissen et al., 1988) and Raju's area method (Raju, Van der Linden, & Fleer, 1995).

Non-IRT based methods use the summed score as the matching criteria. These methods are easy with calculation and interpretation and, they do not require very large sample sizes. Most commonly used non-IRT based DIF detection methods are the Mantel-Haenszel (MH) method (Holland & Thayer, 1988), the SIBTEST method (Shealy & Stout, 1993) and the Logistic Regression (LR) method (Swaminathan & Rogers, 1990).

Depending on the data characteristic (dichotomously or polytomously scored items) and the type of DIF of interest (uniform, nonuniform or both), the methods have been preferred in the studies. Clauser and Mazor (1998), Holland and Thayer (1988), Penfield and Camilli (2007) summarized many early two groups DIF methods including the ones mentioned above and some others with their purpose of use. In this study, only dichotomously scored items are of interest with both uniform and nonuniform DIF for two groups and, for multiple groups.

Multiple Groups DIF Detection Methods

Many two groups DIF detection methods recently developed for multiple groups DIF detection. However, as mentioned in the first chapter, there are very limited numbers of studies for multiple groups DIF identification. Some IRT based DIF detection methods among multi-groups that have been studied so far are Lord's Chi-Square (Kim et al., 1995), Langer-Improved Wald test for multiple groups DIF (Woods et al., 2013) Raju's Differential Item Functioning of Items and Tests (DFIT) (Oshima et al., 2015), and the multivariate outlier approach (Magis & De Boeck, 2011). Most recently developed non-IRT based DIF detection methods are the Generalized Mantel-Haenszel (Fidalgo & Scalon, 2009; Penfield, 2001) and the Generalized Logistic Regression (Magis et al., 2011) that are both used in this study and will be explained at the end in details.

Kim et al. (1995) presented a method that was closely related to Lord's chi-square method for comparing vectors of item parameters estimated in two groups and provided a real data example. They investigated the effect of calculator use in mathematics items among three groups. Two groups had calculator during their test; whereas, one group did not have calculator in their test. They found that several items were biased against the group with no calculator. Their emphasis was to show the importance of assessing DIF simultaneously instead of doing pairwise comparison. Same data was reanalyzed by Magis and De Boeck (2011) by using a multivariate outlier approach for multiple groups. Their purpose was to present a robust outlier method for multiple groups DIF detection. They found the same items as DIF items Kim et al. (1995) found in their study. Woods et al. (2013) compared three-group improved Wald testing with pairwise comparison with IRT-LR since the improved Wald test was not evaluated in simulations for more than two groups. Svetina and Rutkowski (2014) conducted a simulation study in the context of international large-scale assessment (ILSA) and used only the GLR. In their simulation study, they manipulated number of groups, magnitude of DIF, percent of DIF, the nature of DIF and the percent of affected groups with DIF. Their finding suggested that the number of groups did not have an effect on the performance of GLR (for 10 groups versus 20 groups). However, the accuracy was affected by other factors. Kanjee (2007) used logistic regression to demonstrate the use of it to multiple groups by conducting a simulation study. This method has been specifically developed for more than two groups to simultaneously access DIF across multiple groups by Magis et al. (2011). Oshima et al. (2015) further developed Differential item Functioning of Items and Tests (DFIT) method (Raju et al., 1995) to multiple groups (NCDIF Method) since DFIT method was not cable of simultaneous comparison of multiple groups. They conducted a simulation study by considering different level of numbers of groups, both types of DIF, different levels of magnitude of DIF, different levels of group sizes and different levels of impact as a demonstration.

Penfield (2001) also studied DIF methods to simultaneously access DIF across multiple demographic groups by comparing three Mantel-Haenszel procedures via a simulation study. He expanded the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959) to more than 2 groups. In his study, he used the MH chi-square statistic with and without adjustment to the alpha level (α) (Bonferroni adjusted alpha level of 0.05) along with the Generalized Mantel-Haenszel (GMH) for a single reference group and 1, 2, 3 and 4 focal groups. A consistent magnitude of DIF, add 0.4 to the *b* parameter for all focal groups experiencing DIF, was used in the study. His results showed that the GMH was the most appropriate procedure. He also used the number of focal groups experiencing DIF as one of his factors in his study. When all focal groups experienced DIF, the GMH was still effective to detect DIF with a poor power compared to the MH, reaching as low as 0.65. The question "How does GMH perform when DIF is nonuniform?" was suggested to study in the future which the current study does. As a further extension to this study and improve some limitations, Fidalgo and Scalon (2009) conducted a simulation study including both dichotomously and polytomously scored items. They specifically used the GMH for both dichotomously scored items and polytomously scored items with 4 ordinal response categories (partial credit model) and compared the performance of different statistics for the GMH under different conditions in a small simulation study.

Magis et al. (2011) used the Generalized Logistic Regression (GLR) as an extension of Swaminathan and Rogers (1990) two-group DIF detection method to detect DIF among multiple groups. In their study, the method is based on maximum likelihood estimation and implemented within the *R* package *difR* (Magis, Béland, Tuerlinckx, & De Boeck, 2010) and they compared the GLR with the GMH and generalized Lord's χ^2 test by using real data collected from a language skill assessment. The main purpose of the study was to show that the LR could be easily extended to multiple-groups DIF testing. First, they concluded that there was not much difference between Wald test and likelihood ratio test for the logistic regression procedure. The interest was both uniform and nonuniform DIF. One biggest advantage of this method is that it does not require choosing one specific reference group. Hence, it is naturally suitable for large-scale assessment studies or international surveys (Magis et al., 2011).

Generalized Mantel-Haenszel Test

The Mantel-Haenszel procedure (Mantel & Haenszel, 1959) was first applied to detect DIF by Holland and Thayer (1988) with only two groups. It is used for both dichotomously and polytomously scored items (Fidalgo & Madeira, 2008) and it is one of the most commonly used DIF detection methods (Mannocci, 2012). When a population is subdivided into *K* groups, it tests the null hypothesis of independence between two dichotomous variables by analyzing 2x2x*K* contingency tables. 2 x 2 tables are the frequency tables of correct and incorrect responses between focal and reference groups at each *k*th score level (Zwick, 2012) for the studied item *j*. A contingency table can be shown by Table 1 where a_{jk} and c_{jk} are the observed number of examinees who answer the item correctly for reference and focal group respectively; whereas, b_{jk} and d_{jk} are the observed number of examinees who answer the items incorrectly for reference and focal group respectively.

Table 1:

The 2 × 2 Contingency Table at the kth Score Level

	Score on the Studied Item *j*		
	Correct (1)	Incorrect (0)	Total
Reference	a_{jk}	b_{jk}	$a_{jk} + b_{jk}$
Focal	c_{jk}	d_{jk}	$c_{jk} + d_{jk}$
Total	$a_{jk} + c_{jk}$	$b_{jk} + d_{jk}$	$a_{jk} + b_{jk} + c_{jk} + d_{jk}$

MH statistic is computed as

$$\mathrm{MH} \ = \frac{\left[\left|\sum a_{jk} - \sum E(a_{jk})\right| - \ 0.5\right]^2}{\sum Var(a_{jk})} \tag{5}$$

where

$$E\left(a_{jk}\right) = \frac{(\mathrm{a}jk + \mathrm{c}jk) \quad (\mathrm{a}jk + \mathrm{b}jk)}{(\mathrm{a}jk + \mathrm{b}jk + \mathrm{c}jk + \mathrm{d}jk)}, \tag{6}$$

$$\mathrm{Var}\left(\mathrm{a_{jk}}\right) = \frac{(\mathrm{ajk + cjk}) \quad (\mathrm{cjk + djk})(\mathrm{bjk + djk}) \quad (\mathrm{ajk + bjk})}{(\mathrm{ajk + bjk + cjk + djk})^2(\mathrm{ajk + bjk + cjk + djk}-1)}, \tag{7}$$

and 0.5 is the Yates' correction for continuity (Yates, 1934).

The MH statistic follows a chi-square distribution with (K – 1) degrees of freedom (df). When the MH statistic is negative, it indicates the studied item's being more difficult for the members in the focal group than for the members in the reference group. When the MH statistic is 0, it indicates no DIF; whereas the positive value of it indicates the disadvantages toward the members in the reference group (The studied item is more difficult for the members in the reference group than for the members in the focal group). The MH can detect uniform DIF but it is not so powerful to detect nonuniform DIF (Swaminathan & Rogers, 1990). The MH was initially used only for 2 group DIF detection. Penfield (2001) extended it to investigate the presence of DIF with more than 2 groups in the form of the Generalized Mantel-Haenszel (GMH) test. It is an extension of the chi-square test of association that allows the comparisons of items responses by conditioning on matching subtest scores across multiple levels. The GMH test statistic is calculated as

$$\text{GMH} = (n_t - \mu_t)'V - 1(n_t - \mu_t), \tag{8}$$

where n_t is the vector of observed number of target responses (e.g. correct) summed across examinees with total score *t* in the reference group, μ_t is the expected number of target responses (e.g. correct) summed across examinees with total score *t* in the reference group, and *V* is the covariance matrix of *n*. GMH test statistic is distributed as a chi-square statistic with G-1 degrees of freedom, where G is the number of groups. If the GMH is significant, the null hypothesis of no DIF present is rejected and the MH is used as a follow up analysis to do pairwise comparison between the reference groups and each of the focal groups. Unlike the GLR method, reference group should be assigned at the beginning of the analysis and the rest of the groups are all set as focal groups for the GMH. Even though it is known that 2 groups MH can only detect uniform DIF, there are not many studies to generalize it for GMH with multiple groups. The current study will consider both uniform and nonuniform DIF for multiple groups for this purpose as well.

Generalized Logistic Regression

The method Logistic Regression was first proposed by Swaminathan and Rogers (1990) to assess uniform DIF between two groups. Later, Narayanon and Swaminathan (1996) demonstrated it for nonuniform DIF for two groups. Swaminathan and Rogers (1990) defined the LR model as

$$P\ (u{=}1){=}\ \frac{e^{z}}{1+e^{z}}, \tag{9}$$

where

$$z = \beta_0 + \beta_1\theta + \beta_2 g + \beta_3(\theta g). \tag{10}$$

In the model θ is the observed trait level (usually the total score) of the examinee and *g* is the group membership. The parameters β_2 and β_3 are the group differences in the performance on the item and the interaction between group and the trait level (Swaminathan & Rogers, 1990). Uniform DIF exists when $\beta_2 \neq 0$ and $\beta_3 = 0$, and nonuniform DIF exists if $\beta_3 \neq 0$ (regardless of β_2). The hypothesis of interest is

$$H_0 : \beta_2 = \beta_3 = 0. \tag{11}$$

And the statistic for testing this null hypothesis follows a chi square distribution with 2 df (Swaminathan & Rogers, 1990).

The LR was extended to defect DIF among multi-groups by Magis et al. (2011) in the form of the GLR. One advantage of this method that distinguishes it from other methods is that it does not require a reference groups to be selected. Another advantage of it is that since the method is naturally modified for the simultaneous comparison, no merging of focal groups is necessary (Svetina & Rutkowski, 2014). The GLR checks the intercepts (α_g) and slopes (β_g) for each group in addition to intercept common across groups (α) and slope common across groups (β) where g indicates the group. The model used is

$$ln\left(\frac{\pi_{ig}}{1\text{-}\,\pi_{ig}}\right)=\begin{cases}\alpha+\beta\,S_i & if\,g=R\\ \left(\alpha+\alpha_g\right)+(\beta+\beta_g)\,S_i & if\,g\neq R\end{cases} \tag{12}$$

where π_{ig} is the probability of correct response of examinee i in group g, S_i is the matching score for examinee i and R is the reference group. Even though the reference group is included in the logistic model, the intercept and slope of it are constrained to 0. If the intercept for at least one focal group is significantly different than 0, then the item is flagged as uniform DIF. In the same way, if the slope for at least one focal group is significantly different than 0, then the item is flagged as nonuniform DIF. The null hypothesis for each case describe above are given as;

$$\begin{aligned} &H_0: \alpha_1 = \cdots = \alpha_G = \beta_1 = \cdots = \beta_G = 0 && \text{(DIF)}\\ &H_0: \beta_1 = \cdots = \beta_G = 0 && \text{(NUDIF)}\\ &H_0: \alpha_1 = \cdots = \alpha_G = 0 \,|\beta_1 = \cdots = \beta_G = 0 && \text{(UDIF)} \end{aligned} \tag{13}$$

As Magis et al. (2011) explained the method estimates the groups parameters (mentioned above) simultaneously that eliminates the pairwise comparison process. Once the item is flagged as either type of DIF, follow up analyses are done by using two statistics; the Wald chi-square and the likelihood ratio test. Magis et al. (2011) recommended using both test simultaneously and when both tests have similar results, Wald chi-square was suggested to interpret the results. However, when they are different, the interpretation should be done with great caution. This method was included in the study since it was appropriate when items are dichotomously scored and the interest was both uniform and nonuniform DIF for more than two group cases (Magis et al., 2011).

Research Questions

The study was guided by two main research questions:

Research Question 1: In investigating uniform DIF, does the magnitude of DIF affect the performance of the Generalized Mantel-Haenszel and the Generalized Logistic Regression under different number of total groups and different groups experiencing DIF? What are the Type I error, power, and precision rates of these two methods for these conditions?

Research Question 2: In investigating nonuniform DIF, does the type of nonuniform DIF affect the performance of the Generalized Mantel-Haenszel and the Generalized Logistic Regression under different number of total groups and different groups experiencing DIF? What are the Type I error, power, and precision rates of these two methods for these conditions?

Chapter 3: METHODOLOGY

The main purpose of this study was to compare the performance of two non-IRT based multi-group DIF detection methods on type I error, precision and power rates. To serve this purpose, a simulation study was conducted in addition to a real data analysis. In this chapter, first, type I error, power and precision are explained in the context of study. Second, real data used in the analysis is described and the mean ability estimates of 12 PISA countries are calculated to be used in the simulation study. Finally, the simulation study, study outcomes and the software used are explained.

Type I Error, Precision and Power

Type I error and power are two commonly used terms in educational measurement. Type I error (false positive) means the rejection of null hypothesis when the null hypothesis in fact is true. In this study, it means labeling non-DIF items as having DIF that indicates a false detection. In DIF study context, it is the identification of an item as displaying DIF without any between-group performance difference in the population (Clauser & Mazor, 1998)

Power (true positive) means the rejection of null hypothesis when the null hypothesis in fact is false. As opposite to type I error, in this study it means labeling DIF items as having DIF that indicates a true detection. Precision is another term that should be considered in DIF studies. In this study context, it means the accuracy of DIF items detected among other items labeled as having DIF because of their also being above the detection threshold. In the study, Bradley's liberal criterion (Bradley, 1978) is considered as the nominal level for type I error rate (a nominal α level of .05). According to Bradley (1978), a test is robust if Type I error rate is approximately equal to the nominal α level.

Real Data Analysis

PISA

The PISA is an international study by the Organization for Economic Co-operation and Development (OECD) (http://www.oecd.org). It was first performed in 2000 and then it has been repeated in every three years. The exam is administered to 15-years old students in OECD member or non-member counties. It measures the performance on mathematics, science and reading. Even though the test is given in every country's native languages, the instrument being measured for the translated items are same. Hence, regardless of the language, examinees from different countries with the same ability level should be still able to answer the item with measuring the same construct with the same probability. In this way, comparison and order done after each PISA administration will be valid and be meaningful. For this reason, it is important to assure that test is fair among all countries. Because of the different cultural and language settings of different countries, these translated tests may not be functioning in the same way in all countries (Ercikan, 1998; 2002). Hence, DIF study is the only way to see it.

Data

The real data analysis of this study examined 34 dichotomously scored mathematics items from PISA 2012, Booklet 4 in 12 countries. Polytomously scored items were not included in the data set since the main interest was only dichotomously scored items. All multiple-choice items were dummy coded accordingly and matched with the booklet for scored items. In the booklet, the coding was 0 for no credit, 1 for full credit, 7 for N/A and 8 for not reached.

These twelve countries including the USA and Turkey were chosen to represent a wide range of ranking and native languages. All twelve countries had close sample sizes (a minimum 351 and maximum 419) to minimize the effect of sample size on the mean of ability estimates of each country (see Table 2).

Table 2:

Sample Sizes of Twelve Countries

Albania	383
Austria	365
Germany	351
France	355
Hong Kong-China	360
Hungary	382
Ireland	390
Israel	385
Korea	386
Sweden	368
Turkey	379
United States of America	419

To estimate the mean of ability of each country, the equating by using concurrent calibration method was conducted. The differences across different countries in terms of ability estimates were important to control the effect of impact on the simulation study. Table 3 gives the mean of ability estimates of the countries.

Table 3:

Mean of Ability Estimates of Twelve Countries

Albania	-0.73241
Austria	0.165933
Germany	0.223196
France	-0.01447
Hong Kong-China	0.617665
Hungary	-0.04876
Ireland	0.133444
Israel	-0.18899
Korea	0.512893
Sweden	-0.14589
Turkey	-0.27212
United States of America	-0.03133

Simulation Study

The manipulated factors in the simulation study were a) the type (uniform versus nonuniform DIF) and the magnitude (none, small, medium and large magnitude of DIF) of DIF items, b) the total number of groups and c) the groups experiencing DIF. The test length, the percent of DIF items, the sample size for each group and the items having DIF were held constant for all conditions. 50 dichotomously scored items were generated by using 2-parameter logistic Item Response Theory (2 PL IRT) model for all conditions. 1000 data sets for each condition were used in the study. This number has been used in many DIF studies (Edwards, 2016; Finch, 2015; Güler & Penfield, 2009; Welkenhuysen-Gybels & Billiet, 2002) and it is also known that smaller sample size yielded poorer type I error (Edwards, 2016). Dichotomous type of items are common in the context of knowledge or aptitude test (Welkenhuysen-Gybels & Billiet, 2002) and the 50 items test length is considered as a long test (Shealy & Stout, 1993) that is very likely to occur for large scale assessments.

The 2 PL IRT model from Equation 3 is

$$P_i\,(\theta) = \frac{e^{(a_i(\theta - b_i))}}{1 + \; e^{(a_i(\theta - b_i))}}$$

Where $P_i\,(\theta)$ is the probability of a correct response to a dichotomous item *i*, a_i is the discrimination parameter and b_i is the difficulty parameter for item *i* and θ is the ability.

Item parameters are taken from certain distributions; *a* ~log-N(0, 1), *b* ~ $\mathcal{N}$ (0, 1). For ability, the means of ability estimates from real data analysis explained above were used in the normal distribution with a standard deviation of 1. Ability parameter for each person was randomly drawn from the normal distributions with these means (See Table 4 for groups and mean of ability estimate used). The means used in the simulation study varied across the groups to represent the likely cases in real data set. For two groups, the minimum and maximum mean values were considered; whereas, for 12 groups all real mean values found were considered. For 6 groups, random selection of mean of ability estimates was considered. Item parameter distributions, ability (latent trait) distribution were also the independent variables in the study.

Table 4:

Mean of Ability Estimates for Groups

2 Groups	-0.73241, 0.617665
6 Groups	-0.73241,0.617665, -0.27212, 0.223196, 0.133444, -0.03133
12 Groups	-0.73241,0.165933, 0.223196, -0.01447, 0.617665, -0.04876, 0.133444, -0.18899, 0.512893, -0.14589, -0.27212, -0.03133

For the binary responses (0 for false and 1 for true) for dichotomous scoring, a random number is drawn from a uniform (0, 1) distribution for each item. If the draw value is less than or equal to $P_i\,(\theta)$, the item response is 1, if the draw number is greater than $P_i\,(\theta)$, the item response is 0. DIF contamination (percent of DIF items) was held constant in the simulation design. 10% of total items were generated as having DIF (the first five items) for all conditions similar to the study of Welkenhuysen-Gybels (2004). Also Shirley (2014) conducted a differential item functioning analysis to examine question/ item bias of PISA 2009 science items between two groups; non-language learners (non-LLs) versus language learners (LLs) across 45 countries by using IRT based DIF techniques on the 35 science items from PISA 2009, Booklet 3 and found that at least 5 items may have DIF. Hence, 5 as the number of DIF items were used in this study to represent the total number of DIF items.

The type and the magnitude of DIF items: Both uniform and nonuniform DIF items were investigated in the study. The magnitude of uniform DIF was examined at three different levels. For uniform DIF items, b parameters were manipulated by adding 0.4, 0.6 and 0.8 to the original b parameters to represent the small, medium and large level of uniform DIF respectively (Zumbo, 1999). These values have been used in prior research with Finch (2015), Penfield

(2001) and Rogers & Swaminathan, (1993). For nonuniform DIF items, two cases were investigated: only *a* parameter difference and, both *a* and *b* parameters difference. For only *a* parameter difference, *a* parameter was manipulated by adding 0.4 to the original *a* parameter, and for both *a* and *b* parameter difference, *a* parameter was manipulated by adding 0.4 to the original *a* parameter and *b* parameter was manipulated by adding 0.7 to the original *b* parameter (Oshima et al., 2015). Also, the non-DIF conditions were considered for all the groups to be able to calculate the type I error rate when no DIF items were present. Without changing any item parameters for any groups, the analysis was run to see how many times non-DIF items were falsely labeled as DIF items. Power and precision rates were not applicable for these cases.

Number of groups: As mentioned, early DIF studies mainly focused on only two groups. However, the purpose of this study was to assess DIF with multiple groups. Although two groups case was included as a baseline for earlier two group studies (Güler & Penfield, 2009; Narayanon & Swaminathan, 1996; Swaminathan & Rogers, 1990), more than two groups was interest. Hence, three conditions were simulated for the number of groups: 2, 6 and 12. 6 groups were also purposely included as a baseline to recently done Finch's study (2015). 12 group cases were included to represent large number of groups seen in large-scale assessments with ethnicity or in cross-cultural studies with different countries.

Groups experiencing DIF: Since early DIF studies considered 2 groups, only case was one group's (focal group) experiencing DIF. However, as the number of focal groups increased other possibilities arose. As a result, 5 different cases of groups experiencing DIF were investigated in the study; only one focal group (first focal group) experiencing DIF, all focal groups experiencing DIF, half of the total groups including reference group experiencing DIF, half of the focal groups experiencing DIF and lastly, just reference group experiencing DIF.

Group size ratio was held constant in the study (1:1). For all conditions, focal groups have the same sample sizes with reference group (1000 examinees). Hence 2,000-, 6,000- and 12,000 total sample sizes were considered for 2-, 6- and 12 groups respectively.

There were 55 cases (5x5x2 for 6 and 12 groups and 5x1x1 for 2 groups) for DIF conditions plus 3 cases for non-DIF conditions in the simulation study (Total 58 cases). Table 5 summarizes the all the conditions used in the study.

Table 5:

Simulation Conditions Considered for the Study

Variables	**Levels**
Number of Groups	Two, Six and Twelve
Reference Sample Size	1000
Group Size	Equal
Type of DIF	Uniform and Nonuniform
Level of DIF	0, 0.4, 0.6, 0.8 added to *b* parameters for Uniform DIF (0, 0.4), (0.7, 0.4) added to *a* and *b* parameters respectively
Impact	True Mean of Ability Estimates to be used in Normal Distributions (See Table 3)
Groups Experiencing DIF	One Focal Group, All Focal Groups, First Half of the Groups including Reference Group, Second Half of the Focal Groups Only Reference Group
Number of Items	50
Target Items	First Five Items
DIF Methods	Generalized Mantel-Haenszel test, Generalized Logistic Regression
Number of Iterations	1000

Study Outcomes

There were 3 study outcomes of interest; type I error rate, power and precision rate. There were also the dependent variables of the study. Type I error rates for each condition were calculated as the proportion of items within a replication that were originally non-DIF but flagged as having DIF in the analysis. In the same way, power rates for each condition were calculated as the proportion of items that were originally DIF items and flagged as having DIF in the analysis. Rates were averaged across the number of replication. 1000 replications were used for each case to provide the accuracy in determining the outcomes and ensure the stability. Precision rate was also calculated as the proportion of DIF items within a replication that were above a certain threshold to all those items above that threshold. Even though the precision is not widely used DIF studies, I believe that it is important to investigate how accurately the methods detect DIF items among other items that are above a certain threshold to be a potential false positive. For example when there are five items above a certain threshold and 4 of them are real DIF items then the accuracy is 4 divided by 5 that is 80%. An alpha level .05 was used to decide the significance of the results. No repeated measure of ANOVA was conducted in the study since the only level and type of DIF were between-subject effects; whereas, type I error, power or precision were dependent variables. Instead each parallel condition with 6 and 12 groups was interpreted and compared individually by considering the level of uniform and nonuniform DIF separately. Each line graph compared type I error, power and precision rates across either magnitude uniform DIF or different types of nonuniform DIF for each case. For example Figure 1 gives type I error rate comparison for 6 Groups uniform DIF case when one focal group experienced DIF and this graph was compared with Figure 39 (type I error rate comparison for 12 Groups uniform DIF case when one focal group experienced DIF).

Software and Methods Used

IBM SPSS, version 21 was used to organize the real data, choose booklet and countries. Dummy coding was also done. Data generation for the simulation study was done through R statistical software (R Development Core Team, 2014). DifR package under genDichoDif (methods are “genLogistic” for GLR and “DifGMH” and for the GMH) function (Magis et al., 2010) was used to compare the performance of the Generalized Mantel-Haenszel DIF method and the Generalized logistic regression DIF method for both simulation study and real data analysis. Also each method was run separately to get the significance level and p values for each item for the real data analysis. The reason of using these two methods is that they are both non-IRT based DIF detection methods with easy application and interpretation. Both of them were initially developed for two groups cases. The Generalized Mantel-Haenszel DIF method is the developed version of the Mantel-Haenszel method for two groups DIF and the Generalized Logistic Regression is the developed version of the Logistic Regression for two groups DIF. They both use total test scores in matching examinees from different groups and they are Chi square methods (Holland & Thayer, 1988; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Among IRT based and non-IRT based methods, the MH is the one best performing with the smaller error rates and high power (Gómez-Benito & Navas-Ara, 2000). Also Gómez-Benito and Navas-Ara (2000) showed that non-IRT based methods performed better than IRT based methods. However, one disadvantage of using the MH they stated is that it can only detect uniform DIF items. There are many other studies that compare the performance of the MH and the LR under several conditions (Clauser & Mazor, 1998; Gierl, Jodoin, & Ackerman, 2000; Narayanon & Swaminathan, 1996; Rogers & Swaminathan, 1993). On the other hand, when they were developed to simultaneously detect DIF across groups, they are not many studies that compare the performance of the GMH and the GLR in both a real study and simulation study. Hence, current study aimed to do it by both using a real data analysis and a simulation study.

Chapter 4: RESULTS

This chapter summarizes the results of the real data analysis and the results of simulation study for dichotomously scored data to compare the performance of two non-IRT based multi-group DIF detection methods on the type I error, power and precision rates under three main conditions: Different types and levels of DIF, different number of total groups and different number of groups experiencing DIF.

Results of Real Data Analysis

For analyzing real data, in R statistical software (R Development Core Team, 2014) the difR package (Magis et al., 2010) was used for the comparison of the GMH and the GLR (function genDichoDif). From PISA 2012, 34 dichotomously scored Mathematics items were used from Booklet 4. Surprisingly, all 34 items were found to have DIF across 12 countries from both methods. For the GLR, both Wald statistic and the likelihood ratio statistics had the same results. Table 6 gives results for the GLR by using the likelihood ratio statistics and for the GMH with the DIF detection threshold and with significance level of 0.05. Detection threshold for the GMH was 19.6751 and for the GLR (likelihood ratio statistics) it was 33.9244.

Table 6:

Results of Real Data Analysis

	GMH		GLR	
PM00FQ01	169.6592	0.0000 ***	248.4271	0.0000 ***
PM00KQ02	138.4227	0.0000 ***	155.6433	0.0000 ***
PM903Q03	173.1774	0.0000 ***	193.7371	0.0000 ***

PM905Q01T	79.5815	0.0000 ***	97.9162	0.0000 ***
PM905Q02	45.5959	0.0000 ***	62.5699	0.0000 ***
PM906Q01	56.8866	0.0000 ***	70.3861	0.0000 ***
PM915Q01	155.6671	0.0000 ***	192.9139	0.0000 ***
PM915Q02	197.8418	0.0000 ***	223.5867	0.0000 ***
PM918Q01	92.2958	0.0000 ***	127.3143	0.0000 ***
PM918Q02	102.3192	0.0000 ***	133.6291	0.0000 ***
PM918Q05	264.5722	0.0000 ***	309.0842	0.0000 ***
PM919Q01	40.7210	0.0000 ***	76.5810	0.0000 ***
PM919Q02	22.9561	0.0179 *	38.6217	0.0156 *
PM923Q01	68.8864	0.0000 ***	79.2076	0.0000 ***
PM923Q03	113.5119	0.0000 ***	165.9272	0.0000 ***
PM923Q04	30.7982	0.0012 **	42.8061	0.0050 **
PM924Q02	61.4149	0.0000 ***	81.5241	0.0000 ***
PM943Q01	54.8512	0.0000 ***	113.0764	0.0000 ***
PM943Q02	153.2168	0.0000 ***	161.1312	0.0000 ***
PM953Q02	46.0019	0.0000 ***	62.7969	0.0000 ***
PM953Q03	50.1386	0.0000 ***	68.7455	0.0000 ***
PM954Q01	74.7871	0.0000 ***	85.2337	0.0000 ***
PM954Q02	190.2703	0.0000 ***	203.9296	0.0000 ***
PM954Q04	99.9166	0.0000 ***	120.3490	0.0000 ***
PM982Q01	43.8132	0.0000 ***	67.0460	0.0000 ***
PM982Q02	113.7763	0.0000 ***	143.2207	0.0000 ***

PM982Q03T	52.2716	0.0000 ***	74.6137	0.0000 ***
PM982Q04	53.8242	0.0000 ***	76.6954	0.0000 ***
PM992Q01	55.3239	0.0000 ***	78.2386	0.0000 ***
PM992Q02	239.6271	0.0000 ***	271.0791	0.0000 ***
PM992Q03	127.6755	0.0000 ***	180.5091	0.0000 ***
PM995Q01	146.0386	0.0000 ***	158.4782	0.0000 ***
PM995Q02	45.0589	0.0000 ***	81.7388	0.0000 ***
PM995Q03	92.9103	0.0000 ***	136.8326	0.0000 ***
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Detection threshold: 19.6751 (significance level: 0.05)		Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Detection threshold: 33.9244 (significance level: 0.05)	

However, identifying the possible source of the DIF was beyond the scope of this study. Hence, no further investigations with the items were done nor any pairwise analyses were conducted among countries. However, it gave the inspiration to simulation study that type I error could be highly inflated when large number of total groups were considered.

Result of the Simulation Study

In the study, the type I error, power and precision were averaged across 1000 iterations for all conditions. For the type I error rate, the ratio of misidentified non-DIF items over all non-DIF items was calculated for each case within a replication and then the sum of these ratios for one condition across 1000 iteration was divided by 1000. For the power rate, the ratio of true identified DIF items over all DIF items for each condition within a replication was calculated and then the sum of these ratios was divided by 1000. For the precision, the ratio of true identified DIF items over all items identified as DIF was calculated for each condition within a replication and the sum of these ratios were divided by 1000. For some iteration, the values of precision were not available since there were no DIF items found (zero divided by zero). In these cases, the precision rate was calculated across the applicable iterations (less than 1000 iterations). Table 7 shows how the items were identified for 6 group small uniform DIF condition for one iteration and the calculations of the type I error, power and precision are given with one iteration for both methods. An alpha level of .05 is used for all identifications.

Table 7:

Results from One Iteration from One Focal Group out of Six Groups Experiencing Small Uniform DIF

	M.-H.	Logistic	# DIF
Item1	DIF	DIF	2/2
Item2	DIF	DIF	2/2
Item3	DIF	DIF	2/2
Item4	DIF	DIF	2/2
Item5	NoDIF	NoDIF	0/2
Item6	DIF	DIF	2/2
Item7	DIF	DIF	2/2
Item8	NoDIF	NoDIF	0/2
Item9	NoDIF	NoDIF	0/2
Item10	NoDIF	NoDIF	0/2
Item11	NoDIF	NoDIF	0/2

Item11	NoDIF	NoDIF	0/2
Item12	DIF	DIF	2/2
Item13	NoDIF	NoDIF	0/2
Item14	NoDIF	NoDIF	0/2
Item15	NoDIF	NoDIF	0/2
Item16	NoDIF	NoDIF	0/2
Item17	DIF	NoDIF	1/2
Item18	DIF	DIF	2/2
Item19	NoDIF	NoDIF	0/2
Item20	DIF	NoDIF	1/2
Item21	DIF	DIF	2/2
Item22	NoDIF	NoDIF	0/2
Item23	DIF	DIF	2/2
Item24	NoDIF	NoDIF	0/2
Item25	NoDIF	NoDIF	0/2
Item26	NoDIF	NoDIF	0/2
Item27	NoDIF	DIF	1/2
Item28	DIF	DIF	2/2
Item29	NoDIF	NoDIF	0/2
Item30	NoDIF	NoDIF	0/2
Item31	NoDIF	NoDIF	0/2
Item32	NoDIF	NoDIF	0/2
Item33	DIF	DIF	2/2

Item34	NoDIF	NoDIF	0/2
Item35	NoDIF	NoDIF	0/2
Item36	NoDIF	NoDIF	0/2
Item37	NoDIF	NoDIF	0/2
Item38	NoDIF	DIF	1/2
Item39	NoDIF	NoDIF	0/2
Item40	NoDIF	NoDIF	0/2
Item41	NoDIF	NoDIF	0/2
Item43	NoDIF	NoDIF	0/2
Item44	NoDIF	NoDIF	0/2
Item45	NoDIF	NoDIF	0/2
Item46	NoDIF	NoDIF	0/2
Item47	NoDIF	NoDIF	0/2
Item48	NoDIF	NoDIF	0/2
Item49	DIF	DIF	2/2
Item50	NoDIF	NoDIF	0/2

From table 7, for both GMH and GLR, power was 4/5 = 0.8 (Four of first 5 DIF items were detected truly as having DIF). Type I error for GMH and GLR was 11/45 (11 of last 45 non-DIF items were detected falsely as having DIF). Precision for both GMH and GLR = 4/15 (There were 15 items above the threshold and 4 of them were true DIF items).

The simulation study sought answers for two research questions below. The answer of each research question was given separately for total number of groups with uniform and nonuniform DIF cases. Figures for results were also provided for each related area.

Research Question 1: In investigating uniform DIF, does the magnitude of DIF affect the performance of the Generalized Mantel-Haenszel and the Generalized Logistic Regression under different number of total groups and different groups experiencing DIF? What are the Type I error, power, and precision rates of these two methods for these conditions?

Research Question 2: In investigating nonuniform DIF, does the type of nonuniform DIF affect the performance of the Generalized Mantel-Haenszel and the Generalized Logistic Regression under different number of total groups and different groups experiencing DIF? What are the Type I error, power, and precision rates of these two methods for these conditions?

Two Groups

Uniform DIF

When there were only two groups in total (one reference group and one focal group), both methods had very similar results on the type I error rates for uniform DIF. (See Figure 3). As the magnitude of uniform DIF increased, type I error rate for both methods increased as well. When no DIF existed, type I error rate was still high (around 0.15) that indicated at least 6 items out of 45 non-DIF items were falsely identified as having DIF. This error rate was around 0.3 when large magnitude of uniform DIF (0.8) existed that meant at least 13 items out of 45 non-DIF items were falsely identified as having DIF in average.

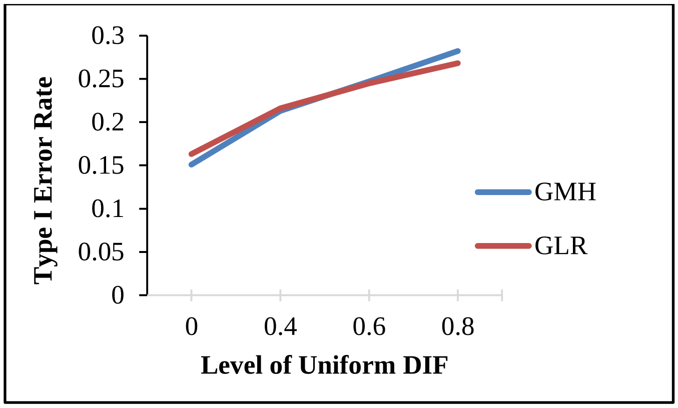

Figure 3: Type I Error Rates for 2 Groups Uniform DIF Case

Power rate for two groups was also very similar for both methods (See Figure 4). As the magnitude of DIF increased, the performance of both GMH and GLR was really high (reaching 1 for large magnitude of DIF). When small magnitude of DIF existed, power rate for both methods was around 0.7 that meant at least 3 items out of 5 DIF items were truly identified as having DIF.

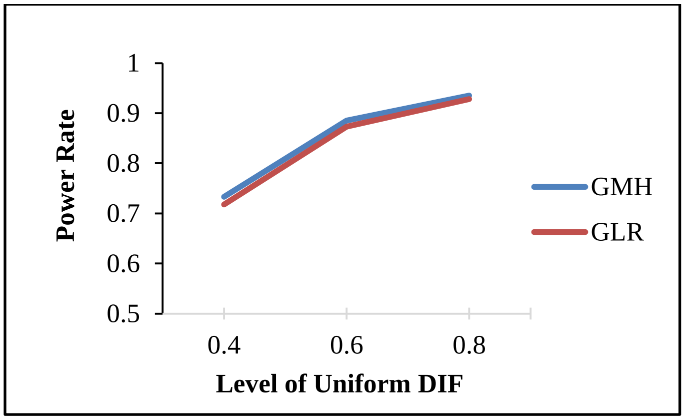

Figure 4: Power Rates for 2 Groups Uniform DIF Case

Precision rate for two groups had almost identical results for both methods (See Figure 5). As expected from type I error rates and power rates for same condition - large magnitude of

DIF (0.8), the precision rates for both methods were really high. There was not much difference between small magnitude of uniform DIF and medium magnitude of uniform DIF on their precision rates (around 0.3 for both methods).

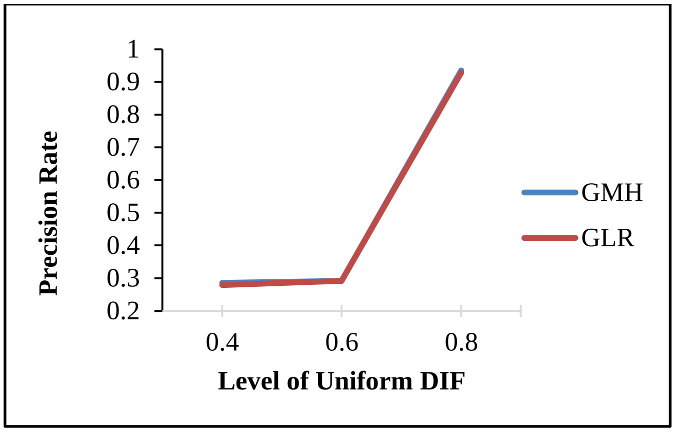

Figure 5: Precision Rates for 2 Groups Uniform DIF Case

Nonuniform DIF

When it was nonuniform DIF with only *a* parameter change, type I error rate for the GLR was higher than the GMH and, when it was nonuniform DIF with both *a* parameter and *b* parameter change, type I error rate for the GMH was higher than the GLR (See Figure 6). There was not much difference between two types of nonuniform DIF for the GLR; whereas, nonuniform DIF with both *a* parameter and *b* parameter change was higher than nonuniform DIF with only *a* parameter for GMH. Type I error rates of both nonuniform DIF were highly inflated (with a minimum 0.25 and maximum 0.35).

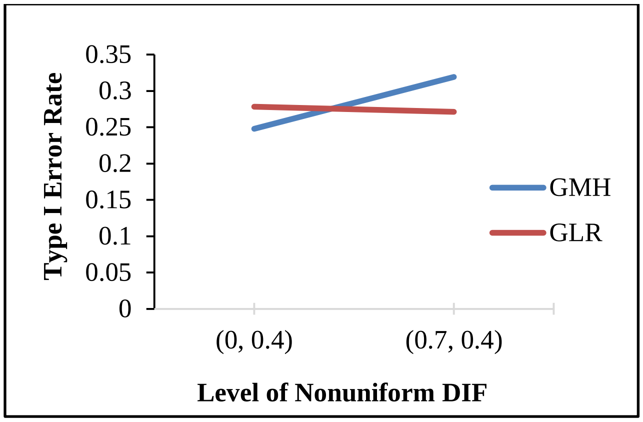

Figure 6: Type I Error Rates for 2 Groups Nonuniform DIF Case

Power rate of the GLR for nonuniform DIF with only *a* parameter change was higher than power rate of the GMH for nonuniform DIF with only *a* parameter change (See Figure 7). There was not much difference on the GLR’s power rate for two type of nonuniform DIF; whereas, power rate of the GMH for both *a* parameter and *b* parameter change was higher than the power rate of GMH for nonuniform DIF with only *a* parameter change . GLR was able to detect almost all the DIF items for both types of nonuniform DIF; whereas, GMH was able to detect 2 out of 5 DIF items in average for nonuniform DIF with only *a* parameter change and detect 3 out of 5 DIF items in average for nonuniform DIF with both *a* parameter and *b* parameter change.

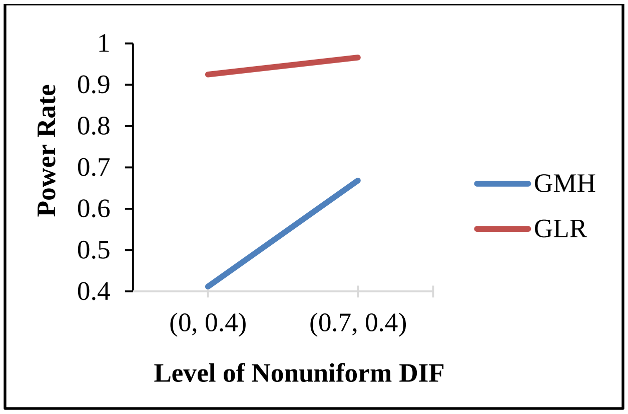

Figure 7: Power Rates for 2 Groups Nonuniform DIF Case

Precision rate of the GLR for nonuniform DIF with only *a* parameter change was higher than power rate of GMH for nonuniform DIF with only *a* parameter change (See Figure 8). There was not much difference on GLR's precision rate for two type of nonuniform DIF; whereas, precision rate of GMH for both *a* parameter and *b* parameter change was higher than the precision rate of GMH for nonuniform DIF with only *a* parameter change.

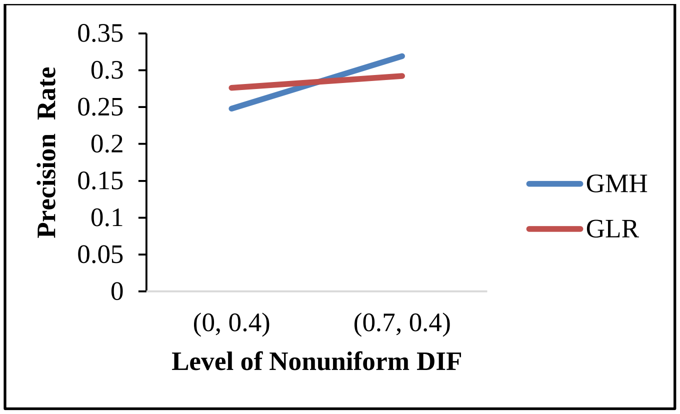

Figure 8: Precision Rates for 2 Groups Nonuniform DIF Case

Summary

Overall, there was not much difference between the GMH and the GLR for uniform DIF case on all magnitudes. Both methods' power rates were high with a minimum value of around 0.7. As the magnitude of uniform DIF increased, the power rate of both methods increased as well. For nonuniform DIF with only *a* parameter change, the GLR was more powerful than the GMH, and the GMH was still comparable to the GLR for nonuniform DIF with both *a* and *b* parameters change. Type I error rates for all cases were above the nominal level of .05 and highly inflated for many cases with both methods. When there was no DIF item, GMH had slightly better control than GLR (0.151 versus 0.163). Precision rates of both methods were very close to each other for uniform DIF. GLR had consistently higher precision rate for two types of nonuniform DIF. Table 8 gives type I, power and precision rates of all cases for 2 groups.

Table 8:

Results for 2 Groups

	GMH			GLR		
	Type I	Power	Precision	Type I	Power	Precision
0	0.151	-	-	0.163	-	-
0.4	0.213	0.733	0.286	0.216	0.718	0.279
0.6	0.247	0.885	0.292	0.245	0.873	0.292
0.8	0.282	0.935	0.935	0.268	0.928	0.928
(0, 0.4)	0.248	0.412	0.248	0.319	0.668	0.319
(0.7, 0.4)	0.278	0.925	0.276	0.271	0.966	0.292

Six Groups

Uniform DIF

In contrast to 2 groups, for 6 groups, when there was no DIF item, the GMH had a higher type I error rate than the GLR (See Figure 9). When only one focal group experienced DIF, type I error rate of the GMH for all magnitude of uniform DIF was higher than the GLR. When the magnitude of uniform DIF was medium (0.6), both methods had the lowest type I error rates. Even though type I error rates for these cases were above the nominal level of .05, the highest value was around 0.1.

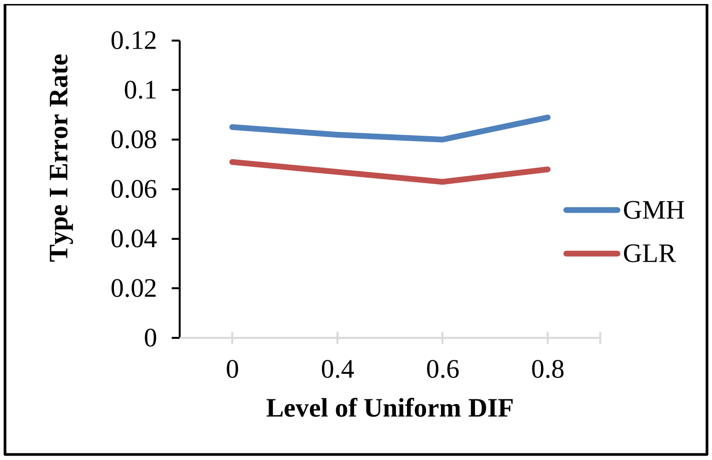

Figure 9: Type I Error Rate for 6 Groups Uniform DIF - One Focal Group Experiencing DIF

Power rate of the GMH was consistently higher than the power rate of GLR for 6 groups uniform DIF case when only one focal group experienced DIF (See Figure 10). Even though there was not much difference among the magnitude of uniform DIF with respect to power rate of the GMH, the GLR had higher power rate for small magnitude of uniform DIF and lowest power rate for the large magnitude of uniform DIF.

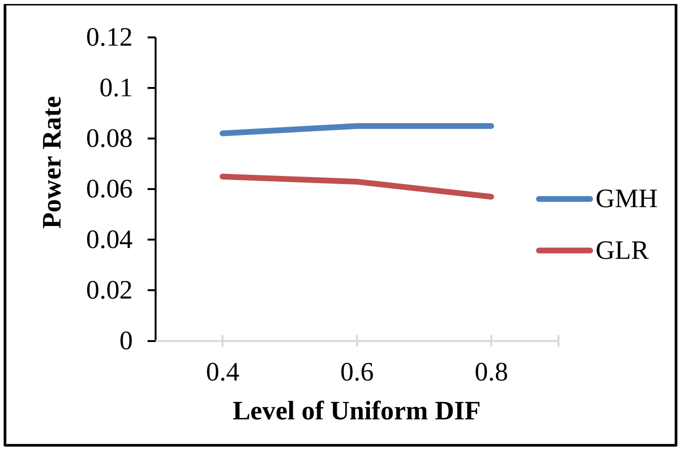

Figure 10: Power Rate for 6 Groups Uniform DIF - One Focal Group Experiencing DIF

Precision rate of the GLR for 6 groups uniform DIF case with small magnitude was higher than the GMH when only one focal group experienced DIF; whereas, it was lower with large magnitude of uniform DIF (See Figure 11). When it was medium magnitude of uniform DIF, both methods had the highest and very similar precision rates.

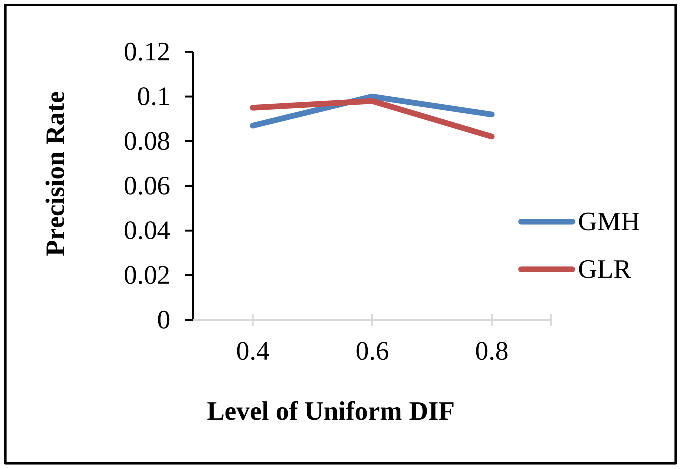

Figure 11: Precision Rate for 6 Groups Uniform DIF - One Focal Group Experiencing DIF

Type I error rate of the GMH was consistently higher than type I error rate of the GLR for 6 Groups Uniform DIF cases when half of the groups including reference group experienced DIF (See Figure 12). When the magnitude of uniform DIF was medium, both methods had highest type I error rates. However, there was not much difference among the magnitude of DIF with respect to each method’s type I error rate. It was above the nominal level of .05 with the highest value of 0.8.

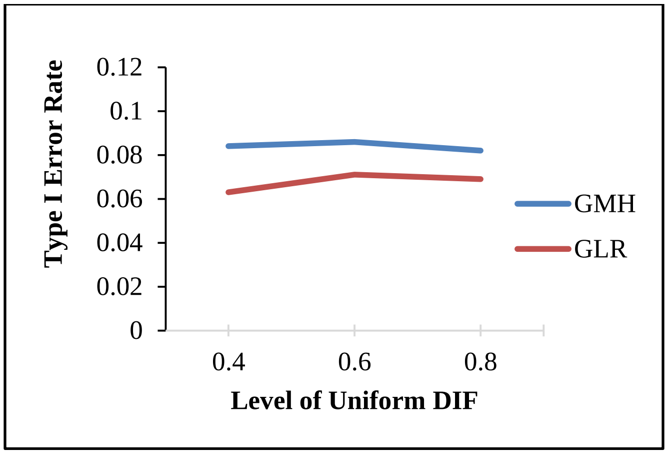

Figure 12: Type I Error Rate for 6 Groups Uniform DIF –Half of the Groups including Reference Group Experiencing DIF

Power rate of the GMH was also consistently higher than the power rate of the GLR for 6 Groups Uniform DIF cases when half of the groups including reference group experienced DIF (See Figure 13). When the magnitude of uniform DIF was medium, both methods had the highest power rates and they decreased when it was large magnitude of uniform DIF. However, none of the methods were really powerful; the highest value was around 0.1.

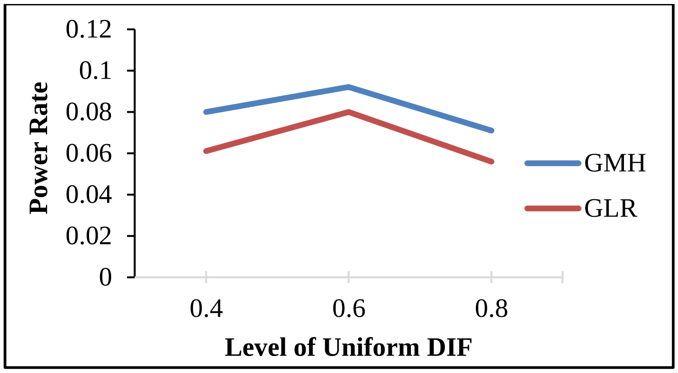

Figure 13: Power Rates for 6 Groups Uniform DIF – Half of the Groups including Reference Group Experiencing DIF

Precision rate of the GLR was slightly higher than the precision rate of the GMH for 6 groups small magnitude of uniform DIF when half of the groups including reference group experienced DIF (See Figure 14). There was not much difference between small and medium magnitude of uniform DIF for both methods. Precision rates of both methods got decreased for large magnitude of uniform DIF.

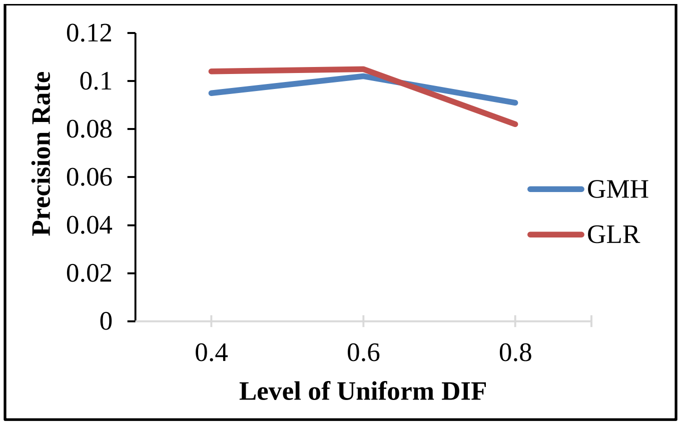

Figure 14: Precision Rates for 6 Groups Uniform DIF - Half of the Groups including Reference Group Experiencing DIF

When half of the focal groups experienced DIF for 6 groups, type I error rate of the GMH was consistently higher than the type I error rate of GLR (See Figure 15). Type I error rate of the GLR was around 0.06 that was slightly higher than the nominal level of 0.05; whereas, it was around 0.08 for GMH.

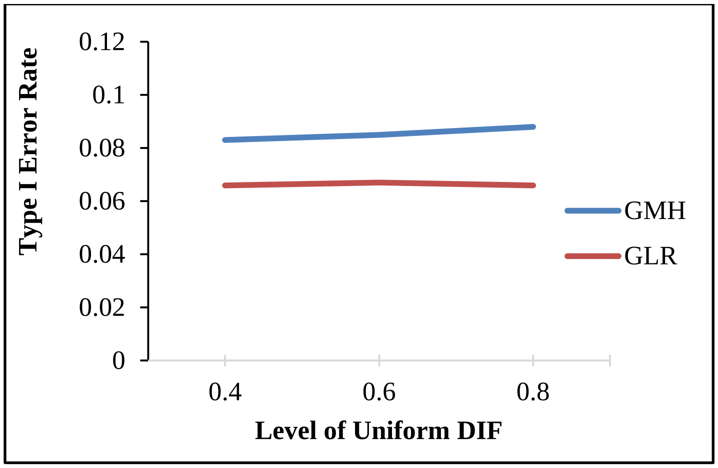

Figure 15: Type I Error Rates for 6 Groups Uniform DIF – Half of the Focal Experiencing DIF

Power rate of the GMH for 6 groups when half of the focal groups experienced DIF was higher than the power rate of the GLR for all magnitude of uniform DIF (See Figure 15). For GMH, power rate got decreased as the magnitude of uniform DIF increased. For GLR, the power rate for small and large magnitude of uniform DIF was similar; whereas, the power rate for medium magnitude of uniform DIF was the smallest. Power rate was below 0.15 for both methods.

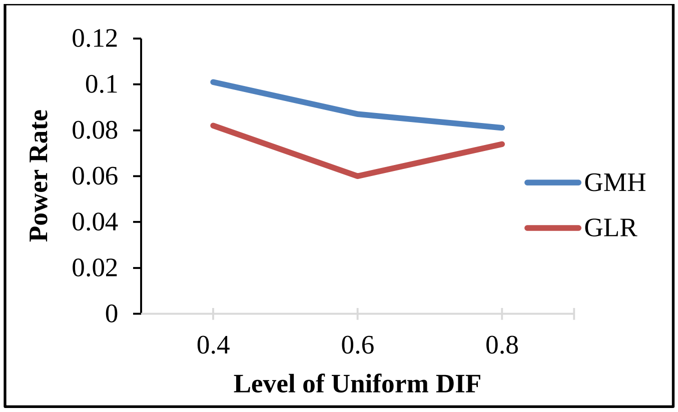

Figure 16: Power Rates for 6 Groups Uniform DIF- Half of the Focal Groups Experiencing DIF

Precision rate of the GMH for 6 groups when half of the focal groups experienced DIF was higher than the precision rate of the GLR in general (See Figure 17). The precision rate of the GLR was smallest when the magnitude of uniform DIF was medium and it was highest when the magnitude of uniform DIF was small. For the GMH, the highest value was around 0.12 and the lowest value was around 0.09; whereas, for GLR, the highest value was around 0.125 and lowest value was around 0.08.

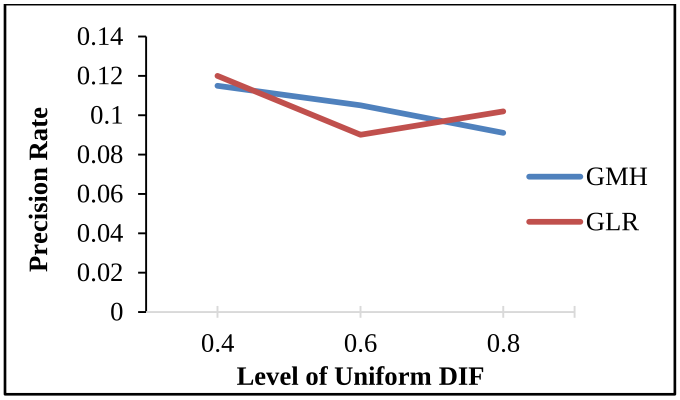

Figure 17: Precision Rates for 6 Groups Uniform DIF – Half of the Focal Groups Experiencing DIF

Type I error rate of the GMH for 6 groups uniform DIF when all focal groups experiencing DIF was higher than the type I error rate of the GLR in general (See Figure 18). When the magnitude of uniform DIF was medium, the type I error rate of the GMH was the smallest. Type I error rate of the GLR slightly increased as the magnitude of uniform DIF increased. It was slightly above to the nominal level of .05 for the GLR.

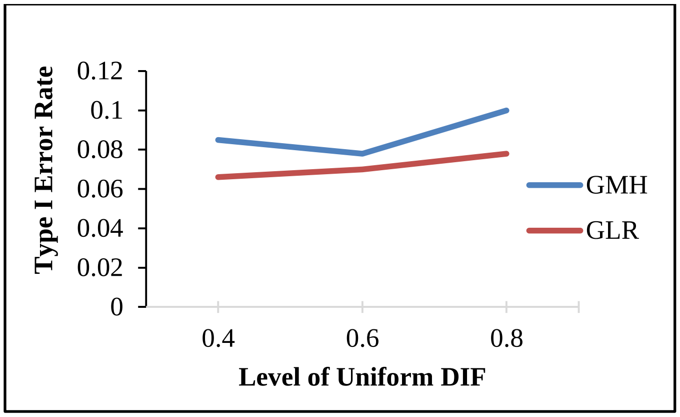

Figure 18: Type I Error Rates for 6 Groups Uniform DIF – All Focal Groups Experiencing DIF

Power rate of the GMH was consistently higher than the power rate of the GLR for 6 groups uniform DIF when all focal groups experienced DIF (See Figure 19). Both methods had highest power rate when it was small magnitude of uniform DIF. There was not much difference between medium and large magnitude of uniform DIF with respect to power rate of each method.

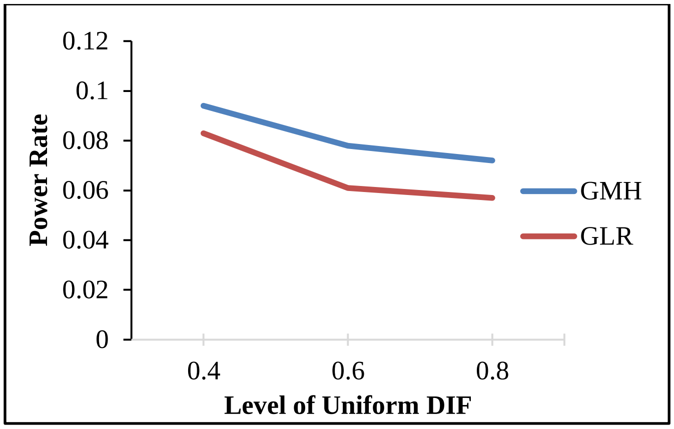

Figure 19: Power Rate for 6 Groups Uniform DIF – All Focal Groups Experiencing DIF

Precision rate of the GLR was higher than the precision rate of the GLR for 6 group small magnitude of uniform DIF when all focal groups experiencing DIF (See Figure 20). Precision rate of the GMH was higher than the GLR when it was medium magnitude of uniform DIF. Both methods had same precision rates when it was large magnitude of uniform DIF.

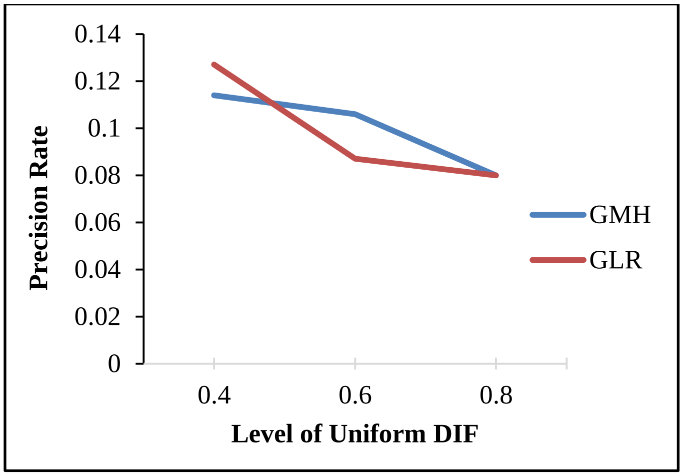

Figure 20: Precision Rates for 6 Groups Uniform DIF –All Focal Groups Experiencing DIF

When it was reference group experiencing DIF with 6 groups, type I error rate of the GMH was consistently higher than the type I error rate of the GLR for all magnitude of uniform DIF (See Figure 21). There was not any difference between small and medium magnitude of uniform DIF for both methods' type I error rates. When it was large magnitude of uniform DIF, both methods had the highest type I error rates that was around 0.08 for the GMH and 0.07 for the GLR.

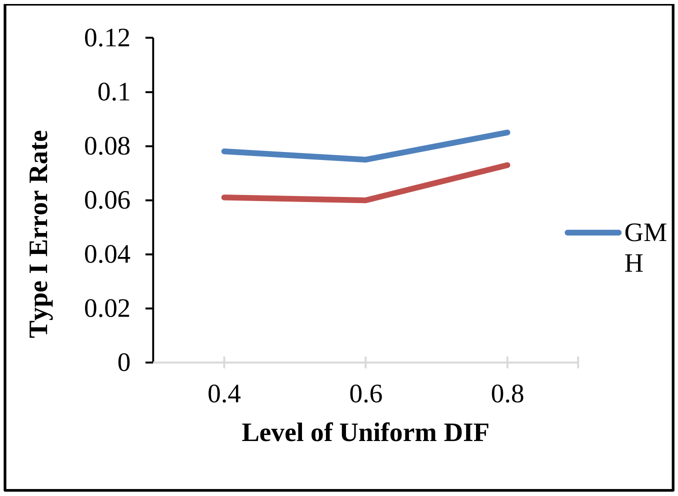

Figure 21: Type I Error Rates for 6 Groups Uniform DIF- Reference Group Experiencing DIF

Power rate of both methods were really high and very similar when it was reference group experienced DIF (See Figure 22). The power rate ranged between 0.6 and 0.8 for both methods. There was not much difference between medium and large magnitude of uniform DIF for both methods.

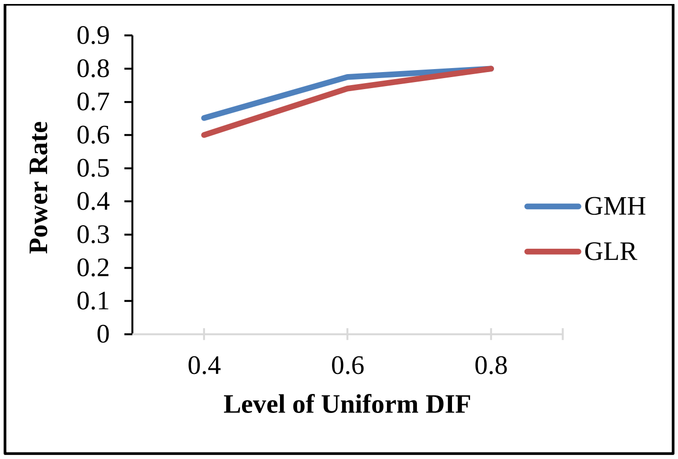

Figure 22: Power Rates for 6 Groups Uniform DIF – Reference Group Experiencing DIF

Precision rate of the GLR was slightly higher than the precision rate of the GMH for 6 groups all levels of uniform DIF when reference group experienced DIF (See Figure 23). When it was medium magnitude of uniform DIF, both methods' precision rates were highest. Precision rate of the GMH was around 0.65 and it was around 0.55 for the GLR.

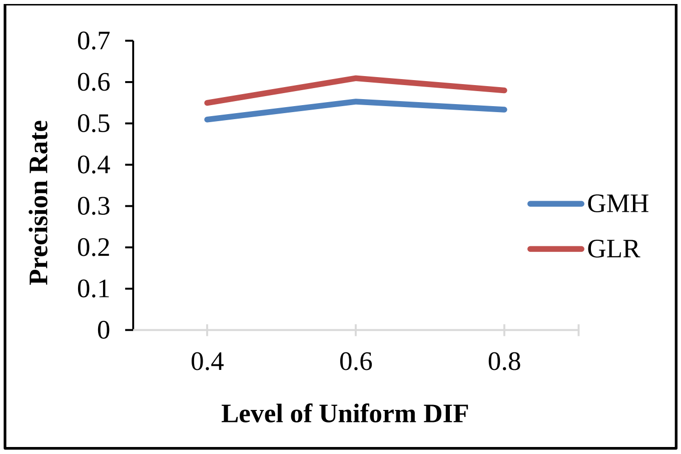

Figure 23: Precision Rates for 6 Groups Uniform DIF – Reference Group Experiencing DIF

Nonuniform DIF

Type I error rate of the GMH was consistently higher than the type I error rate of the GLR for both types of nonuniform DIF for 6 groups when one focal group experienced DIF (See Figure 24). There was no difference between nonuniform DIF with only *a* parameter change and, nonuniform DIF with both *a* and *b* parameter change. The rate was around 0.7 for the GLR and 0.9 for the GMH.

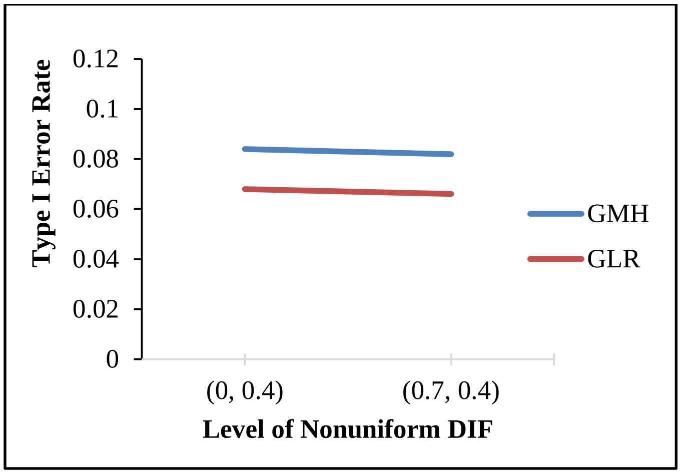

Figure 24: Type I Error Rates for 6 Groups Nonuniform DIF – One Focal Groups Experiencing DIF

Power rate of the GLR was slightly higher than the power rate of the GMH for nonuniform DIF with only *a* parameter change (See Figure 25). There was no difference between power rates of the GMH and the GLR when it was nonuniform DIF when one focal group experienced DIF with 6 groups. GLR also had the same power rate for both types of nonuniform DIF. It was not higher than 0.1 for both methods.

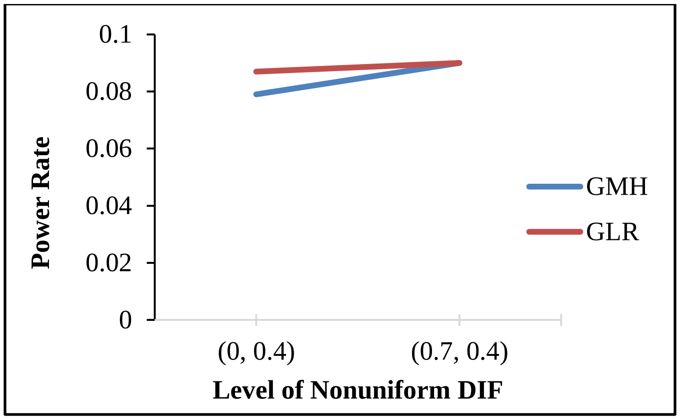

Figure 25: Power Rates for 6 Groups Nonuniform DIF - One Focal Groups Experiencing DIF

Precision rate of the GMH when it was nonuniform DIF with only *a* parameter change when one focal group experienced DIF with 6 groups was around 0.1 and it was around 0.125 for the GLR (See Figure 26). When it was nonuniform DIF with both *a* and *b* parameter change, both methods had the same precision rate that was around 0.11.

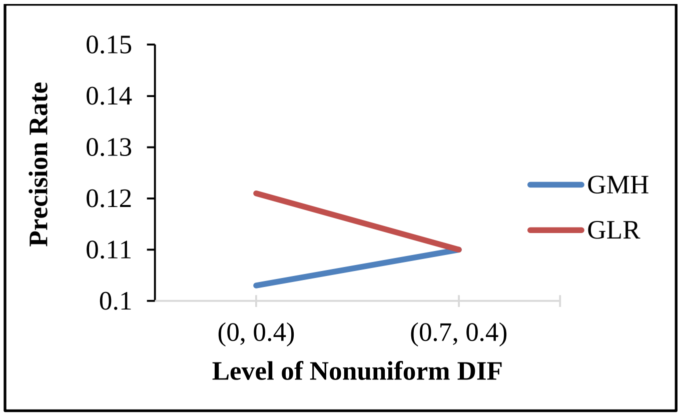

Figure 26: Precision Rates for 6 Groups Nonuniform DIF – One Focal Groups Experiencing DIF

Type I error rate of the GMH was consistently higher than the type I error rate of the GLR for both type of nonuniform DIF when half of the groups including reference group experienced DIF for 6 groups (See Figure 27). For both GMH and GLR, type I error rate of nonuniform DIF with only *a* parameter change was slightly higher than the type I error rate of nonuniform DIF with both *a* and *b* parameter change.

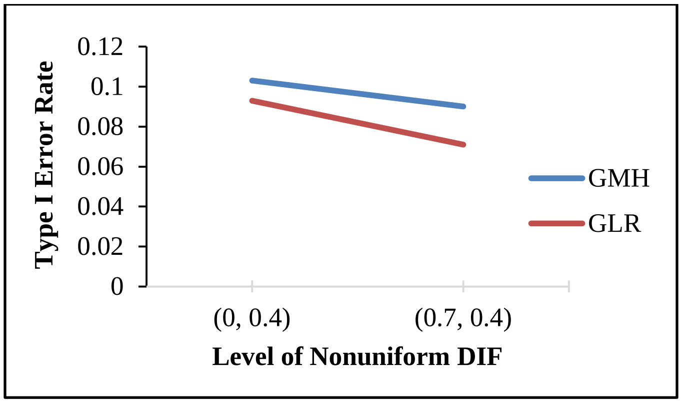

Figure 27: Type I Error Rates for 6 Groups Nonuniform DIF – Half of the Groups including Reference Group Experiencing DIF

Power rate of the GLR was higher than the power rate of the GMH when it was nonuniform DIF with only *a* parameter change when half of the groups including reference group experienced DIF for 6 groups (See Figure 28). There was no power rate difference for the GMH between two types of nonuniform DIF. However, power rate of the GLR decreased when it was nonuniform DIF with both *a* and *b* parameters change.

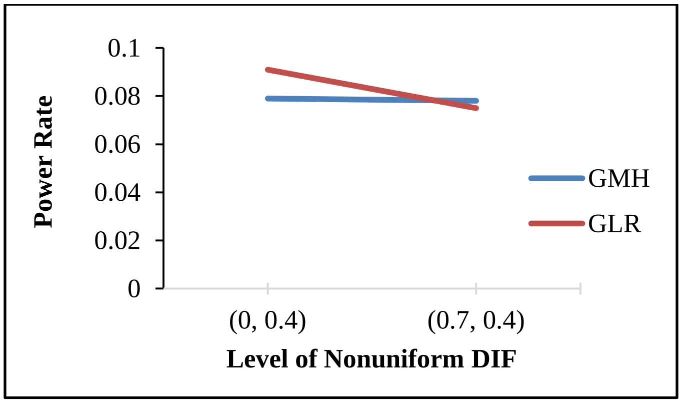

Figure 28: Power Rates for 6 Groups Nonuniform DIF – Half of the Groups including Reference Group Experiencing DIF

Precision rate of the GLR was higher than the precision rate of the GMH for both type of nonuniform DIF when half of the groups including reference group experienced DIF for 6 groups (See Figure 29). GMH had higher precision rate when it was nonuniform DIF with both *a* and *b* parameter change than with only *a* parameter change; whereas, it was almost same for the GLR. The lowest value was around 0.08 and the highest value was 0.095 for the GMH.

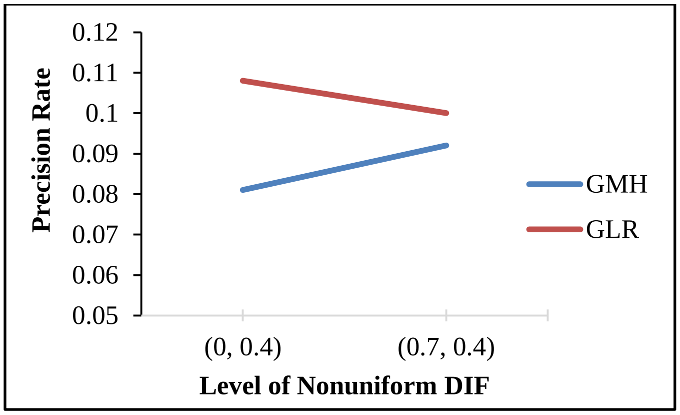

Figure 29: Precision Rates for 6 Groups Nonuniform DIF – Half of the Groups including Reference Group Experiencing DIF

When half of the groups experienced nonuniform DIF for 6 groups, type I error rate of the GMH was consistently higher than the type I error rate of GLR (See Figure 30). There was not much difference between two type of nonuniform DIF with respect to type I error rate of each method. The rate was above the nominal level of .05 for all cases.

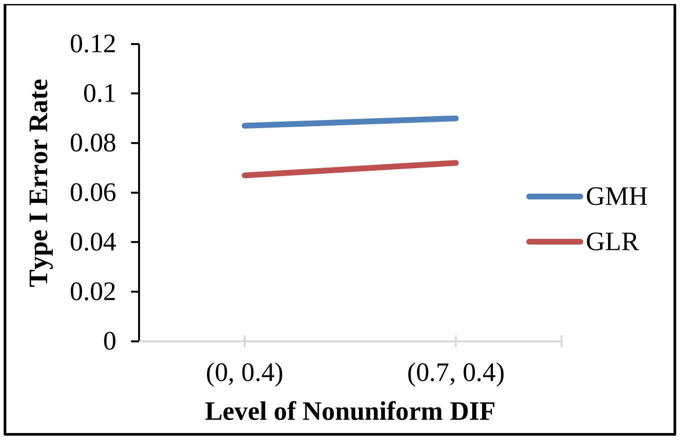

Figure 30: Type I Error Rates for 6 Groups Nonuniform DIF – Half of the Focal Groups Experiencing DIF

For power rate, both methods had similar value for nonuniform DIF with only a parameter change when half of the focal groups experienced DIF for 6 groups (See Figure 31). GLR had the similar values for both type of nonuniform DIF. GLM had slightly higher value for nonuniform DIF with both *a* and *b* parameter change.

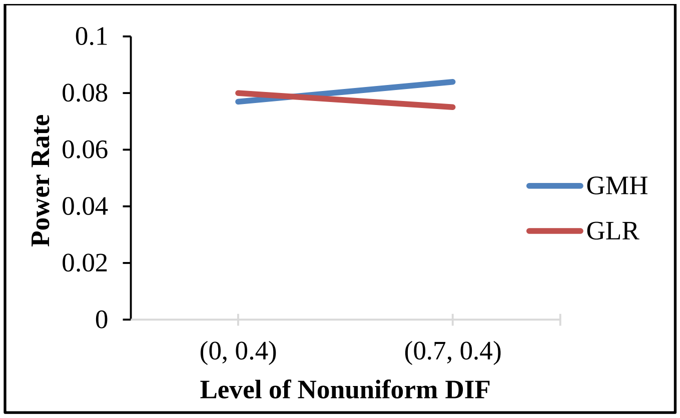

Figure 31: Power Rates for 6 Groups Nonuniform DIF – Half of the Focal Groups Experiencing DIF

Precision rate of the GLR was higher than the precision rate of the GMH for nonuniform DIF with only *a* parameter change when half of the focal groups experienced DIF for 6 groups (See Figure 32). Two methods had the same value when it was nonuniform DIF with both *a* and *b* parameter change. It was around 0.1.

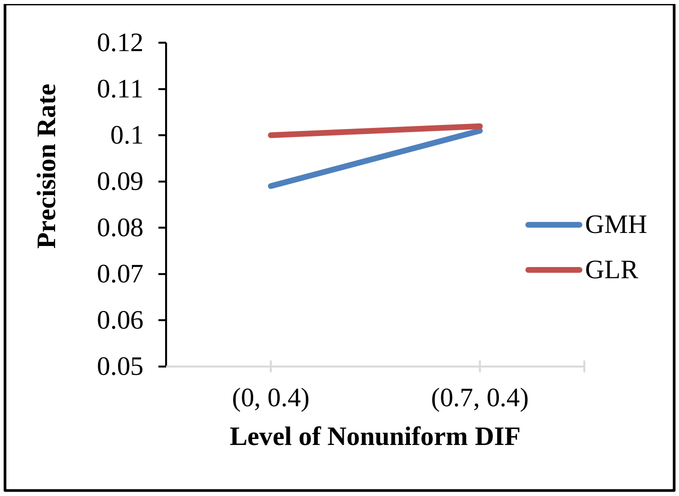

Figure 32: Precision Rates for 6 Groups Nonuniform DIF – Half of the Focal Groups Experiencing DIF

When all focal groups experienced nonuniform DIF for 6 groups, type I error rate of the GMH was consistently higher than the type I error rate of the GLR (See Figure 33). There was no difference between two types of nonuniform DIF with respect to type I error rate of each method.

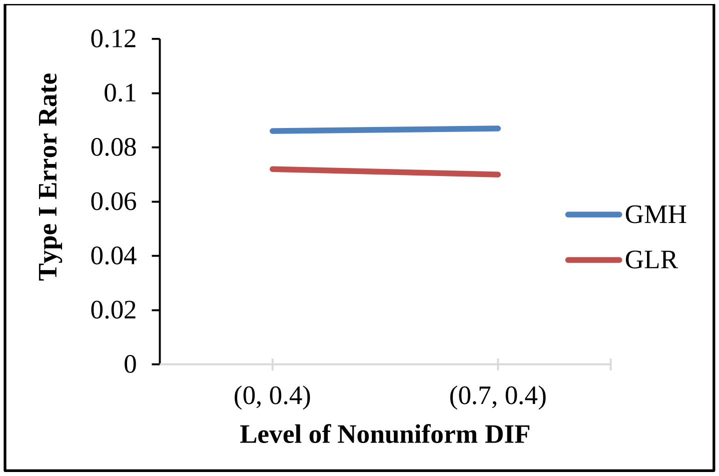

Figure 33: Type I Error Rates for 6 Groups Nonuniform DIF – All Focal Groups Experiencing DIF

Power rates of both methods were close to each other when it was nonuniform DIF with only *a* parameter change when all focal groups experienced DIF for 6 groups (See Figure 34). The values of two methods were same for nonuniform DIF with both *a* and *b* parameter change. Both methods had higher power rates when it was nonuniform DIF with only *a* parameter change.

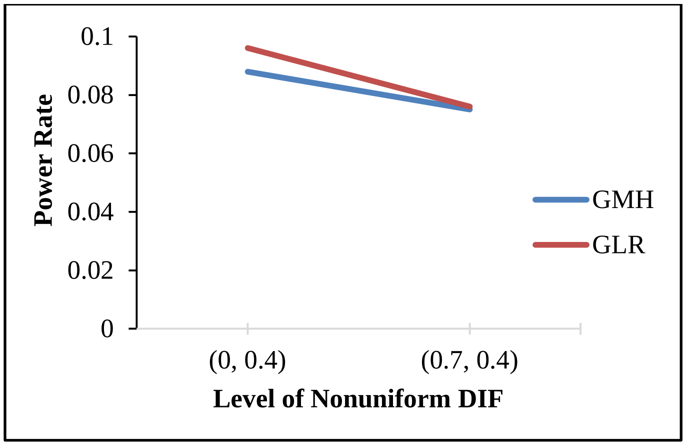

Figure 34: Power Rates for 6 Groups Nonuniform DIF –All Focal Groups Experiencing DIF

Precision rate of the GLR was higher than the precision rate of the GLR for nonuniform DIF with only *a* parameter change when all focal groups experienced DIF for 6 groups (See Figure 35). GMH had the same power rate for both types of nonuniform DIF; whereas, power rate of the GLR was lower for nonuniform DIF with both *a* and *b* parameter change.

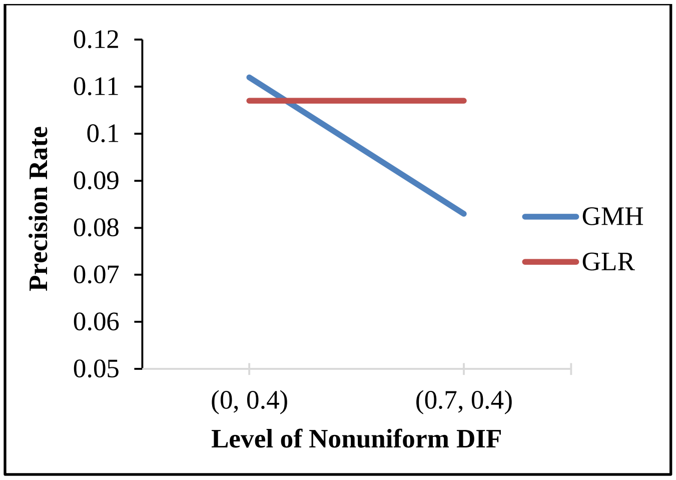

Figure 35: Precision Rates for 6 Groups Nonuniform DIF – All Focal Groups Experiencing DIF

When it was only reference group experienced DIF for 6 groups, both methods had the similar type I error rates (See Figure 36). They had higher type I error rates when it nonuniform DIF with both *a* and *b* parameter change.

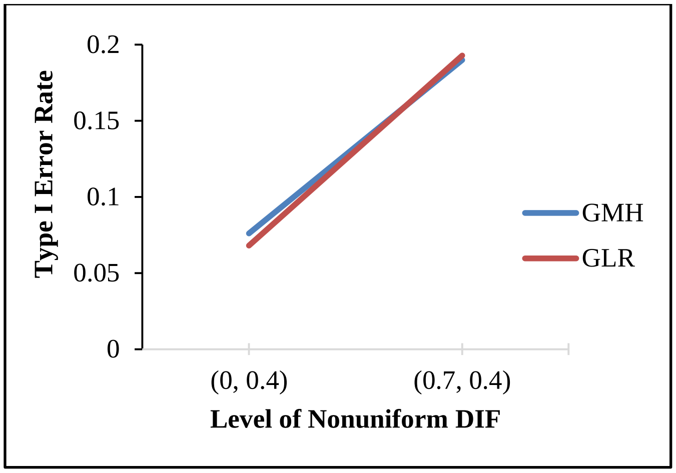

Figure 36: Type I Error Rates for 6 Groups Nonuniform DIF – Reference Group Experiencing DIF

Power rate of GLR was higher than the power rate of the GMH for nonuniform DIF with only a parameter change when only reference group experienced DIF (See Figure 37). When it was nonuniform DIF with both *a* and *b* parameter change, both methods had similar power rates that was almost 1. Both methods were able to detect all DIF items when it was nonuniform DIF with both *a* and *b* parameter change.

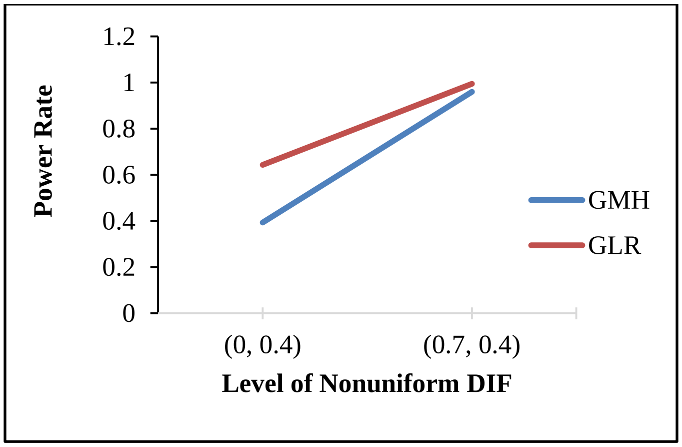

Figure 37: Power Rates for 6 Groups Nonuniform DIF – Reference Group Experiencing DIF

Lastly, precision rate of the GLR was consistently higher than the precision rate of the GMH for both type of nonuniform DIF (See Figure 38). The rates were higher when it was nonuniform DIF with both *a* and *b* parameter change for both methods. The lowest precision rate was around 0.4 and the highest rate was around 0.55 for the GMH; whereas, it was around 0.6 for the GLR for both type of nonuniform DIF.

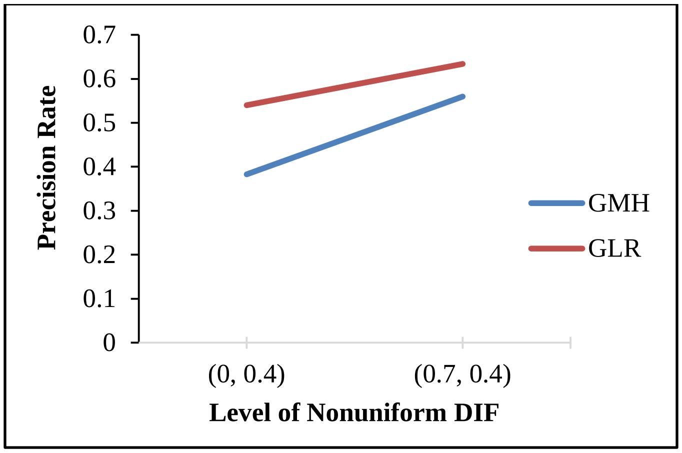

Figure 38: Precision Rates for 6 Groups Nonuniform DIF – Reference Group Experiencing DIF

Summary

In general, GLR had lower type I error rate for all cases with uniform DIF and when no DIF existed it also had lower type I error rate. GMH had lower precision rate for all cases with uniform DIF. For nonuniform DIF case, when it was only a parameter change, GLR usually had higher power and precision rates. However, when it was nonuniform DIF with both a and *b* parameter change, both methods had similar power rates; whereas; GLR had better precision rate. Type I error rate for both methods with all condition was above the nominal level of 0.05. Power rate ranged between 0.06 and 0.95 for both methods. When it was reference group experiencing any kind of DIF, power and precision rate was the highest for both methods. Table 9 displays all the values for type I, power and precision rates of all cases for 6 groups.

Table 9:

Results for 6 Groups

		GMH			GLR		
		Type I	Power	Precision	Type I	Power	Precision
0		0.085	-	-	0.071	-	-
	F1	0.082	0.082	0.087	0.067	0.065	0.095
	R1, F1, F2	0.084	0.08	0.095	0.063	0.061	0.104
0.4	F3, F4, F5	0.083	0.101	0.115	0.066	0.082	0.12
	F1, F2, F3, F4, F5	0.085	0.094	0.114	0.066	0.083	0.127
	R	0.078	0.651	0.510	0.061	0.600	0.550
	F1	0.08	0.085	0.10	0.063	0.063	0.098
	R1, F1, F2	0.086	0.092	0.102	0.071	0.08	0.105
0.6	F3, F4, F5	0.085	0.087	0.105	0.067	0.060	0.09
	F1, F2, F3, F4, F5	0.078	0.078	0.106	0.07	0.061	0.087
	R	0.075	0.775	0.553	0.060	0.740	0.610
	F1	0.089	0.085	0.092	0.068	0.057	0.082
	R1, F1, F2	0.082	0.071	0.091	0.069	0.056	0.082
0.8	F3, F4, F5	0.088	0.081	0.091	0.066	0.074	0.102
	F1, F2, F3, F4, F5	0.10	0.072	0.08	0.078	0.057	0.08
	R	0.085	0.80	0.534	0.073	0.80	0.58

	F1	0.084	0.079	0.103	0.068	0.087	0.121
	R1, F1, F2	0.103	0.079	0.081	0.093	0.091	0.108
(0, 0.4)	F3, F4, F5	0.087	0.077	0.089	0.067	0.08	0.10
	F1, F2, F3, F4, F5	0.086	0.088	0.112	0.072	0.096	0.107
	R	0.076	0.393	0.383	0.068	0.643	0.540
	F1	0.082	0.09	0.11	0.066	0.09	0.11
	R1, F1, F2	0.09	0.078	0.101	0.071	0.075	0.102
(0.7, 0.4)	F3, F4, F5	0.090	0.084	0.101	0.072	0.075	0.102
	F1, F2, F3, F4, F5	0.087	0.075	0.083	0.07	0.076	0.107
	R	0.190	0.96	0.560	0.193	0.995	0.634

Twelve Groups

Uniform DIF

For 12 groups, when only one focal group experienced uniform DIF, type I error rate of both methods were very similar (See Figure 39). When there was no DIF, type I error rate was around 0.12 for two methods. There was not much difference between small and large magnitude of uniform DIF (around 0.1); whereas it was below the nominal level of 0.05 when it was small magnitude of uniform DIF.

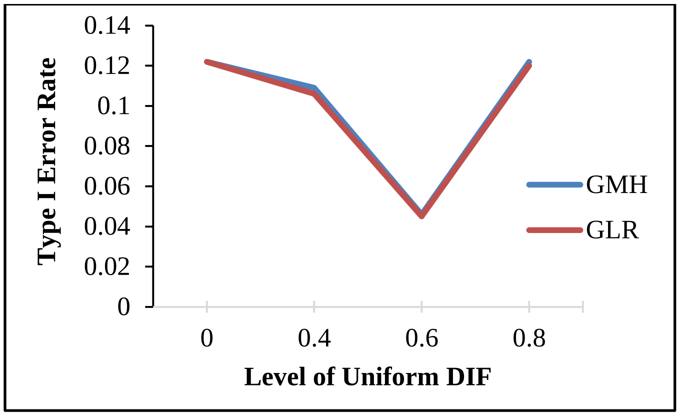

Figure 39: Type I Error Rates for 12 Groups Uniform DIF – One Focal Group Experiencing DIF

Power rate of both methods were also similar when one focal group experienced DIF for 12 groups (See Figure 40). When it was medium magnitude of uniform DIF, both methods' power performance was really low (around 0.03). The highest value was around 0.13 for the GLR and it was around 0.15 for the GMH.

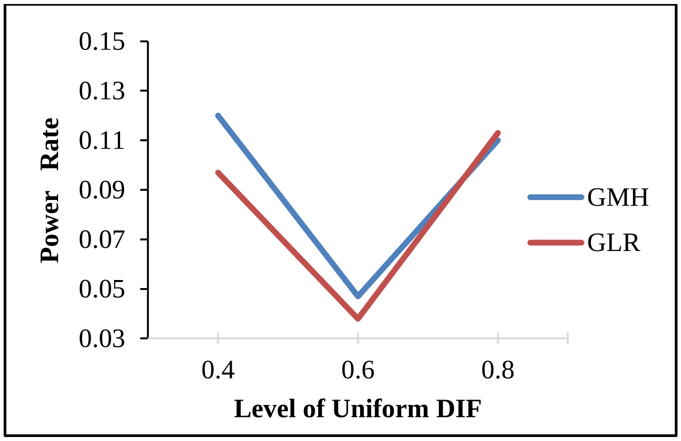

Figure 40: Power Rates for 12 Groups Uniform DIF – One Focal Group Experiencing DIF

Precision rate of the GMH was slightly higher than the precision rate of the GLR only one focal experienced small magnitude of uniform DIF for 12 groups (See Figure 41). For medium and large magnitude of uniform DIF both methods had similar precision rates and there was not much difference between these two magnitudes of uniform DIF with respect to precision rates. It was around 0.1 for all cases.

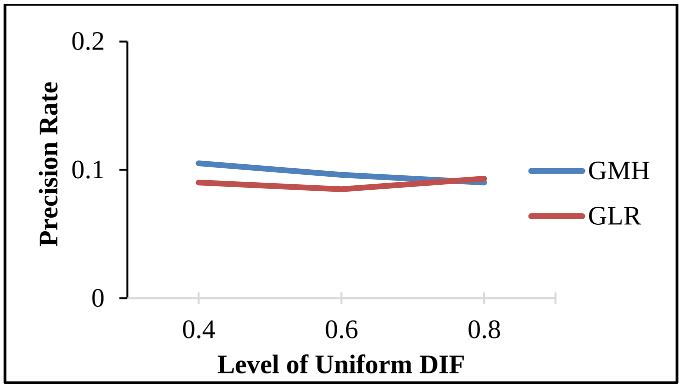

Figure 41: Precision Rates for 12 Groups Uniform DIF – One Focal Group Experiencing DIF

Type I error rate of GMH was 0.02 more than the type I error rate of the GLR when half of the groups including reference group experienced DIF (See Figure 42). When it was medium and large magnitude of uniform DIF, both methods had similar type I error rates. It was around 0.05 for medium magnitude of uniform DIF and 0.12 for large magnitude of uniform DIF.

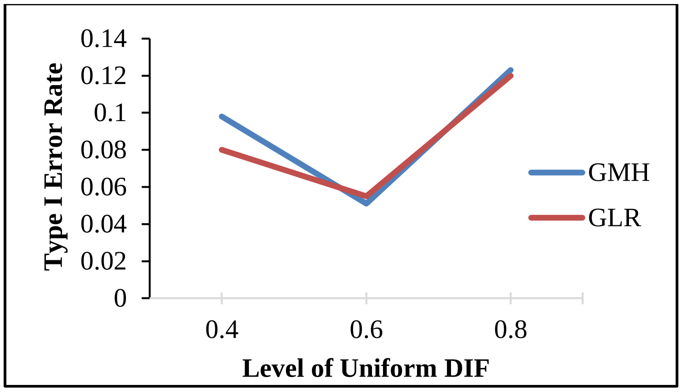

Figure 42: Type I Error Rates for 12 Groups Uniform DIF - Half of the Groups including Reference Group Experiencing DIF

Power rate of the GLR was slightly higher than the power rate of the GMH for medium magnitude of uniform DIF when half of the groups including reference group experienced DIF (See Figure 43). Both methods had same and higher power rates for small and large magnitude of uniform DIF that was around 0.11.

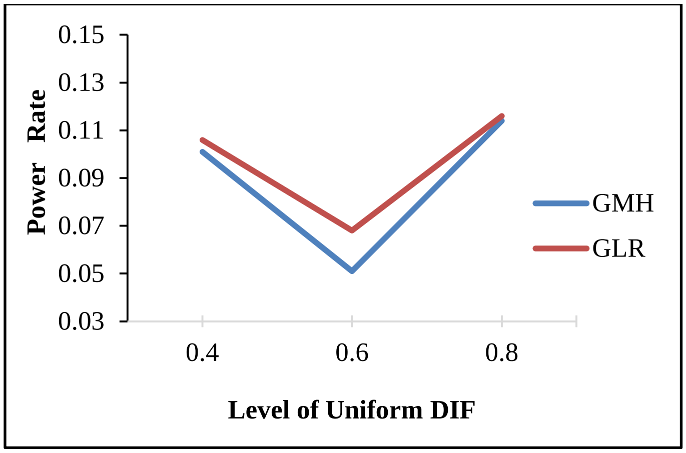

Figure 43: Power Rates for 12 Groups Uniform DIF – Half of the Groups including Reference Group Experiencing DIF

Precision rate of the GLR was slightly higher than the precision rate of the GMH for all magnitudes of uniform DIF when half of the groups including reference groups experienced DIF (See Figure 44). There was not much difference between small and medium magnitude of uniform DIF; whereas both methods had smaller value for large magnitude of uniform DIF.

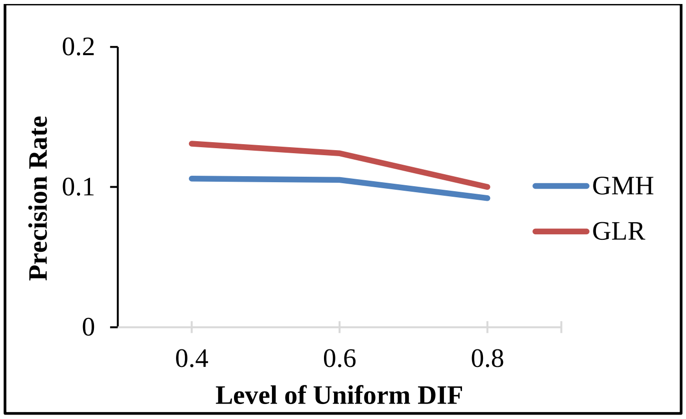

Figure 44: Precision Rates for 12 Groups Uniform DIF – Half of the Groups including Reference Group Experiencing DIF

Both methods perform very similar with respect to their type I error rates when half of the focal groups experienced DIF for 12 groups. They had the smallest value when it was medium magnitude of uniform DIF that was at the nominal level of 0.05. For large magnitude of uniform DIF, this value was around 0.12.

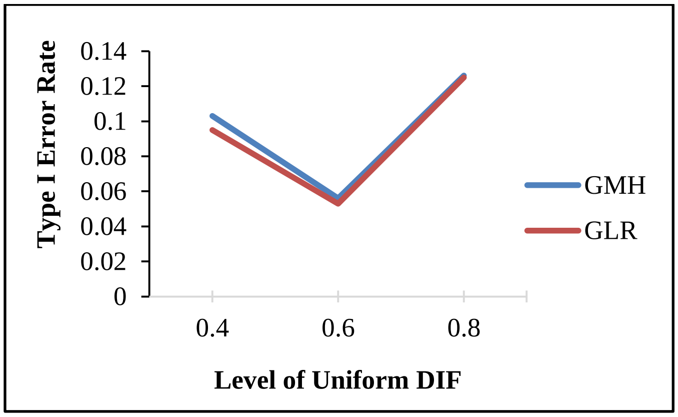

Figure 45: Type I Error Rates for 12 Groups Uniform DIF – Half of the Focal Groups Experiencing DIF

Power rate of the GMH was higher than the power rate of the GLR for small magnitude of uniform DIF (0.11 versus 0.09) when half of the focal groups experienced uniform DIF for 12 groups (See Figure 46). Both methods had the smallest power rates when it was medium magnitude of uniform DIF (around 0.05) and biggest power rate when it was large magnitude of uniform DIF (around 0.13).

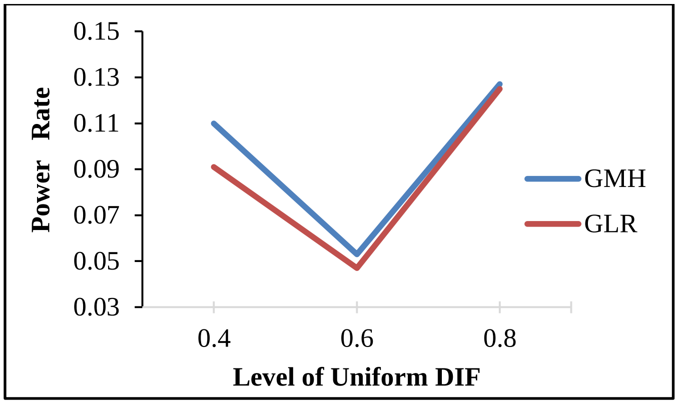

Figure 46: Power Rates for 12 Groups Uniform DIF – Half of the Focal Groups Experiencing DIF

Precision rates of two methods were quite similar when half of the focal groups experienced DIF for 12 groups (See Figure 47). There was not much difference among the magnitudes of uniform DIF. It was around 0.1 for all cases.

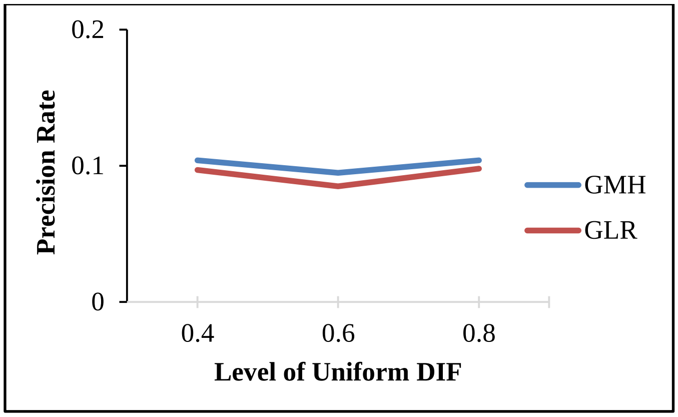

Figure 47: Precision Rates for 12 Groups Uniform DIF – Half of the Focal Groups Experiencing DIF

Type I error rate of both methods were very close to each other when all focal groups experienced DIF for 12 groups (See Figure 48). For medium magnitude of uniform DIF, both methods had the smallest type I error rate that was at the nominal level. There was not much difference between small and large magnitude of uniform DIF for both methods, it was around 0.1.

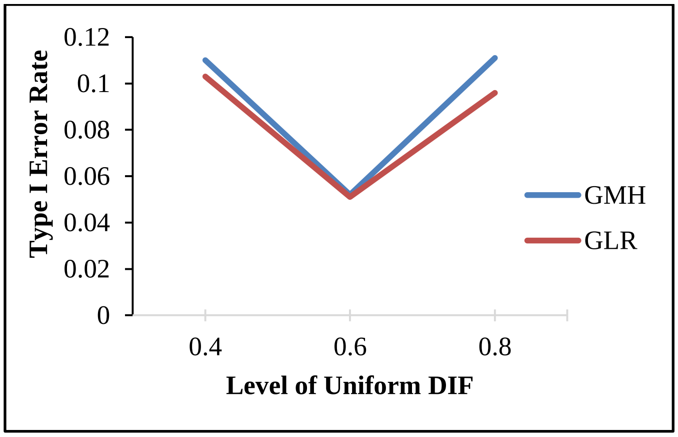

Figure 48: Type I Error Rates for 12 Groups Uniform DIF – All Focal Groups Experiencing DIF

Both methods had same power rates for all magnitude of uniform DIF when all focal groups experienced DIF for 12 groups (Figure 49). There was no difference between small and large magnitude of uniform DIF (around 0.13). Both methods had the smallest power rate when it was medium magnitude of uniform DIF (around 0.05).

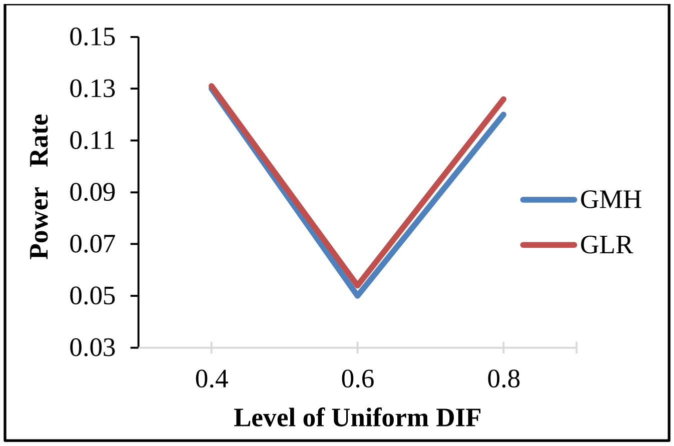

Figure 49: Power Rates for 12 Groups Uniform DIF – All Focal Groups Experiencing DIF

Precision rate of both methods were similar when all focal groups experienced small magnitude of uniform DIF for 12 groups (See Figure 50). For medium and large magnitude of uniform DIF, precision rate of the GLR was higher than the precision rate of the GMH. GMH had same precision rate for medium and large magnitude of uniform DIF.

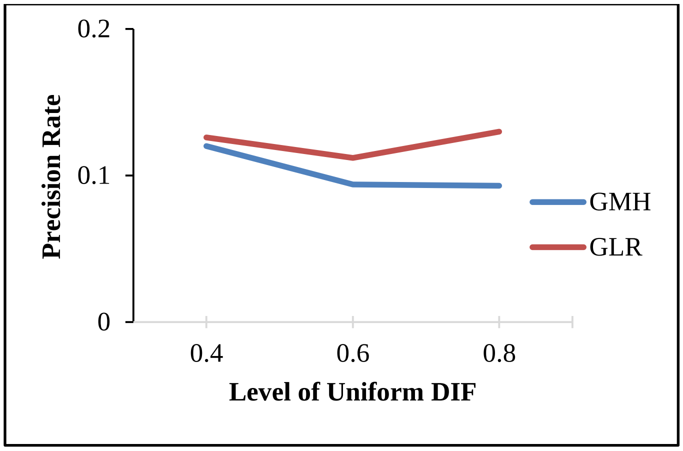

Figure 50: Precision Rates for 12 Groups Uniform DIF – All Focal Groups Experiencing DIF

Both methods had almost identical type I error rates when it was reference group experiencing DIF (See Figure 51). When it was medium magnitude of uniform DIF, type I error rates of both methods were at the nominal level of 0.05. For large magnitude of uniform DIF, the rate was around 0.13.

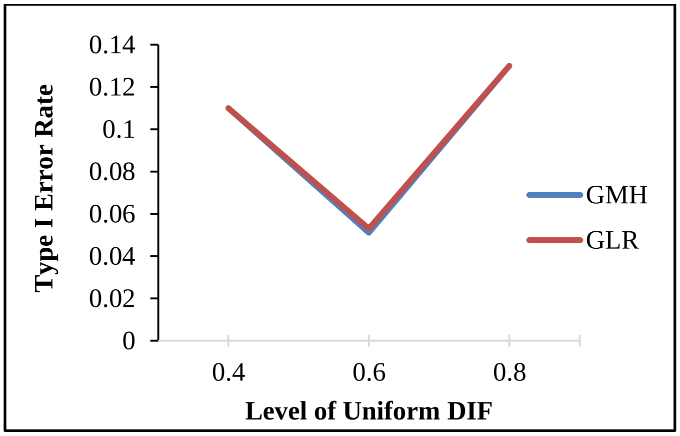

Figure 51: Type I Error Rates for 12 Groups Uniform DIF - Reference Group Experiencing DIF

Power rates of two methods were also almost identical when it was reference group experiencing DIF for 12 groups (See Figure 52). It was really high for large magnitude of uniform DIF (around 0.8) and lowest for medium magnitude of uniform DIF (around 0.11).

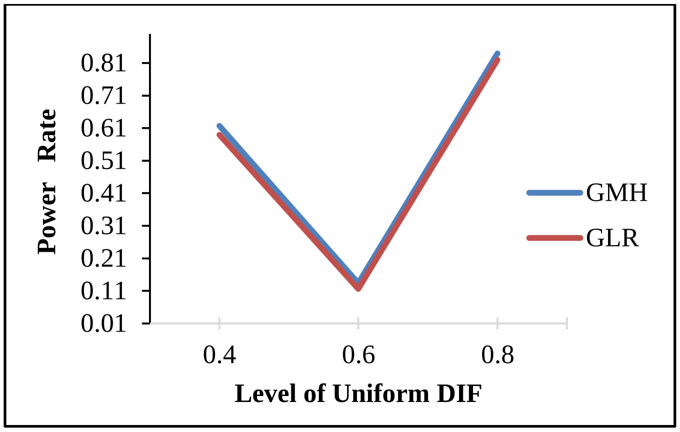

Figure 52: Power Rates for 12 Groups Uniform DIF – Reference Group Experiencing DIF

As expected from type I error and power rates of both methods' being so close to each other, precision rates of both methods were also very similar to each other for all magnitude of uniform DIF when reference groups experienced DIF for 12 groups (See Figure 53). Precision rate of the GMH for small magnitude of uniform DIF was slightly higher than the precision rate of the GLR.

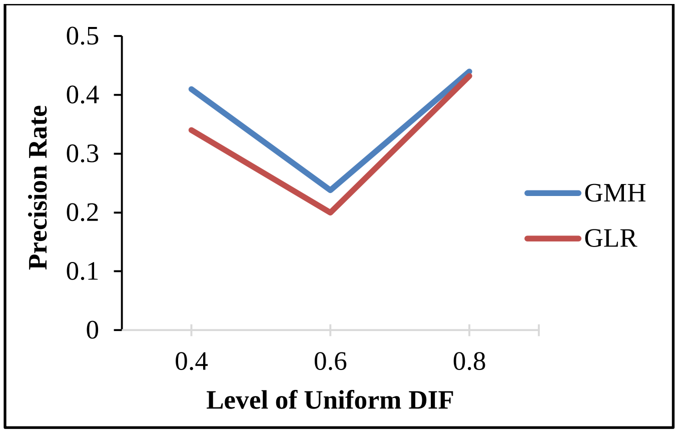

Figure 53: Precision Rates for 12 Groups Uniform DIF – Reference Group Experiencing DIF

Nonuniform DIF

Type I error rate of the GMH was slightly than type I error rate of the GLR for nonuniform DIF with only a parameter change when one focal group experienced DIF for 12 groups (See Figure 54). There was not much difference between two types non uniform DIF for each method with respect to their type I error rates.

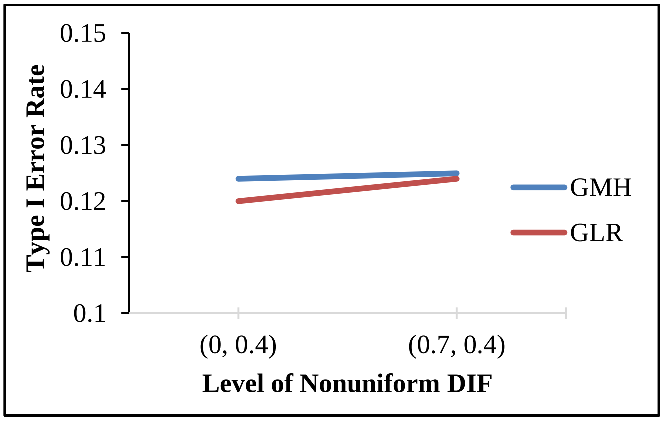

Figure 54: Type I Error Rates for 12 Groups Nonuniform DIF –One Focal Group Experiencing DIF

Power rate of the GLR was higher than the power rate of the GMH for nonuniform DIF with only *a* parameter change than power rate of the GMH when one focal group experienced DIF for 12 groups (See Figure 55). They had similar power rates when it was nonuniform DIF with both *a* and *b* parameter change that was around 0.135.

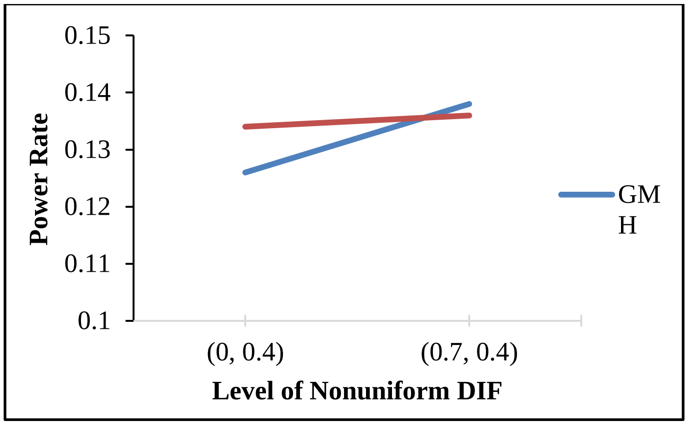

Figure 55: Power Rates for 12 Groups Nonuniform DIF –One Focal Group Experiencing DIF

Precision rate of two methods were identical when one focal group experienced DIF (See Figure 56). There was no difference between two types of nonuniform DIF for each method. The rate was around 0.1.

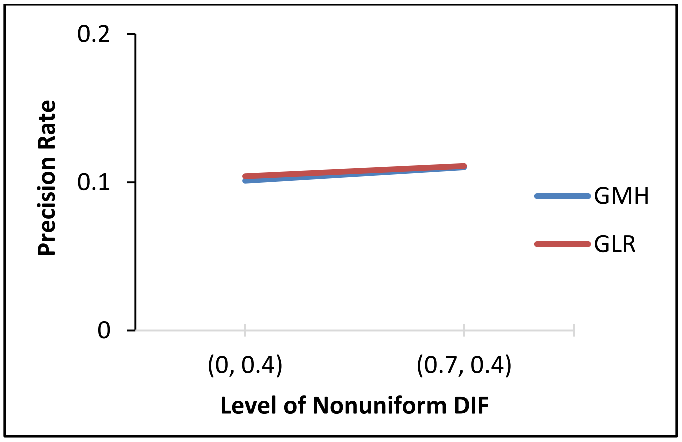

Figure 56: Precision Rates for 12 Groups Nonuniform DIF – One Focal Group Experiencing DIF

When half of the groups including reference group experienced nonuniform DIF, type I error rate of the GMH was consistently higher than type I error rate of the GMR for both types nonuniform DIF (See Figure 57). Type I error rate of nonuniform DIF with both *a* and *b* parameter change was slightly higher than type I error rate of nonuniform DIF with only *a* parameter change.

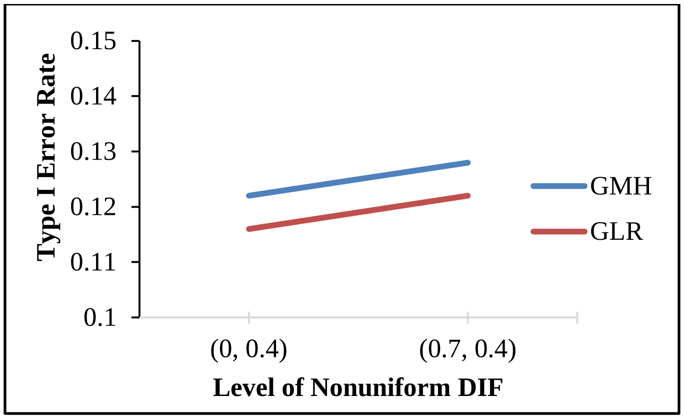

Figure 57: Type I Rates for 12 Groups Nonuniform DIF – Half of the Groups including Reference Group Experiencing DIF

Power rates of the GMH was slightly higher than the power rate of the GLR when half of the groups including reference group experienced nonuniform DIF with only a parameter change for 12 groups (See Figure 58). Both methods had similar power rates when it was nonuniform DIF with both *a* and *b* parameter change (around 0.13).

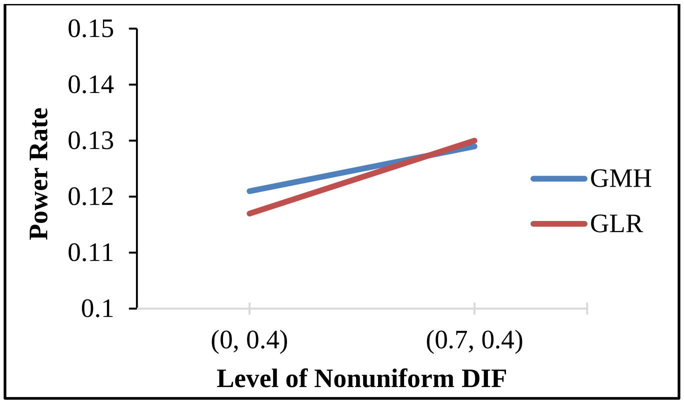

Figure 58: Power Rates for 12 Groups Nonuniform DIF – Half of the Groups including Reference Group Experiencing DIF

Precision rates of both methods were very similar for both types of nonuniform DIF when half of the groups including reference group experienced DIF (See Figure 59). There was not much difference for both types of nonuniform DIF. The rate was around 0.11.

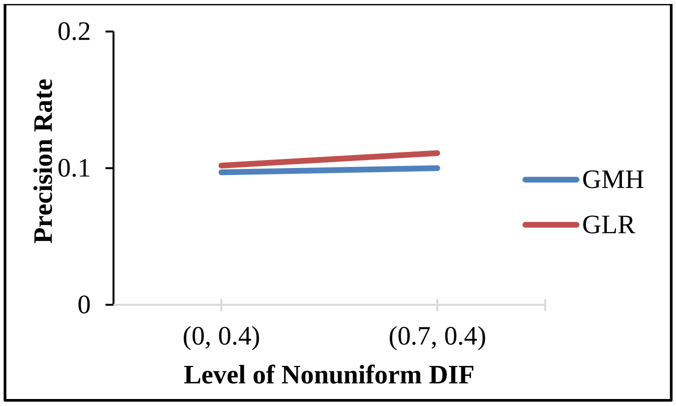

Figure 59: Precision Rates for 12 Groups Nonuniform DIF – Half of the Groups including Reference Group Experiencing DIF

When half of the focal groups experienced nonuniform DIF, type I error rate of the GMH was slightly higher than the type I error rate GLR (See Figure 60). Type I error rate of the GMH was around 0.125 and type I error rate of GLR was around 0.12 for both type of nonuniform DIF.

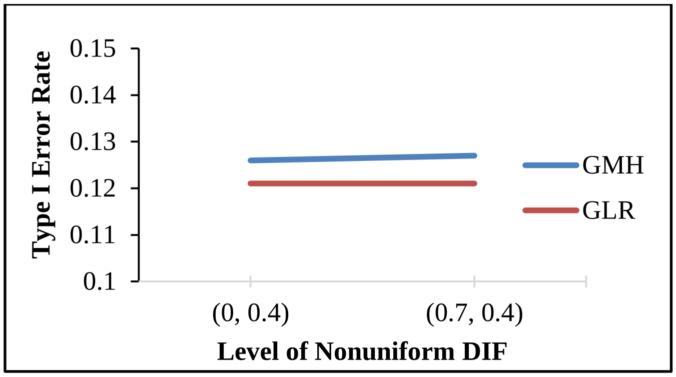

Figure 60: Type I Error Rates for 12 Groups Nonuniform DIF –Half of the Focal Groups Experiencing DIF

Power rate of the GMH was slightly higher than the power rate of the GLR when half of the focal groups experienced nonuniform with only a parameter change for 12 groups (See Figure 61). There was no difference on type I error rate of GMH between two types of nonuniform; whereas, power rate of GLR got decreased when it was nonuniform DIF with both *a* and *b* parameter change.

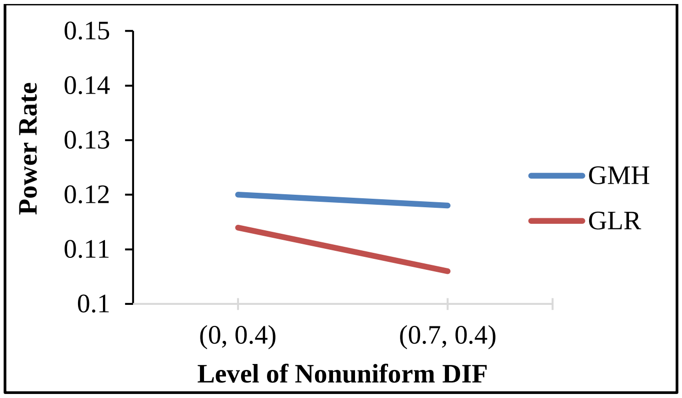

Figure 61: Power Rates for 12 Groups Nonuniform DIF – Half of the Focal Groups Experiencing DIF

For precision rate, the GLR had slightly higher value for nonuniform DIF with only a parameter change; whereas, the GMH had slightly higher value for nonuniform DIF with both a and b parameter change when half of the focal groups experienced DIF for 12 groups (See Figure 62). There was not big difference between two types of nonuniform DIF for GMH with respect to precision rate. For the GLR, the precision rate was around 0.11 for nonuniform DIF with only *a* parameter change and around 0.09 for nonuniform DIF with both *a* and *b* parameter change.

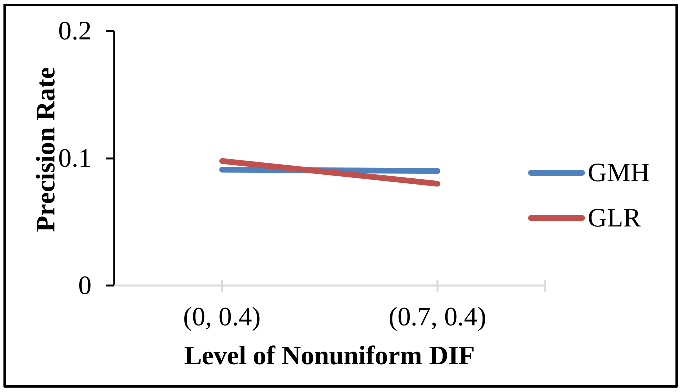

Figure 62: Precision Rates for 12 Groups Nonuniform DIF – Half of the Focal Groups Experiencing DIF

Type I error rate of the GMH was higher than the type I error rate of the GLR when all focal groups experienced nonuniform DIF with only *a* parameter change for 12 groups (See Figure 63). Both methods had similar type I error rates when it was nonuniform DIF with both *a* and *b* parameter change. It was around 0.125.

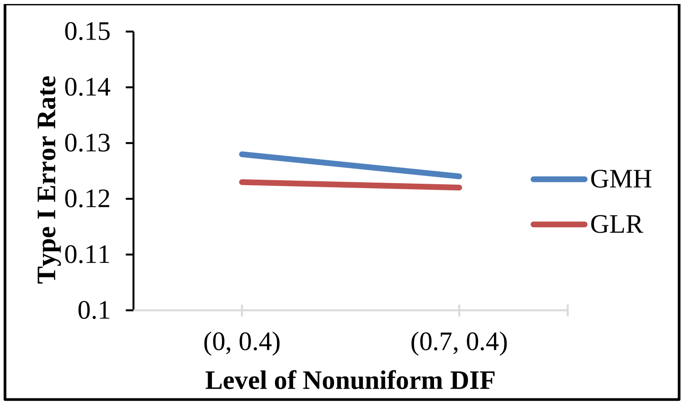

Figure 63: Type I Error Rates for 12 Groups Nonuniform DIF – All Focal Groups Experiencing DIF

For the power rate, the GMH had higher value than the GLR for nonuniform DIF with only *a* parameter chance when all focal groups experienced DIF (See Figure 64). The rate was around 0.12 for the GMH and around 0.105 for the GLR. For nonuniform DIF with both *a* and *b* parameter chance, the GLR had higher value than the GMH (0.12 versus 0.11).

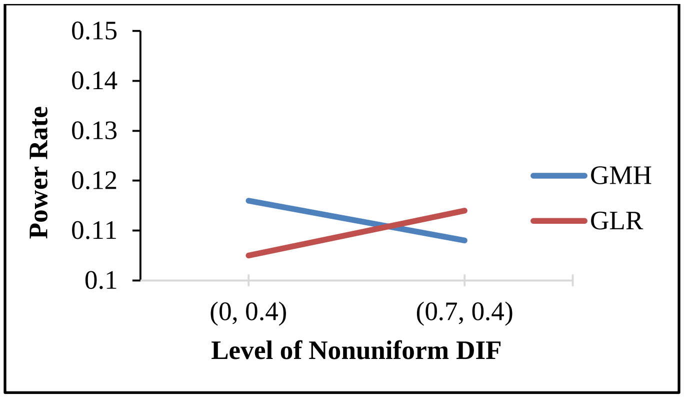

Figure 64: Power Rates for 12 Groups Nonuniform DIF – All Focal Groups Experiencing DIF

Precision rate of both methods were almost identical for both types of nonuniform DIF when all focal groups experienced DIF (See Figure 65). There was not difference between two types nonuniform DIF with respect to precision rates of each method. The rate was around 0.1.

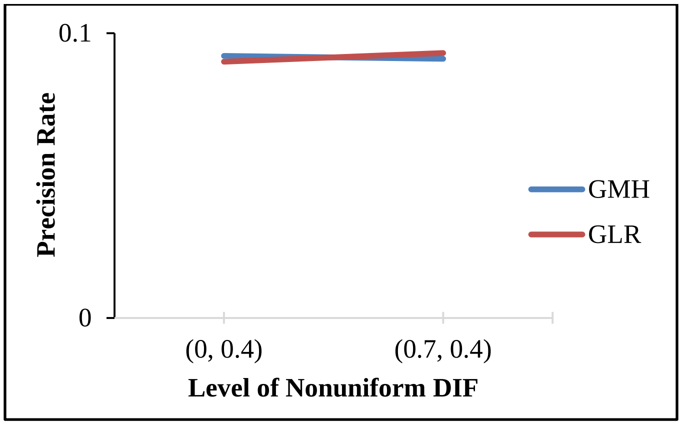

Figure 65: Precision Rates for 12 Groups Nonuniform DIF – All Focal Groups Experiencing DIF

Lastly, when it was reference group experienced nonuniform DIF with only a parameter change, both methods had same type I error rates that was around 0.11 for 12 groups (See Figure

66). GMH had higher type I error rate than the GLR when it was nonuniform DIF with both *a* and *b* parameter change (0.14 versus 0.12).

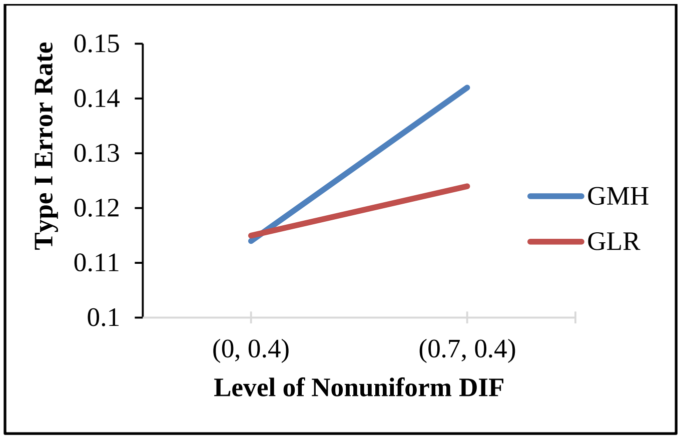

Figure 66: Type I Error Rates for 12 Groups Nonuniform DIF – Reference Group Experiencing DIF

For power rate, the GLR had higher value than the GMH for nonuniform DIF with only *a* parameter change when reference group experienced DIF for 12 groups (See Figure 67). The power rate was around 0.65 for the GLR and around 0.4 for the GMH when it was nonuniform DIF with only *a* parameter change. Two methods had close power rates when it was nonuniform DIF with both *a* and *b* parameter change. The power rate was almost 1.

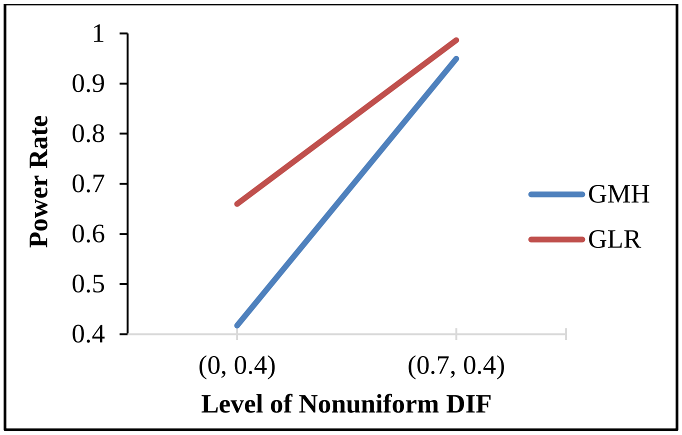

Figure 67: Power Rates for 12 Groups Nonuniform DIF –Reference Group Experiencing DIF

Precision rate of the GLR was higher than the precision rate of the GMH when it was reference group experiencing nonuniform DIF with only *a* parameter change DIF for 12 groups (See Figure 68). The rate was around 0.4 for the GLR and 0.3 for the GMH. Two methods had same precision rates for nonuniform DIF with both *a* and *b* parameter change. The rate was around 0.45.

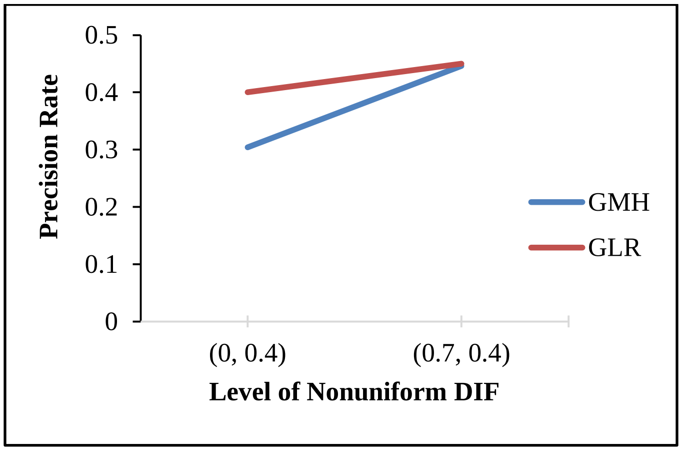

Figure 68: Precision Rates for 12 Groups Nonuniform DIF –Reference Group Experiencing DIF

Summary

In contrast to 6 groups, for 12 groups there was not much difference on two methods' type I error rate for many cases. Only for nonuniform DIF for both *a* and *b* parameter change, the GLR had lower type I error rate than the GMH. Similar to 6 groups, when it was reference group any kind of DIF, power and precision rate of two methods were really high. Overall, precision rate of the GLR was higher than the precision rate of the GMH. Table 10 displays the values of type I error, power and precision rates of all cases for 12 groups.

Table 10:

Results for 12 Groups

		GMH			GLR		
		Type I	Power	Precision	Type I	Power	Precision
0		0.122	-	-	0.122	-	-
0.4	F1	0.109	0.120	0.105	0.106	0.097	0.090
	R1, F1, F2, F3, F4, F5	0.098	0.101	0.106	0.080	0.106	0.131
	F6, F7, F8, F9, F10, F11	0.103	0.111	0.104	0.095	0.091	0.097
	F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11	0.110	0.130	0.120	0.103	0.131	0.126
0.6	R	0.110	0.617	0.410	0.110	0.590	0.340
	F1	0.046	0.047	0.096	0.045	0.038	0.085
	R1, F1, F2, F3, F4, F5	0.051	0.051	0.105	0.055	0.068	0.124
	F6, F7, F8, F9, F10, F11	0.056	0.053	0.095	0.053	0.047	0.085
	F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11	0.052	0.05	0.094	0.051	0.054	0.112

	R	0.051	0.134	0.238	0.053	0.116	0.200
0.8	F1	0.122	0.110	0.090	0.120	0.113	0.093
	R1, F1, F2, F3, F4, F5	0.123	0.114	0.092	0.120	0.116	0.100
	F6, F7, F8, F9, F10, F11	0.126	0.127	0.104	0.125	0.125	0.098
	F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11	0.111	0.12	0.093	0.096	0.126	0.130
	R	0.130	0.840	0.440	0.130	0.820	0.432
(0, 0.4)	F1	0.124	0.126	0.101	0.120	0.134	0.104
	R1, F1, F2, F3, F4, F5	0.122	0.121	0.097	0.116	0.117	0.102
	F6, F7, F8, F9, F10, F11	0.126	0.120	0.091	0.121	0.114	0.098
	F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11	0.128	0.116	0.092	0.123	0.105	0.090
	R	0.114	0.417	0.304	0.115	0.660	0.400
(0.7, 0.4)	F1	0.125	0.138	0.110	0.124	0.136	0.111
	R1, F1, F2, F3, F4, F5	0.128	0.129	0.100	0.122	0.13	0.111
	F6, F7, F8, F9, F10, F11	0.127	0.118	0.090	0.121	0.106	0.080
	F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11	0.124	0.108	0.091	0.122	0.114	0.093
	R	0.142	0.950	0.446	0.124	0.987	0.450

Chapter 5: DISCUSSION

This study compared the type I error, power and precision rates of the GMH and the GLR under a variety of uniform and nonuniform DIF conditions for multiple groups. To address the goals of this study, 50 dichotomously scored items were simulated with 10% of the items contaminated with DIF. Possible conditions within a real testing data, such as different groups experiencing DIF, was included in the study to provide a guideline for researchers with respect to particular DIF detection method that might work best. Also different number of groups was another manipulated variable. Two research questions given below were investigated in the study:

Research Question 1: In investigating uniform DIF, does the magnitude of DIF affect the performance of Generalized Mantel-Haenszel and Generalized Logistic Regression under different number of total groups and different groups experiencing DIF? What are the Type I error, power, and precision rates of these two methods for these conditions?

Research Question 2: In investigating nonuniform DIF, does the type of nonuniform DIF affect the performance of Generalized Mantel-Haenszel and Generalized Logistic Regression under different number of total groups and different groups experiencing DIF? What are the Type I error, power, and precision rates of these two methods for these conditions?

To answer each of these research questions, type I error, power and precision rates of parallel conditions were compared individually. The answer of first part of both research questions was ‘yes’. However, there were no generalized results for all cases regarding the magnitude of uniform DIF and the type of nonuniform DIF. Type I error, power and precision rate differences across groups was discussed separately to gain a better perspective of where two methods had some advantages over other.

Type I Error Rate Difference

One surprising result found in the study regarding the type I error rate was that the type I error rate was not controlled below the nominal level of .05 in many situations regardless of type and magnitude of DIF, and the number of total groups. However, the results demonstrated that the GLR method controlled the type I error rate better than the GMH for multiple groups regardless of the other simulation condition except the no-DIF condition and the nonuniform DIF with only *a* parameter change condition for 2 groups. When no-DIF existed, the MH had a better control of type I error than the LR (Type I error rate is 0.151 for the GMH is 0.163 for the GLR). When nonuniform DIF with only a parameter change existed, type I error rate was 0.248 for the GMH and 0.319 for the GLR. For different groups experiencing uniform DIF, Penfield (2001) found that type I error rate of GMH was mostly at the nominal level of .05 for a sample size of 1000 when groups had same N(0,1) ability distribution. In this study, type I error rate of the GMH was at the nominal level of .05 only when it was medium level of uniform DIF for 12 groups. One possible reason for differing results between Penfield's (2001) study and this study is due to Penfield (2001) using N(0,1) ability distribution and this study using the means of different groups in the normal distributions from a real data PISA 2012. Svetina and Rutkowski (2014) also found that GLR yielded an increased type I error in the absence of DIF. That was not the case in this study especially for the 6 groups condition. In his study, Finch (2015) found that the GMH and the GLR had an inflated type I error for some situations. This finding was consistent with the finding of this study. However, what was inconsistent between Finch (2015) and this study was the method that had the highest type I error inflation. Finch (2015) reported that the GLR had the highest type I error inflation; whereas, this study found that the GMH had the highest type I error inflation. One potential reason for these different findings is that this study investigated both uniform and nonuniform DIF items, while Finch (2015) investigated only uniform DIF items. In this study, the GMH had the highest type I error inflation for 12 groups nonuniform DIF with both *a* and *b* parameter change when reference groups experienced DIF (0.142). Another possible reason for differing results is that Finch (2015) simulated only one item to contain DIF. However, although same items were simulated in this study as well to contain DIF, five items were considered instead of one item.

To summarize, both methods had inflated type I error rates for many situations. Overall, the GMH had higher inflated type I error rates than the GLR. Type I error rate of both methods were strongly affected by the type of DIF and the number of groups that experienced DIF. Six groups had lower type I error rates than 12 groups in general.

Power Rate Difference

For two groups, as the magnitude of uniform DIF increased, the powers of two methods have increased as well. These findings were consistent with Swaminathan and Rogers (1990) whom also stated that the MH was slightly better with the power for uniform DIF that the current study supports it with its findings as well. However, even though they indicated that the LR procedure was specifically intended to detect nonuniform DIF hence it might not be effective in detecting strictly uniform DIF, the LR was as powerful as the MH for two groups uniform DIF cases in this study.

For nonuniform DIF with only *a* parameter change, the power rate of the GLR is a lot higher than the GMH (0.668 versus 0.412). GMH was still able to detect these items in contrast to Swaminathan and Rogers (1990)'s findings but the power rate was really low compared to the GLR. For nonuniform DIF with both *a* and *b* parameters change, both methods' power rates were higher with the GLR's being better. These findings were consistent with Narayanon and

Swaminathan (1996) and, Güler and Penfield (2009). They both agreed that the GMH's overall performance to detect noncrossing – nonuniform (only *a* parameter change) was comparable with the GLR but not to detect crossing-nonuniform DIF (both *a* and *b* parameters change). Rogers and Swaminathan (1993) also stated that LR was more powerful than MH for detecting nonuniform DIF and as powerful as detecting uniform DIF that is agreed in this study. It was surprising that the power rate of nonuniform DIF case for both six and 12 groups were highest when reference group experienced DIF. However, there is no research related to found results.

Also the study findings of Svetina and Rutkowski (2014) suggested that number of groups did not matter much for detecting DIF items that was mostly true for 6 and 12 groups. The results of this study demonstrated that overall the GMH had the highest power rates for uniform DIF conditions. This finding was consistent with Magis and De Boeck (2011) and, Finch (2015). Magis and De Boeck (2011) found that the MH had the highest power rates for two groups. Besides, Magis et al. (2011) indicated that the GMH and the GLR would perform similarly for uniform DIF with respect to power that was mainly consistent with the findings of this study. Finch (2015) showed that the GLR and the GMH had comparable power rates across magnitude of uniform DIF conditions. That was consistent with the findings of this study. However, Finch (2015) did not investigate nonuniform DIF items.

Precision Rate Difference

Precision rate was not investigated in any prior research so no direct comparison of results was possible with early DIF studies. However, when methods have similar power rates but different type I error rates, it is expected that the method with higher type I error rate will have lower precision rate than the method with lower type I error rate. In the similar way, when two methods had similar type I error rates, the method with higher power rate will have higher precision rate than the method with lower power rate. From this perspective, this study's findings with respect to precision rates could be compared with early studies. Güler and Penfield (2009); (1996) and Swaminathan and Rogers (1990) showed that the GLR was more powerful than the GMH for nonuniform DIF items and the GMH had inflated type I error rates in some situations. This supports the finding of this study with respect to precision rate. Overall, precision rate of the GLR was higher than the precision rate of the GMH for many situations. Especially for uniform DIF cases, both methods had similar power rates but different type I error rates that resulted different precision rates of two methods.

Limitation and Future Directions

Although the results of this study indicated that the GLR was better than the GMH with respect to precision rate, there are several limitations of these results that deserve recognition. This chapter discusses the limitations and the future directions of current study and its findings.

The first limitation of this study is that as with any simulation study, the generalizability of findings is limited by the selection of the manipulated factors in the design. The conditions were purposely aimed for PISA data. To eliminate or minimize the effect of impact, real means of ability estimates from twelve countries that participated PISA 2012 exam were used in the study. Each mean of ability estimates was used in normal distributions with a standard deviation of 1 and each examinee's ability was randomly chosen from the pool of these ability values.

The second limitation of this study is about the sample size. As Swaminathan and Rogers (1990) showed in their study, sample size affects the power of all procedures. Sample size was held constant as in this study to represent a reasonable sample for the cross-cultural studies.

However, in real- life data, as opposed to the simulation study, researchers rarely get equal sample sizes in their groups. This is especially true with minority groups, focal groups, which by definition are much smaller than the larger reference group. Hence different sample sizes should be studied in the future for simultaneously assessing DIF across multiple groups so that researcher may investigate groups sizes that are more accurately reflect real-life test data.

The third limitation of the study is that, in both the real study and in the simulation study, only dichotomously scored items were considered. For the real data analysis, polytomously scored items were excluded from the analysis. However, it is important to analyze the data as a whole not to distort the structure of it. Also, for the simulation study and for the ability estimation the 2 PL IRT model was used. However, in real life, there are items polytomously scored, and the guessing parameter (lower asymptote parameter) would be of interest especially for multiple-choice items that are dichotomously scored. The PISA had 11 polytomously scored items that were excluded from the real data analysis. In the future, any analysis with the PISA should consider all items in the study. Also for simulation studies, both dichotomously scored and polytomously scored items should be considered.

The interest of this study was only to detect DIF items in real analysis as an illustration. However, in practical applications once the DIF items are detected, further investigation should be done by experts to identify the sources of DIF. It is important to know why DIF occurred. Next, necessary steps should be taken; for example, item revision or item removal may be needed after the items are decided to be biased. Besides for large scale cross cultural studies like PISA and TIMMS experts specifically should be chosen from linguistic and cultural experts. Item revision and even item removal finalizes the whole DIF study that will make score comparability and any inferences regarding groups' performance more meaningful.

Another thing is in this study always the first five items had DIF for all conditions. However, it is quite likely that different groups may have different DIF items with respect to the reference group. In PISA real analysis, all 34 items displayed DIF that may be because of different groups' having different items with DIF or the type I error rate's being inflated. The scope of this study was on the hypothesis of type I error's being inflated when same DIF items existed. Hence, in future studies the condition of different groups having different DIF items should be considered as a manipulated factor. Also different level of total number of DIF items should be considered. Computing time required per analysis was another limitation to the simulation study. Although in real analysis, it is not an issue since only one dataset is considered, for the simulation study, 1000 replication was considered for each condition to ensure the accuracy of the finding.

Conclusion

The results of this study may be considered more comprehensive than other research done in the field of GMH's use in simultaneous DIF detection within a dichotomous IRT framework, because nonuniform DIF were examined in addition to uniform DIF. This study added to the existing research by examining the DIF methods' precision rates in the context of accurately detecting true DIF items among all items that are above the detection threshold. The study demonstrated an inflated type I error rate and relatively low power rates in general exhibited by both GMH and GLR for simultaneously detecting uniform DIF and nonuniform DIF in dichotomous items under different number of total groups (especially for 12 groups) and different groups experiencing DIF. Although the GLR performed better than the GMH in many situations, due to mixed results for the type I error, power and precision rates, there is no best method for detecting both uniform and nonuniform DIF for 2 PL IRT model. However, there are two guidelines than can help determining when a particular method should be used with its strength and weakness.

First, if the power is the main concern, then the GMH is the optimal option to investigate uniform DIF since the GMH overall had the best power rates. However, the GMH did have higher type I error rates and did not control type I error rates very well. Hence, when the type I error and precision were the main concerns, the GMH would not be the best option.

Second, if the precision is the main concern, then the GLR is the optimal option to investigate both uniform and nonuniform DIF. The GLR had better precision rate than the GMH in many conditions that indicated that the GLR is a better choice than the GMH for simultaneously assessing DIF for multiple groups when precision is the main concern. However, the GMH had still some advantages for uniform DIF items and even for nonuniform DIF items.

If the budget and time are the main concerns to a researcher with respect to further investigation of items such as item revision and item removal, then the method with lower type I error rate may have advantageous such as saving time without checking or even revising falsely detected nonDIF items. However, the potential danger is not to revise true DIF items that are not detected if the power of the method is low. For this reason, when assessing DIF simultaneously among multiple groups, using more than one method (even three methods if possible) may help researcher to get more accurate results.

To summarize, the purpose of the study was to investigate the performance of the GMH and the GLR to simultaneously assess DIF under several conditions. As a conclusion, this study proposes using both GMH and GLR in combination with other methods to assess DIF since each method has some advantages and disadvantages for different conditions. When it is strictly uniform DIF, the GMH performs slightly better than the GLR with respect to power and performs as well as the GLR when it is specifically crossing- nonuniform. When precision is the main concern, the GLR is the optimal option for assessing DIF simultaneously among multiple groups.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology.

Birnbaum, A. (1968). “Some Latent Trait Models and Their Use in Inferring an Examinee’s Ability.” In F Lord, M Novick (eds.), “Statistical Theories of Mental Test Scores,” Addison- Wesley, Reading, MA.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152.

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

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the items in a test*: Educational Testing Service.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*: ERIC.

De Ayala, R. J. (2009). The Theory and Practice of Item Response Theory. *New York: Guilford Publications Incorporated*.

DeVellis, R. F. (2006). Classical test theory. *Medical care, 44*(11), S50-S59.

Dorans, N. J., & Holland, P. W. (1992). DIF DETECTION AND DESCRIPTION: MANTEL-HAENSZEL AND STANDARDIZATION. *ETS Research Report Series, 1992*(1).

Dorans, N. J., & Holland, P. W. (1993). Dorans, N. J., & Holland, P. W. (1993). DIF detection and description:

Mantel–Haenszel and standardization. In P. W. Holland &

H. Wainer (Eds.), Differential item functioning (pp. 35-66). Hillsdale,

NJ: Erlbaum.

Edwards, J. M. (2016). *Assessment of the multivariate outlier approach for differential item functioning detection: A Monte Carlo simulation study.* (10108882 Ph.D.), Ball State University, Ann Arbor. Retrieved from https://search.proquest.com/docview/1793940408?accountid=14556

http://vv6tt6sy5c.search.serialssolutions.com/?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQuest+Dissertations+%26+Theses+Global&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&rft.genre=dissertations+%26+theses&rft.jtitle=&rft.atitle=&rft.au=Edwards%2C+Julianne+M.&rft.aulast=Edwards&rft.aufirst=Julianne&rft.date=2016-01-01&rft.volume=&rft.issue=&rft.spage=&rft.isbn=9781339722160&rft.btitle=&rft.title=Assessment+of+the+multivariate+outlier+approach+for+differential+item+functioning+detection%3A+A+Monte+Carlo+simulation+study&rft.issn=&rft_id=info:doi/ ProQuest Dissertations & Theses Global database.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, *77*(2), 177.

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

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(6), 543-553.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3-4), 199-215.

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 68*(6), 940-958.

Fidalgo, A. M., & Scalon, J. D. (2009). Using generalized Mantel-Haenszel statistics to assess DIF among multiple groups. *Journal of Psychoeducational Assessment*.

Finch, W. H. (2015). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied Measurement in Education, 29*(1), 30-45.

Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21*(2), 131-145.

Gierl, M., Jodoin, M., & Ackerman, T. (2000). *Performance of Mantel-Haenszel, SIBTEST, and Logistic Regression when the number of DIF items is large.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

Gómez-Benito, J., & Navas-Ara, M. J. (2000). A Comparison of χ2, RFA and IRT Based Procedures in the Detection of DIF. *Quality & Quantity, 34*(1), 17-31.

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

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on. *Educational Measurement: issues and practice, 12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*: Sage.

Hidalgo, M. D., & LÓPez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance

and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun

(Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Erlbaum.Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of cross-cultural psychology, 16*(2), 131-152.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations fidelity across languages. *Journal of cross-cultural psychology, 18*(2), 115-142.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement*, 209-225.

Kalaycioğlu, D. B., & Berberoğlu, G. (2011). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment, 29*(5), 467-478.

Kanjee, A. (2007). Using logistic regression to detect bias when multiple groups are tested. *South African Journal of Psychology, 37*(1), 47-61.

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

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European journal of psychology of education, 16*(3), 385.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*: Sage.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*: Routledge.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862.

Magis, D., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research, 46*(5), 733-755.

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

Mannocci, A. (2012). The Mantel-Haenszel procedure. 50 years of the statistical method for confounders control. *Italian Journal of Public Health, 6*(4).

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease.

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

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

Messick, S. (1990). Validity of test interpretation and use. *ETS Research Report Series, 1990*(1), 1487-1495.

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

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

Penfield, R. 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., & Camilli, G. (2007). Differential item functioning

and item bias. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics:

Vol. 26. Psychometrics (pp. 125-167). Amsterdam: Elsevier.

Poortinga, Y. H. (1989). Equivalence of Cross-Cultural data: an overview of basic issues. *International Journal of Psychology, 24*(6), 737-756.

Poortinga, Y. H., & Van de Vijver, F. J. (1987). Explaining cross-cultural differences bias analysis and beyond. *Journal of cross-cultural psychology, 18*(3), 259-282.

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

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*: Routledge.

R Development Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/

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

Sari, H. I., & Huggins, A. C. (2015). Differential Item Functioning Detection Across Two Methods of Defining Group Comparisons Pairwise and Composite Group Comparisons. *Educational and Psychological Measurement, 75*(4), 648-676.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*(3), 143-152.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1992). Evaluating hypotheses about differential item functioning. *ETS Research Report Series*, *1992*(1).

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

Shirley, T. A. (2014). *Language Learners and the Impact of Reading Ability and Socioeconomic Inequality on Science and Math Achievement: Discriminate Function Analysis and Differential Item Functioning (Item Bias) Studies of the PISA Across 45 Countries*: UNIVERSITY OF CALIFORNIA, DAVIS.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology*, *75*(5), 1350.

Svetina, D., & Rutkowski, L. (2014). Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. *Large-scale Assessments in Education, 2*(1), 4.

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

Team, R. (2015). RStudio: integrated development for R. *RStudio, Inc., Boston, MA URL http://www. rstudio. com*.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines.

Van de Vijver, F. J., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European review of applied psychology*, *47*, 263-279.

Walstad, W. B., & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *The Journal of Economic Education, 28*(2), 155-171.

Welkenhuysen-Gybels, J. (2004). The performance of some observed and unobserved conditional invariance techniques for the detection of differential item functioning. *Quality and quantity*, *38*(6), 681-702.

Welkenhuysen-Gybels, J., & Billiet, J. (2002). A comparison of techniques for detecting cross-cultural inequivalence at the item level. *Quality & Quantity*, *36*(3), 197-218.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods: Institutionen för beteendevetenskapliga mätningar, Umeå.

Woods, C. M. (2009). Testing for differential item functioning with measures of partial association. *Applied Psychological Measurement*.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532-547.

Yates, F. (1934). Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society, 1*(2), 217-235.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*.

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).