Differential Item Functioning for Polytomous Response Items Using Hierarchical Generalized Linear Model

by

Meng Hua

B.S., Southwest University of China, 2008

M.S., State University of New York at Albany, 2010

Submitted to the Graduate Faculty of the

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Meng Hua

It was defended on

October 31, 2019

and approved by

Clement A. Stone, PhD, Professor, Department of Psychology in Education

Lan Yu, PhD, Associate Professor, Department of Medicine

Dissertation Advisor: Feifei Ye, PhD, Senior Scientist, RAND Corporation, Pittsburgh

Suzanne Lane, PhD, Professor, Department of Psychology in Education

Differential Item Functioning for Polytomous Response Items Using Hierarchical Generalized Linear Model

Meng Hua, PhD

University of Pittsburgh, 2019

Hierarchical generalized linear model (HGLM) as a differential item functioning (DIF) detection method is a relatively new approach and has several advantages; such as handling extreme response patterns like perfect or all-missed scores and adding covariates and levels to simultaneously identify the sources and consequences of DIF. Several studies examined the performance of using HGLM in DIF assessment for dichotomous items, but only a few exist for polytomous items. This study examined the DIF-free-then-DIF strategy to select DIF-free anchor items and the performance of HGLM in DIF assessment for polytomous items. This study extends the work of Williams and Beretevas (2006) by adopting the constant anchor item method as the model identification method for HGLM, and examining the performance of DIF evaluation with the presence of latent trait differences between the focal and reference group. In addition, the study extends the work of Chen, Chen, and Shih (2014) by exploring the performance of HGLM for polytomous response items with 3 response categories, and comparing the results to logistic regression and Generalized Mantel-Haensel (GMH) procedure.

In this study, the accuracy of using iterative HGLM with DIF-free-then-DIF strategy to select DIF-free items as anchor was examined first. Then, HGLM with 1-item anchor and 4-item anchor were fitted to the data, as well as the logistic regression and GMH. The Type I error and power rates were computed for all the 4 methods. The results showed that compared to dichotomous items, the accuracy rate of HGLM methods in selecting DIF-free item was generally

lower for polytomous items. The HGLM with 1-item and 4-item anchor methods showed decent control of Type I error rate, while the logistic regression and GMH showed considerably inflated Type I error. In terms of power, HGLM with 4-item anchor method outperformed the 1-item anchor method. The logistic regression behaved similarly to HGLM with 1-item anchor. The GMH was generally more powerful, especially under small sample size conditions. However, this may be a result of its inflated Type I error. Recommendations were made for applied researchers in selecting among HGLM, logistic regression, and GMH for DIF assessment of polytomous items.

Table of Contents

List of Tables

List of Figures

Preface

The road to a doctorial degree is long and hard, and I would never have made it without the help and support of my advisor and friend Dr. Feifei Ye. I cannot thank her enough. I would also like to thank Dr. Suzanne Lane, Dr. Clement Stone, and Dr. Lan Yu, for the guidance and insight they provided while on my dissertation committee. And Dr. Kevin Kim, whom I miss to this day.

I must also thank my dear parents, whose unconditional love and support make me who I am. And my cat Captain Blue. D. McMeowmers, although utterly unhelpful, but her company got me through all those long dark nights.

1.0 Introduction

1.1 Background

For the past few decades, measurement equivalence has been an increasing concern in psychological and health studies. If measurement bias is present, a measurement scale is no longer invariant across groups, which means the measures perform differently for different groups of participants, thereby threatening cultural fairness and the accurate estimation of treatment effects, and may lead to flawed public policies (McHorney & Fleishman, 2006). Measurement equivalence can be viewed from various perspectives (Borsboom, 2006); it is often examined under the item response theory (IRT) framework, which is essentially an examination of differential item functioning (Embretson & Reise, 2000).

Differential item functioning (DIF) refers to a situation in which an item functions differently in two groups of participants conditioned by the latent measured trait. The presence of DIF is an indication of measurement bias; over the years, it has rendered concerns from numerous researchers. DIF assessment has a long history in education testing and is well-developed for dichotomous response items, possibly due to the popularity of multiple choice items that are scored as correct or incorrect, but it is less so for polytomous response items, which are scored on multiple points.

DIF assessment, although originated in education testing, is now becoming popular in health studies (Teresi, 2006). McHorney and Fleishman (2006) argued that DIF assessment is fundamental to health-related studies; as modern society becomes more culturally diverse in its age, racial, and socioeconomic status composition, it is crucial that health-related outcome

instruments are culturally fair. In 2017, a search of PubMed using the term “differential item functioning” resulted in 1271 articles, while in 2010, Scott et al. (2010) conducted a similar search that resulted in 211 articles. Researchers have been using DIF to evaluate the performance of measurement scales across participants of different age, gender, race, language, country, socioeconomic status, education, employment status, health care settings, and other characteristics. DIF has been identified in many health-related areas, such as mental health status, physical functioning and functional ability, patient satisfaction, and quality of life (McHorney & Fleishman, 2006; Scott, et al, 2010). Polytomous items are common in psychological evaluations and health studies, as the instruments often employ a Likert-scale type of measure. Therefore, researchers have been increasingly interested in DIF assessment for polytomous items.

1.1.1 DIF assessment for polytomous items

DIF assessment for polytomous items, however, presents its unique challenges. Penfield and Lam (2000) discussed three issues pertaining to the extension of DIF assessment from dichotomous items to polytomous items. First, reliabilities are typically lower in polytomous items. This effect results from a combination of shorter scale length, inconsistency of the rater scores, and more dissimilar content domains, all of which are common in polytomous items. Lower reliability is often related to inaccuracies in the trait estimates, which leads to false identification of DIF items, known as the Type I error.

Second, DIF assessment requires a matching variable to match examinees from the focal and reference groups with equal levels of latent trait so they are comparable. Traditionally, this is done by using the total score or some function of the total score as the matching variable. There are two classes of DIF procedures: the observed score approach uses the observed score as the

matching variables, while the latent trait approach uses an estimate of latent trait, which is a function of the observed score (Potenza & Dorans, 1995). The matching variable should be a sufficient estimate of the trait; in other words, information in the latent trait variable should be captured by the matching variable. In addition, the matching variable should be a reliable estimate of the latent trait (Meredith & Millsap, 1992). A mismatch between the latent trait and observed scores can inflate Type I error with the presence of different group abilities or trait levels (DeMars, 2010). However, due to the typically shorter scale length, lower reliability, and potential multidimensionality, defining a matching variable is less straightforward for polytomous items (Zwick, Donoghue, & Grima 1993).

One possible solution is using an external criterion to match the groups of examinees (Zwick, et al., 1993), as the chosen external variable can have high reliability. However, the main problem with this approach is that the external matching variable is not necessarily highly correlated with the target test; in other words, the external matching variable and the target test may not be measuring the same construct. Another solution is to improve the performance of the matching variable. Zwick et al. (1993) suggested including the studied item in the matching variable. Purifying the matching variable has also been found to result in more accurate results in polytomous DIF assessment (Hildago-Montesinos & Gómez-Benito, 2003; Su & Wang, 2005). Another way is to use a matching variable based on the estimated latent trait instead of the observed score (DeMars, 2008).

Third, creating a measure of item performance is more complex for polytomous items. For dichotomous items, item performance can be assessed by estimating the probability of a correct response. However, for polytomous items with multiple response categories, there is no single

measure, but rather, several degrees of correct response; in addition, there is a potential group difference in each category of response.

There are several solutions to this problem. One approach is to place the polytomous responses on an interval scale and compare the group mean score at each level of a matching variable, as adopted by the Mantel test (Mantel, 1963), and standardized mean difference (SMD) statistic (Dorans & Schmitt, 1991). The major problem with this approach lies in the appropriateness of treating ordinal responses on an interval scale. Sometimes the score categories may be nominal in nature, meaning the adjacent categories do not necessarily represent ordered levels of performance, making this approach more problematic. Another approach is to test for the group-by-score dependence at each level of matching variable, thus preserving the categorical nature of the rating scales. This is the method adopted by the generalized Mantel-Haenszel (GMH) approach (Somes, 1986). A third approach is to dichotomize the polytomous scale using various strategies and to assess the group difference in odds of a certain response as in the dichotomous scale. This is the logic used by the polytomous logistic regression procedure (PLR) (Agresti, 2013; French & Miller, 1996). The Mantel test, SMD and GMH do not specify a parametric form to match the item score at each trait level; this is known as the nonparametric method. As the logistic type of procedures do specify a parametric form to match item score at each trait level using a mathematical function, they are known as the parametric method.

1.1.2 DIF assess under hierarchical generalized linear model framework

One of the relatively new methods for DIF assessment is to use the hierarchical generalized linear model, which has received increasing attention. The hierarchical generalized linear model (HGLM), also known as the generalized linear mixed model, is a general form for nested data that

models nonlinear relationships. The popular hierarchical linear model (HLM) is a special case of HGLM in which the sampling data is normal, the link function is canonical, and the structure model is linear (Raudenbush & Bryk, 2002).

The relationship between IRT and HGLM has long been demonstrated by various researchers (Adams, Wilson, & Wu, 1997; Kamata, 2001). Kamata (2001) showed that the Rasch model is mathematically equivalent to a 2-level HGLM with fixed item parameters and random person parameters. Items are treated as repeated measures nested within participants. Willams and Beretvas (2006) expanded Kamata's work and demonstrated the equivalence of polytomous HGLM and a constrained form of Muraki's rating scale model. Since then, researchers have examined the HGLM for accounting for item dependence (Beretvas & Walker, 2012; Fukuhara & Paek, 2015; Paek & Fukuhara, 2015; Xie, 2014), and to account for both person and item dependence (Jiao, Kamata, Wang, & Jin, 2012; Jiao & Zhang, 2015).

HGLM has several advantages over the traditional IRT approach for DIF evaluation. First, in the HGLM framework, DIF can be interpreted as the difference between item parameter estimates in the focal group and the reference group, specified as the cross-level interactions between group indicators and item parameters (Chen, Chen, Shih, 2014); thus, the model allows assessment of multiple sources of DIF by examining the variability of DIF across items (Beretvas, Cawthon, Lockhart & Kaye, 2012; Van den Noortgate & De Boeck, 2005). Furthermore, additional covariates can be incorporated into the model to provide alternate explanations for DIF, rather than the descriptive measurement approach that the traditional IRT model takes, which focuses on the performance of the scale at measuring the participant' trait level (De Boeck & Wilson, 2004; Swanson, Clauser, Case, Nungester & Featherman, 2002). Thus, the model is more general, flexible, and conceptually useful. Second, the extreme response patterns of perfect scores

and all-missed scores can both be used for parameter estimation. Third, examination of the variability in person-level scale scores allows researchers to explore the consequences of DIF (Cheong & Kamata, 2013). Fourth, additional levels can be added to the model to account for higher-level clusters, such as doctor or hospital, while studies have shown that ignoring such nested structures can yield consequences such as inflated Type I error rate (French & Finch, 2010). Last, implementation of HGLM is more straightforward with widely available software. In addition, HGLM and its extensions can simultaneously handle item and person parameters, DIF, effect of covariates, as well as local item and person dependence. Thus, it has great potential for practical use (Ravand, 2015).

1.2 Statement of the Problem

One of the common issues for DIF detection is scale indeterminacy. Scale indeterminacy refers to the estimation of a DIF parameter that is not absolute but related to the other DIF parameters in the same scale (de Ayala, 2009). In order to solve this problem, it is necessary to set constraints to identify the model. Most studies that address this issue are in education testing settings. There are three popular approaches: the mean of the person ability parameter or the mean of the item difficulty parameter can be constrained to an arbitrary value (e.g., zero), or a set of anchor items can be selected to serve as a matching criterion variable (Chen et, al, 2014; Wang, 2004). Some studies examined the effect of different constraining methods using HGLM on DIF detection for dichotomous items. Cheong and Kamata (2013) explored the performance of the equal mean difficulty method and the constant anchor item method and compared the results to the well-researched Mantel-Haenszel procedure. Chen et al. (2014) explored the performance of

the equal mean ability method with rank-based strategy (Woods, 2009) and the constant anchor item method with the DIF-free-then-DIF strategy (DFTD; Wang, Shih, & Sun, 2012) for two criteria: first the accuracy in selecting DIF-free items, and then for Type I error rate and power. They found that the equal mean ability method is sensitive to group difference in ability (usually referred to as "impact") and is prone to Type I error under such conditions. As a result, it is not recommended if researchers suspect impact might be present (Chen et al., 2014). The equal mean difficulty method is more robust than the constant anchor item method when there is a violation of assumptions. Thus, it is recommended by Cheong and Kamata (2013). If the constant anchor item method is to be used, it is important that procedures be completed to make sure the reference items selected are free of DIF. However, these studies focused only on dichotomous items, not on polytomous items.

With polytomous items, literature examining the performance of HGLM in DIF assessment is relatively scarce. As previously mentioned, Willams and Beretvas (2006) extended Kamata's (2011) dichotomous HGLM to polytomous items and demonstrated the mathematical equivalence between Muraki's rating scale model (Muraki, 1990) and polytomous HGLM. The authors compared the performance of HGLM and IRT models for parameter recovery and found the two performed similarly. A comparison between HGLM and the generalized Mantel-Haenszel (GMH) approach for DIF detection under the condition of no group ability difference showed the two approaches produced similar results in terms of Type I error rate and statistical power. Ryan (2008) extended this study and found similar results. However, these studies used equal mean person ability method to constrain the model; in addition, the ability of groups of examinees was set to be equal, meaning no impact among groups. Yet impact is most likely present in reality, and plays an

important role in DIF assessment. It is necessary to extend these studies by exploring the performance of HGLM with different constraining methods, as well as with the presence of impact.

1.3 Purpose of the Study

The purpose of this study is to evaluate the performance of HGLM in DIF assessment for polytomous items, and in comparison to the GMH and polytomous logistic regression procedures. Specifically, this study expanded the work of Chen et al. (2014) by applying HGLM with DFTD strategy to polytomous items, using the constant anchor item method. Additionally, this study expanded the work of Williams and Beretvas (2006) by exploring the performance of DIF with the presence of impact.

1.4 Research Questions

This study attempted to answer the following three questions:

1. How accurately can HGLM select DIF-free items as anchor items for DIF analysis?
2. What is the Type I error rate for DIF detection using HGLM, and how does it compare to using GMH and logistic regression?
3. What is the statistical power for DIF detection using HGLM, and how does it compare to using GMH and logistic regression?

1.5 Significance of the Study

DIF assessment, as a tool to evaluate item fairness, is an integrated part of health studies. As the instruments in health studies commonly employ a Likert-type of scale, researchers have been paying more attention to DIF assessment for polytomous items, which is less developed and studied than DIF assessment for dichotomous items. DIF assessment with HGLM is a relatively new approach; in previous studies it has been proved useful for dichotomous items and showed great potential for polytomous items. However, the performance of HGLM in DIF assessment for polytomous items is not yet fully understood. This dissertation study aims to provide more information on this subject and produce useful guidelines for practitioners.

1.6 Organization of the Study

The rest of this dissertation is organized as follows: the second chapter reviews DIF assessment for polytomous items and various detection methods followed by an introduction of using HGLM and its application for DIF detection. The third chapter describes the Monte Carlo study in detail; simulation factors, evaluation criteria, data generation, validation, and analysis are discussed. The fourth chapter reports the results, and the fifth chapter summarizes and discusses the findings.

2.0 Literature Review

This chapter consists of several sections reviewing literature on DIF assessment on polytomous response items under the HGLM framework. First, DIF assessment for polytomous response items was reviewed; then, DIF detection using HGLM were discussed.

2.1 DIF Assessment Methods for Polytomous Items

To formulate DIF for polytomous items, assume $y_1, y_2, \ldots, y_t$ as the T scores of a certain item, where T is the number of possible response category scores. Reference and focal groups are noted as F and R. K is the number of levels of stratification variable. Table 1 presented a $2\times T$ contingency table for the kth stratum, with the row and column marginal totals fixed.

Table 1 Data for the kth Level of a 2×T Contingency Table Tsed in DIF Tetection

Group				Item Score			Total
	y_1	y_2	…	y_t	…	y_T	
Reference	n_{RIk}	n_{R2k}		n_{Rtk}		n_{RTk}	n_{R+k}
Focal	n_{FIk}	n_{F2k}		n_{Ftk}		n_{FTk}	n_{F+k}
Total	n_{+Ik}	n_{+2k}		n_{+2k}		n_{+2k}	n_{++k}

2.1.1 Nonparametric methods

Methods such as the Mantel, SMD and GMH approaches do not assume a particular statistical model to link item scores to the matching variable; instead, they just focus on the group difference in the observed item scores at each level of the matching variable; thus they are known as nonparametric methods (Penfield & Lam, 2000).

2.1.1.1 The Mantel test

The Mantel test is a polytomous extension of the Mantel-Haenszel (MH) procedure, which is one of the most widely-used DIF detection methods. MH for dichotomous items utilizes a series of *K* 2×2 contingency tables for each scoring level after examinees in both groups are matched on the total scores; where *K* is the number of the levels of the matching variable (Mantel & Haenzel, 1959). The Mantel test extended the MH by using a series of K 2×*T* tables with 2 rows and *T* columns, where *T* is the number of possible response category scores (Mantel, 1963). The test statistic is created by calculating weighted sum scores for the focal group and then summed for each response, conditioning on the total test score. Each table is created for each stratum of ability level. The null hypothesis is that the odds of correct response are the same in both the focal and the reference groups.

The weighted sum of scores for the focal group in the *k*th table for the *k*th stratum is

$$F_k = \sum_{t=1}^{T} y_t n_{Ftk} \quad (1)$$

where y_t is the item score for the *T* possible score on the item. n_{Ftk} is the number of focal group members with score *t* in the *k*th stratum. Under the hypothesis of no association, the rows and columns are considered independent; thus the rows and columns of frequencies for each group are

distributed as multivariate hypergeometric variables; that is, n_{Fk} is a multivariate hypergeometric variable with parameters n_{F+k} while n_{+k} with parameters n_{++k}. Thereby the expected value of F_k is

$$\mathrm{E}(\mathrm{F_k}) = \frac{n_{F+k}}{n_{++k}} \sum_{t=1}^{T} y_t n_{Ftk} \tag{2}$$

where a plus sign (+) indicates marginal sums meaning summation over the index; for example, n_{F+k} represents summation over all the numbers of focal group members in the *k*th score level.

The variance of F_k is

$$\mathrm{V}(\mathrm{F_k}) = \frac{n_{R+k} n_{F+k}}{n_{++k}^2 (n_{++k}-1)} [(n_{++k} \sum_{t=1}^{T} y_t^2 n_{+tk}) - (\sum_{t=1}^{T} y_t n_{+tk})^2] \tag{3}$$

Under the hypothesis of no association, the frequency counts can be viewed as following a multivariate hypergeometric distribution, and a chi-square test can be conducted to test the hypothesis, where the test statistic is distributed as a chi-square variable with 1 degree of freedom. To test the null hypothesis, a chi-square statistic is

$$\chi^2 = \frac{[\sum_{k=1}^{K} F_k - \sum_{k=1}^{K} E(F_k)]^2}{\sum_{k=1}^{K} V(F_k)} \tag{4}$$

with 1 degree of freedom for the χ^2 statistic.

A rejection of the null hypothesis provides evidence that even after the focal and reference group members are matched on the stratification variable of trait measures, there is still a group difference in the responses, indicating the presence of DIF. The Mantel test is easy to compute; however, it is designed to detect uniform DIF. Uniform DIF refers to a DIF when there is no interaction effect between group membership and item performances. In other words, the group difference in the measured property is constant among trait levels (Mellenbergh, 1982). The Mantel test is not defined for DIF with interactions between group membership and item performances (i.e., nonuniform DIF). A measure of overall DIF can be developed using odds ratios as described by Zwick et al. (1993) and Liu & Agresti (1996).

2.1.1.2 Standardized mean difference (SMD) statistic

SMD was originally proposed to condense information into a single value for dichotomous items (Dorans & Kulick, 1983, 1986). The null hypothesis is at each level of the matching variable, there is no group difference in proportion of the correct response, which is equivalent to the null hypothesis used by MH statistics (Dorans & Holland, 1993; Potenza & Dorans, 1995). Weighted difference in expected item scores are summed over levels of the matching variables to form DIF statistics.

Zwick & Thayer (1996) extended SMD to polytomous items using the standardized expected item score; the mean item score for each stratum is weighted by the proportion of focal or reference group members at the stratum. The authors presented two different types of standard error; the hypergeometric version was recommended because of the superior performance over the independently distributed multinomial version in terms of standard error ratios. Thereby the DIF statistics can be tested on a standard normal variable. SMD is closely related to the Mantel; both focus on expected test scores at each level of the matching variable.

Using the notations of Table 1, the test statistic for SMD is expressed as

$$\text{SMD} = [\sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}} \frac{\sum_{t=1}^{T} y_t n_{Ftk}}{n_{F+k}}] - [\sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}} \frac{\sum_{t=1}^{T} y_t n_{Rtk}}{n_{F+k}}] \quad (5)$$

under the hypergeometric framework of Mantel (1963), and under the null hypothesis Var (F_k) = Var(R_k); thus the covariance between F_k and R_k is expressed as

$$\text{Cov}(F_k, R_k) = \text{Cov}\left(\sum_{t=1}^{T} y_t n_{Ftk}, \sum_{t=1}^{T} y_t (n_{+tk} - n_{Ftk})\right) \quad (6)$$

$$= -\sum_{t=1}^{T} y_t^2 Var(n_{Ftk}) = -\text{Var}(F_k)$$

Thus there is

$$\text{Var(SMD)} = \sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}}^2 \left[\left(\frac{1}{n_{F+k}}\right)^2 Var(F_k) + \left(\frac{1}{n_{R+k}}\right)^2 Var(R_k) - 2\left(\frac{1}{n_{F+k}}\right)\left(\frac{1}{n_{R+k}}\right) Cov(F_k, R_k)\right]$$

$$= \sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}}^2 \left(\frac{1}{n_{F+k}} + \frac{1}{n_{R+k}}\right)^2 Var(F_k) \quad (7)$$

Using the variance formula, it is possible to test the SMD statistic on a standard normal distribution. A positive SMD indicates that the item favors the focal groups, while a negative SMD indicates that the items favors the reference group, after conditioned on the matching variable. In addition, SMD can also be used as a descriptive statistic to measure the size of DIF (Zwick & Thayer, 1996).

2.1.1.3 Generalized Mantel-Haenszel (GMH)

GMH is an alternate generalization to the MH procedure. GMH is computed by calculating the proportion of group members for each response category, at each level of the matching variable. Under the null hypothesis of no conditional association between item response and group membership, the test statistic is asymptotically distributed as a chi-square variable with *T*-1 degrees of freedom. Unlike the Mantel test which treats the response categories on an ordinal scale, GMH treats response categories on a nominal scale; thus the order of the response is irrelevant. The test statistic for GMH is multivariate normal, while for Mantel it is univariate for the weighted linear combination of item scores that formed the average score (Potenza & Dorans, 1995). In addition, GMH utilizes the entire item response scale to detect nonspecific different patterns across distribution when comparing the performance of focal and reference groups, while the Mantel test and SMD focus on mean item scores across the matching variable. Theoretically, this would

indicate that GMH is sensitive to both uniform and nonuniform DIF, even though it does not produce separate coefficients.

Using the notations of Table 1, there is

$$A'_k = (n_{R1k}, n_{R2k}, \ldots, n_{R(T-1)k}) \tag{8}$$

where A'_k is a 1× (*T*-1) vector consisting of the *T*-1 pivotal cells for the *k*th strata. Let

$$n'_k = (n_{+1k}, n_{+2k}, \ldots, n_{+(T-1)k}) \tag{9}$$

the expected value of A'_k is

$$\mathrm{E}(A'_k) = \frac{n_{R+k} n'_k}{n_{++k}} \tag{10}$$

The variance-covariance matrix of A'_k is

$$\mathrm{V}(A'_k) = n_{R+k} n_{F+k} \frac{n_{++k} diag(n_k) - n_k n'_k}{n^2_{++k}(n_{++k}-1)} \tag{11}$$

where dig(n_k) is a (*T*-1)×(*T*-1) diagonal matrix with elements A_k, The GMH statistic is expressed as

$$\chi^2_{GMH} = [\sum A_k - \sum E(A_k)]'[\sum V(A_k)]^{-1}[\sum A_k - \sum E(A_k)] \tag{12}$$

under the null hypothesis of no association between item response category and group membership conditioned on the matching variable; this statistic follows a chi-square distribution with *T*-1 degrees of freedom.

2.1.2 Parametric methods

Methods such as Polytomous logistic regression (PLR) and close-related logistic discriminant function analysis (LDFA) approaches do use statistical functions to link item scores to the matching variable, thus known as the parametric methods.

2.1.2.1 Polytomous logistic regression (PLR)

Swaminathan and Rogers (1990) proposed to use logistic regression for DIF detection for dichotomous items. Using this approach, the probability of an item response is estimated as a function of the group membership and the person ability using the observed score as a proxy. Both uniform and nonuniform DIF can be incorporated into the model: uniform DIF is specified as the group coefficient, while nonuniform DIF as interaction coefficient for group and item variables (Rogers & Swaminathan, 1993). The null hypothesis is a variation of the SMD definition as a mathematical function is specified to the empirical regression assumed by SMD (Potenza & Dorans, 1995). One approach to test for DIF is to test the significance of the group coefficient and item-by-group interaction coefficient using the Wald test. Another approach is to compare the models using the likelihood ratio test since the models are nested.

The dichotomous logistic model approach can be extended to polytomous items using various multinomial logistic regression (MLR) methods to form a logit construct, so the two response categories or the combination of response categories can be compared in a dichotomous manner (French & Miller, 1996; Miller & Spray, 1993). The most popular MLR methods are the cumulative model, the continuation ratio model, and the adjacent categories model (Agresti, 2013). For the cumulative model, cumulative probabilities of responses equal to or greater than a certain response category are compared to those smaller than the category. For the continuation ratio model, probability of a certain response category is compared to that of all the combined response categories beneath it. For the adjacent categories model, probability of a certain response category is compared to that of the category beneath it. For items with T response categories, there are T-1 models corresponding to one model in the dichotomous situation, where DIF can be evaluated in a similar manner.

The logistic model for dichotomous items can be reparametrized as

$$\text{logit (u)} = \beta_0 + \beta_1\theta \tag{13}$$

$$\text{logit (u)} = \beta_0 + \beta_1\theta + \beta_2 G$$

$$\text{logit (u)} = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3\theta G$$

where u is the response to the given item; θ is the observed ability of the participant represented by the total score, while G is the group membership; θG is the interaction between group and ability. $\beta_0, \beta_1, \beta_2, \beta_3$ are model coefficients for the intercept, ability, group effect and the ability by group interaction effect. The models are nested and thereby uniform DIF can be tested by conducting the likelihood ratio test of Equation 13(1) and Equation 13(2); a significant test indicates the presence of uniform DIF for the item. Likewise, a significant likelihood ratio test of Equation 13(2) and Equation 13(3) indicates the presence of nonuniform DIF for the item.

For the cumulative model,

$$\text{logit } (u_t) = \ln\left[\frac{p_1+\cdots p_t}{p_{t+1}+\cdots+p_T}\right] \tag{14}$$

where u_t is the response to the t_{th} response category in the given item; $p_1, \ldots p_T$ are the probabilities of response for each item category.

For the continuation ratio model, there is

$$\text{logit } (u_t) = \ln\left[\frac{p_t}{p_{t+1}+\cdots+p_T}\right] \tag{15}$$

for the adjacent categories model, there is

$$\text{logit } (u_t) = \ln\left[\frac{p_t}{p_{t+1}}\right] \tag{16}$$

For all three models, for an item with T categories, there are T-1 logistic functions expressed as:

$$\text{logit } (u_{t-1}) = \beta_0^{T-1}+\beta_1^{T-1}\theta + \beta_2^{T-1}G + \beta_3^{T-1}\theta G \tag{17}$$

One possible solution to reduce the number of regression functions is by constraining the slope coefficients across functions to be equal, while freely estimating the intercept coefficients. By assuming equal-slope regression lines across functions, Equation (17) is reduced to

$$\text{logit}\,(u_{t\text{-}1}) = \beta_0^{T-1} + \beta_1\theta + \beta_2 G + \beta_3\theta G \qquad (18)$$

The advantage of the polytomous logistic regression approach is that it has the ability to distinguish uniform and nonuniform DIF; additionally, the group difference in various combinations of response categories can be examined. However, there is no omnibus measure of DIF across all response categories. In addition, the sample size demand is usually large (Miller & Spray, 1993). Moreover, *T*-1 logistic functions produce a large amount of parameter estimates; therefore, the results can be difficult to interpret. Furthermore, some MLR methods contain underlying assumptions, such as equal-slope regression lines, which may not necessarily be met in practice (French & Miller, 1996).

2.1.2.2 Polytomous logistic discriminant function analysis (LDFA)

LDFA predicts the probabilities of group membership as a function of the matching variable (usually total score), item scores, and the interaction between total score and item score (Miller & Spray, 1993). LDFA is essentially a dichotomous logistic model, therefore it requires only one simple logistic function. Instead of predicting the probability of item responses given the total score and group membership, the LDFA function is reversed. The item score can take on continuous values instead of dichotomous ones. The coefficients can then be tested in a similar manner as in the dichotomous logistic regression procedures: uniform DIF can be examined by comparing the model predicting group membership from total score to the model predicting group membership from the total score and item score. Nonuniform DIF can be examined by comparing

the model with total score and item score as predicters of the model with the total score, the item score, and the interaction effect of total and item score as predictors. A detection of DIF indicates that the prediction of being in a certain group by total score is only different from that by total score and item score.

$$\text{Logit}(g) = \alpha + \alpha_1 X \tag{19}$$

$$\text{Logit}(g) = \alpha + \alpha_1 X + \alpha_2 U \tag{20}$$

$$\text{Logit}(g) = \alpha + \alpha_1 X + \alpha_2 U + \alpha_3 XU \tag{21}$$

where g is the group membership; X is the total score; U is the item score; α is the regression coefficient corresponding to β in logistic regression. Uniform DIF can be tested by conducting the likelihood ratio test of Equation (19) and Equation (20); a significant test indicates the presence of uniform DIF for the item. Likewise, a significant likelihood ratio test of Equation (20) and Equation (21) indicates the presence of nonuniform DIF for the item.

Advantages of LDFA are that it does not require multiple regression functions; it can produce an overall estimation of DIF across items and categories, as well as being able to distinguish between uniform and nonuniform DIF. However, LDFA is prone to false identification of DIF items when there is large ability difference between groups, and tends to lose power when the discrimination index is high.

2.1.2.3 IRT likelihood-ratio test (IRT-LR)

Instead of using the observed score as the matching variable, DIF methods based on IRT use the latent trait measures as the matching variable (Potenza & Dorans, 1995). Under the IRT framework, DIF exists when there are differences in the item response functions for the reference and focal groups with the same latent trait measures (Lord, 1980). One of the most popular and flexible IRT methods for DIF is the likelihood-ratio test (IRT-LR) (Thissen, Steinberg, & Gerrard,

1986; Thissen, Steinberg, & Wainer, 1988). For IRT-LR, DIF exists when there are differences in the probabilities of obtaining a certain score category for the reference and focal groups with the same latent trait measures (Bolt, 2002; Thissen et al., 1988). The general form of likelihood ratio test for DIF is a comparison between the log likelihood of the compact model (L_C) and the log likelihood of the augmented model (L_A). For the compact model, an item's parameters for the focal and reference groups are constrained to be equal, while for the augmented model such constraints are relaxed. The test statistics G^2 follows a chi-square distribution, with the null hypothesis of no DIF. The degree of freedom equals to the differences in numbers of parameter estimated for the two models.

$$G^2 = (-2\log L_C) - (-2\log L_A) \quad (22)$$

The IRT-LR can distinguish uniform and nonuniform DIF by estimating certain item parameters for the focal and reference groups during model comparisons. In the IRT frame, uniform DIF is a function of item difficulty parameter *b* while nonuniform DIF is a function of item discrimination parameter *a* (Camilli & Shepard, 1994). Thus IRT-LR can produce separate coefficients for statistical testing. However, the IRT-LR is computationally intensive, since for every testing one model must be fitted twice.

2.1.3 Comparisons of DIF detection methods

DIF assessment originated in education testing; thereby, most of the simulation studies on DIF assessment were conducted in the education setting. DIF assessment in health studies is a relatively new topic; little simulation studies exist to examine the behaviors of various methods and provide guidelines for practitioners. As a result, in this dissertation, the discussion of DIF detection methods were mostly conducted in the education setting.

The Mantel, GMH and SMD are all nonparametric methods. Nonparametric methods typically are simple to compute and involves less assumptions than parametric methods. However, Woods (2011) observed that the two MH statistics seem to have underlying assumptions regarding equal ability and equal item discrimination between groups. The Mantel test, while focusing on mean difference to match examinees from difference groups at each stratum of the matching variable, are more sensitive to group difference in ability. In addition, the Mantel is not designed to detect DIF that involves interactions between group membership and item responses, and thus not powerful to detect nonuniform DIF. This is also true for SMD, which is closely related to the Mantel. For well-behaved items with constant, uniform DIF, SMD and the Mantel are very powerful. However, when the ability between groups are unequal, and there are unparallel response functions between groups, the Mantel and SMD lose power and are more prone to Type I error. Under balanced DIF, the response functions for focal and reference groups are no longer parallel. Thereby the Mantel loses power, and GMH is recommended, for it compares group difference across the entire distribution of response categories. Furthermore, since GMH is designed to detect group difference in overall distribution patterns, it is more capable in detecting complex DIF patterns. It is more robust against presence of impact, and generally more powerful for balanced or nonuniform DIF (DeMars, 2008; Fidalgo and Bartram, 2010; Kristjansson et.al, 2005; Woods, 2011).

PLR and LDFA are parametric methods that rely on a mathematic model to make statistic inference. PLR detects DIF by predicting probability of certain item response as a function of total score and group membership, and then evaluating the group effect for DIF. However, this requires multiple functions to correspond to each item response category. Thereby the computation can become cumbersome and the results can be hard to interpret. LDFA avoids this problem by

reversing the logistic function and evaluate DIF by comparing the prediction of group membership as a function of total score and item score. Theoretically, both methods are capable of detecting uniform as well as nonuniform DIF. Studies of PLR are relatively scarce. Kristjansson et al. (2005) compared the Mantel, GMH, PLR and LDFA with the presence of small to moderate impact, three different magnitude of item discrimination parameter, and different skewness for ability distribution and found that all four methods performed well for uniform DIF. For nonuniform DIF, the power was poor for the Mantel and LDFA, while GMH and PLR performed very well. Performance of PLR under large impact is unclear. LDFA in general performs similarly to the Mantel (Kristjansson et al., 2005; Su and Wang, 2005). When the group ability is equal, LDFA can be more powerful than PLR (Hidalgo & Gómez, 2006).

All the aforementioned methods use the observed score as the matching variable. Another approach is to use an estimate of latent trait as the matching variable (Potenza & Dorans, 1995). The IRT-LR is a latent parametric method that operate using the IRT framework. It is flexible, informative, and powerful when the assumptions are met. Woods (2011) found IRT-LR to be more robust against the nonnormality of latent trait distribution than the Mantel and GMH methods. However, IRT-LR can be computationally intensive. In addition, as a parametric method relying on IRT model for statistical inferences, it is sensitive to model misfit (Bolt, 2002). Since during the model comparison process the IRT-LR relays on anchor items for calibration, the purity of anchor items are crucial (Cohen, Kim, & Baker, 1993; Kim & Cohen, 1998); when the anchor items are not DIF-free, IRT-LR can be prone to inflated Type I error (Elosua & Wells, 2013).

2.1.4 Factors considered in DIF detection studies

2.1.4.1 Examinee factors

Latent trait parameter distribution The mean trait difference between the focal and reference groups is often referred to as "impact". When a large amount of impact is present, the matching variable, which used the observed score to match examinees on their ability level, may not be a sufficient index of the latent proficiency; this mismatch between the observed test score and latent proficiency may cause inflated Type I error, especially for less reliable tests (DeMars, 2010). The Mantel, SMD, GMH and LDFA all showed a tendency to have inflated Type I error rates when impact is present, particularly in combination with a high percentage of DIF items, a shorter matching variable, and smaller magnitude of DIF (Chang, Mazzeo, & Roussos, 1996; DeMars, 2008; Su & Wang, 2005; Wang & Su, 2004; Woods, 2011; Zwick, Thayer, & Mazzeo, 1997). However, Kristjansson, Aylesworth, McDowell, & Zumbo (2005) found the group difference to be of little effect; the authors speculated that it might be because the size of impact is moderate in their simulation studies (mean difference = .5 on standard normal distribution), whereas it is large in other studies (≥1). They speculated that the influence of impact is only strong when the size of impact is large.

Few studies have examined the effect of nonnormality of ability distribution on polytomous DIF detection. Moyer (2013) found a decrease of accuracies in non-normal ability estimation even when test length and sample size increase, suggesting more difficulty involved in estimation with nonnormal ability distribution. Woods (2011) found that even though nonparametric methods, such as the Mantel and GMH, do not make explicit assumptions regarding the distribution of latent variables since the matching variable is matched on observed score, they showed decreased performance when the ability distribution was not normal, which suggests a mismatch between the

matching variable and the latent trait, and that the matching variable is no longer a sufficient proxy. This effect suggests a possible underlying assumption of equality about latent variable distributions between groups. Kristjansson et al. (2005) examined the skewness of ability distribution and found little effect on the performance of DIF detection, again possibly due to the moderate impact size.

Sample size A larger sample size is usually related to higher power (proportion of items with DIF that were detected or true positive); however, in some conditions it may inflate Type I error. This has been observed for the Mantel, GMH and SMD (Chang, et. al, 1996; DeMars, 2008; Woods, 2011) as well as for PLR and LDFA (Hildago & Gómez, 2006; Hidalgo, López-Martínez, & Gómez-Benito, Guilera, 2016; Hildago-Montesinos & Gómez-Benito, 2003), especially in combination with a large impact, high percentage of DIF items, and shorter test.

Wood (2011) examined the Mantel and other nonparametric methods and found that under small sample condition (R40/F40 and R400/F40), power is generally too low for practical use (< 60%). Ryan (2008) found similar results for GMH with the only exception when the magnitude of DIF is really large (.75). It could be concluded that a sample size smaller than 500, especially when the number of examinees for the focal and reference groups is not equal, may be too small for achieving sufficient power.

The parametric methods have larger sample size requirements than the nonparametric methods. For PLR, to acquire a decent power rate, a total sample size of 2000 is necessary (French & Miller, 1996) A sample size smaller than 1000 typically does not produce sufficient power (Elosua & Wells, 2013; Hildago & Gómez, 2006). LDFA seemed to have a similar sample size requirement as PLR, although it seemed to perform slightly better than PLR in smaller sample size conditions. (Hildago & Gómez, 2006; Hildago-Montesinos & Gómez-Benito, 2003).

Sample size ratio The effect of unequal number of examinees for the focal and reference groups is not consistent across studies. Some studies found GMH showed a decrease of power when the ratio between reference and focal group change from 1:1 to 4:1, especially for uniform DIF (Kristjansson et. al, 2005; Ryan, 2008). This result is not surprising; an unequal sample size means smaller subjects in the focal group, and consequentially fewer data at each level of the matching variable, resulting in less reliable matching. However, Wood (2011) found a ratio of 10:1 under small sample condition has larger power for the Mantel.

2.1.4.2 Test factors

Percentage of DIF items Some studies have shown that a high percentage of DIF is related to inflated Type I error. Some researchers argue that it is not the percentage of DIF items but the magnitude of overall DIF for the test, which is a function of the percentage of DIF items and the DIF patterns, that is causing the inflation (Su & Wang, 2005; Wang & Su, 2004; Wang & Yeh, 2003). Woods (2011) found that the Mantel and GMH are more sensitive to the percentage of DIF when the test is short. For PLR and LDFA, an increase in the percentage of DIF items under the condition of nonuniform DIF can result in an increase of both Type I error and power (Hildago & Gómez, 2006; Hildago-Montesinos & Gómez-Benito, 2003).

Test length Studies have shown that with the presence of a large impact, increasing the length of the test can help control Type I error rates (DeMar, 2008; Hidalgo, López-Martínez, & Gómez-Benito, & Guilera, 2016; Wang & Su, 2004; Woods, 2011). DeMars (2008) found that for a 5-item test, Type I error rates were inflated when a large impact was present. Hidalgo et. al (2016) found that for short tests (4 - to 10 - items), LDFA showed inflated type I error especially combined with larger sample size and higher percentage of DIF items. Wang and Su (2004) found that for a 10-item test, the Mantel and GMH could not control Type I error well with the presence

of a large impact. Woods (2011) found that Type I error rates for both methods are acceptable only when there are at least 12 items in the matching variable. For a longer test, test length had little effect (Fidalgo & Bartram, 2010; Su & Wang, 2005; Wang & Su, 2004; Woods, 2011). It seemed that a test with less than 10 items may be too short; observed score based on a short test is less reliable and more likely not a sufficient proxy for the latent ability. This mismatch is more serious when large ability difference between groups is present. For a test with more than 20 items, test length had an insignificant impact on the performance of DIF assessment.

2.1.4.3 DIF factors

DIF pattern For dichotomous items, when DIF items are in favor or against one group constantly across items, it is known as constant pattern. When some DIF items favor the focal group and others favor the reference group, the magnitude of DIF are balanced across items, known as the balanced DIF. For polytomous items, DIF can take on more complex patterns since the patterns can also be exhibited in response categories, resulting in "within-item" patterns as well as "between-item" patterns (Wang & Su, 2004). For an item with unbalanced DIF within categories, it is possible for DIF to only exist in the lower categories or only in the higher categories.

Studies have shown that constant DIF usually has more power, while balanced DIF is harder to detect. When DIF is only present in the highest or lowest response category, power can be very poor, mostly below 50% (Fidalgo & Bartram, 2010; Su & Wang, 2005). Typically, the Mantel is more powerful for constant DIF (Su & Wang, 2005), while GMH performs better for DIF that is not constant (Fidalgo & Bartram, 2010; Woods, 2011; Zwick et al., 1993).

Uniform and nonuniform DIF Uniform DIF refers to the situation when the probability of answering an item does not change at different levels of trait levels for different groups; in other words, there is no interaction between item responses and group membership. For nonuniform

DIF, such interactions exist. Many nonparametric methods are designed to detect uniform DIF. However, GMH is constructed to utilize the entire item response scale to detect DIF with non-specific patterns, hence theoretically it should be sensitive to nonuniform DIF as well, even though it does not produce separate test coefficients. For parametric methods like the logistic regression, uniform and nonuniform DIF can be tested with separate coefficients.

In IRT terms, uniform DIF is a function of item location while nonunifrom DIF is a function of item discrimination for dichotomous items. However, for polytomous items, nonuniform DIF is not necessarily only a function of item discrimination parameters; for example, Su and Wang (2005) had shown that nonuniform DIF can occur for a balanced DIF pattern without the interfering of item discrimination parameter. Thus, the terms “uniform DIF” and “nonuniform DIF” are not necessarily accurate. Many researchers seemed to use the term “uniform DIF” as analogous to “parallel DIF” defined by Hanson (1998) and use the term “nonuniform DIF” when response functions are not parallel between groups.

Some studies found that the presence of high item discrimination was related to inflated Type I error (Chang et.al , 1996; Su & Wang, 2005; Wang & Su, 2004; Woods, 2011; Zwick, et. al, 1997) while some found an insignificant effect on Type I error (Elosua & Wells, 2013; Fidalgo & Bartram, 2010; Hidalgo & Gomex, 2006; Kristjansson et al., 2005). On the other hand, unequal item discrimination parameters with high variation leads to a significant decrease in power (Fidalgo & Bartram, 2010; Kristjansson et. al, 2005; Woods, 2011).

Magnitude of DIF Increasing the magnitude of DIF is typically related to the increase of power; DIF of a small size can be hard to detect (Ryan, 2008; Su & Wang, 2005; Wang & Su, 2004; Zwick et. al., 1993). When the magnitude of DIF is small (.1), power is very poor; for constant DIF, increase the magnitude of DIF can increase the power rate.

For PLR and LDFA, increasing the magnitude of DIF is typically related to the increase in power (Hildago-Montesinos & Gómez-Benito, 2003). The effect is more prominent when combined with a large sample size (Hildago & Gómez, 2006); however, there is also a slight increase of Type I error rate. Elosua & Wells (2013) also found that PLR showed inflated Type I error when the magnitude of DIF is large.

2.2 DIF Assessment Using HGLM

The relationship between HGLM and IRT has long been recognized; however, using HGLM as a DIF assessment method has not drawn much attention until recently. HGLM is a general and flexible approach to DIF assessment, while DIF can be directly manipulated as model parameters. A two-level HGLM with fixed item effect and random person effect is equivalent to the Rasch model, while DIF can be specified as group-by-item interaction terms. Multiple sources of DIF, as well as the consequence of DIF, can be examined simultaneously.

2.2.1 A HGLM framework for DIF

Kamata (2001) verified the mathematical equivalence of IRT and HGLM for dichotomous data in the education testing framework. Under the Rasch model, the probability of a correct response of person *j* to item *i* is a function of the person ability θ_j and the item difficulty parameter b_i:

$$\text{Logit}\,(p_{ij}) = \theta_j - b_i \quad \text{with } \theta_j \sim N(0, \tau) \tag{23}$$

The Rasch model can be re-parameterized into a two-level logistic model. The level-1 model for the probability of correct response p_{ij} of student j (j = 1,…, J) on item i (i = 1, …, I) for a test with I items is

$$\text{Logit}\,(p_{ij}) = \beta_{0j} + \sum_{i=1}^{I-1} \beta_{ij} W_{kij} \tag{24}$$

where W_{kij} is the ith dummy-coded item indicator for student j with W_{kij} = 1 if k = i, otherwise W_{kij} = 0. The level 2 equation at person level with random intercept for β_{0j} across persons is expressed as

$$\beta_{0j} = \gamma_{00} + u_{oj}, \text{ with } u_{0j} \sim N(0, \tau) \tag{25}$$

$$\beta_{ij} = \gamma_{i0} \text{ for i = 1, …, I-1}$$

where u_{0j} is a random component represents the ability for person j. Combine the level-1 and level-2 models and the log-odds of the probability of a correct response to item i for person j is

$$\text{Logit}\,(p_{ij}) = u_{0j} + \gamma_{00} + \gamma_{i0} = u_{0j} - (-\gamma_{00} - \gamma_{i0}) \tag{26}$$

where u_{0j} is the random person effect; $-\gamma_{00} - \gamma_{i0}$ is the fixed item effect.

This two-level logistic model can be extended to assess DIF by modeling the main effect of the group membership as a function of additional dummy-coded covariates for the groups. The level-2 model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01} G_j + u_{oj}, \text{ with } u_{0j} \sim N(0, \tau) \tag{27}$$

$$\beta_{ij} = \gamma_{i0} + \gamma_{i1} G_j \text{ for i = 1, …, I-1}$$

where G_j is the dummy group membership indicator (for example, gender). γ_{01} represents the common effect of being in the focal group compared to the reference group, and γ_{i0} represents the mean item effect. A significant γ_{i1} indicates a significant uniform DIF exists for item i between the two levels of group indicator variable G. γ_{i1} can be tested by performing a t-test or a likelihood ratio test. The combined model is

$$\text{Logit}\,(p_{ij}) = \gamma_{00} + \gamma_{01}G_j + \sum_{i=1}^{I-1}\gamma_{i0}W_{kij} + \sum_{i=1}^{I-1}\gamma_{i1}G_jW_{kij} + u_{0j} \qquad (28)$$

Williams and Beratvas (2006) extended the Kamata (2001) model to polytomous items and demonstrated the mathematical equivalence polytomous HGLM and a constrained form of Muraki's rating scale model (Muraki, 1990).

Muraki's rating scale model is closely related to Samejima's (1969) grade response model. In graded response model, for each item with T response categories, there are a discrimination parameter a_i and $T - 1$ category boundaries b_{it}. The probability of scoring x and above for item i is

$$P_{i1}(\theta) = 1 - \frac{\exp[a_i(\theta-1)]}{1+\exp[a_i(\theta-b_{i1})]},\ t = 1 \qquad (29)$$

$$P_{it}(\theta) = \frac{\exp[a_i(\theta-b_{i(t-1)})]}{1+\exp[a_i(\theta-b_{i(t-1)})]} - \frac{\exp[a_i(\theta-b_{it})]}{1+\exp[a_i(\theta-b_{it})]},\ 1 < t < T$$

$$P_{iT}(\theta) = \frac{\exp[a_i(\theta-b_{it})]}{1+\exp[a_i(\theta-bit)]},\ t = T$$

Muraki's rating scale model is a special case of the graded response model with equal category thresholds for each category across items.

$$P_{it}(\theta) = \frac{\exp[a_i(\theta-b_i+c_t)]}{1+\exp[a_i(\theta-b_i+c_t)]} - \frac{\exp[a_i(\theta-b_i+c_{t-1})]}{1+\exp[a_i(\theta-b_i+c_{t-1})]} \qquad (30)$$

where t is the category score, b_i is the location parameter for item i. c_t is the category threshold parameter and is constant across items for each t.

With the item discrimination parameters set to 1, the constrained form of Muraki's rating scale model is expressed as

$$\text{Logit}\,[p_{ij}(X_i \geq t)] = \theta_j - b_{it} \qquad (31)$$

where b_{it} is the category difficulty parameter for category t with $b_{it} = b_i - c_t$, where b_i is the location parameter for item i and c_t is the category threshold for category t. For an item with three response categories, for the first category boundary, the probability of a response in category 1 over the probability of a response in a category higher than category 1 is

$$\frac{Pr_{ij}(X_i=1)}{Pr_{ij}(X_i=2,3)} = \exp(b_{i2} - \theta_j) = \exp(b_i - c_2 - \theta_j) \quad (32)$$

where b_{i2} is the first category boundary value, c_2 is the category threshold for the second category. For the second category boundary, the probability of a response in category 1 or 2 over the probability of a response in category 3 is

$$\frac{Pr_{ij}(X_I=1,2)}{Pr_{ij}(X_i=3)} = \exp(b_{i3} - \theta_j) = \exp(b_i - c_3 - \theta_j) \quad (33)$$

The constrained Muraki's Rating scale model can be re-parameteriazed as polytomous HGLM. The level 1 model for the probability of a response p_{ij} of student j on item i with three response categories is

$$\text{Logit}(p_{1ij}) = \beta_{0j} + \sum_{i=1}^{I-1} \beta_{ij} W_{kij} \quad (34)$$

$$\text{Logit}(p_{2ij}) = \beta_{0j} + \sum_{i=1}^{I-1} \beta_{ij} W_{kij} + \delta_j$$

where p_{1ij} is the probability of person j responding in category 1; p_{2ij} is the probability of person j responding in category 1 or 2; δ_j is the fixed threshold difference between response categories. The level 2 equation at person level with random intercept for β_{0j} is expressed as

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (35)$$

$$\beta_{ij} = \gamma_{i0} \text{ for i= 1, …, I-1}$$

$$\delta_j = \delta$$

Combining the level 1 and level 2 models and the log odds of the probability of a response in category 1 over the probability of a response in a category higher than category 1 is

$$\frac{Pr_{ij}(X_i=1)}{Pr_{ij}(X_i=2,3)} = \exp(\gamma_{00} - \gamma_{i0} + u_{0j}) \quad (36)$$

$$\frac{Pr_{ij}(X_I=1,2)}{Pr_{ij}(X_i=3)} = \exp(\gamma_{00} - \gamma_{i0} + \delta + u_{0j}) \quad (37)$$

It can be seen that the Equation (35) is equivalent to Equation (31), where the difficulty of responding to Category 2 or 3 is $(b_i - c_2)$ for item i, which corresponds to $(\gamma_{00} - \gamma_{i0})$. Equation (36)

is equivalent to Equation (32), where the difficulty of responding to Category 3 is ($b_i - c_3$) for item i, which corresponds to ($\gamma_{00} - \gamma_{i0} + \delta$). It can also be seen that $\delta = c_2 - c_3$.

2.2.2 Model identification

Scale indeterminacy refers to the estimation of a DIF parameter that is not absolute but related to the other DIF parameters in the same test (de Ayala, 2009). To solve scale indeterminacy, it is necessary to set constrains in order to identify the model. There are three popular approaches: the mean of the person ability parameter or the mean of the item difficulty parameter can be constrained to an arbitrary value (e.g., zero), or a set of anchor items can be selected to serve as a matching criterion variable (Chen et, al, 2014; Wang, 2004).

If the mean person ability is fixed to zero, then γ_{00} is equal to zero; thus, item difficulty parameters can be directly estimated by γ_{k0} without the need to use a reference indicator. Equation (26) reduces to

$$\text{Logit}\,(p_{ij}) = u_{0j} + \gamma_{00} + \gamma_{i0} = u_{0j} - (-\gamma_{i0}) \tag{38}$$

which is an equivalent of the Rasch model where u_{0j} is equal to person ability θ_j and γ_{i0} is the fixed item effect equal to item difficulty b_i. This is referred to as the "person centering" approach (de Ayala, 2009). Although this method can assess all items for DIF, the assumption of equal ability for the reference and focal group is questionable in practice.

If the mean item difficulty is fixed to zero, referred to as the equal mean difficulty method (EMD) by Wang (2004), γ_{00} can be freely estimated and represents the overall ability level in Equation (23). The number of dummy-coded item indicator variables W_{kij} equals to $I - 1$ to ensure the model is identifiable. In the DIF model Equation (26), γ_{00} now represents the mean ability level for the reference group. γ_{i0} is the fixed item effect, while γ_{i1} represents the difference in item effect

between the groups. This is referred to as the "item centering" (de Ayala, 2009) approach. By definition, the EMD method requires the presence of DIF in a test to be balanced in size and direction. In other words, the sum of all DIF should equal zero; otherwise, the equal mean item difficulties for the reference and focal group assumption is violated.

If an unbiased item or item set is used to anchor the scale, referred to as the constant anchor item method by Wang (2004), the magnitude of DIF parameters of the reference items are fixed to establish a common metric and all the other items can be assessed for DIF. For example, if the difficulty of the reference items is assumed to be zero, then γ_{00} can be freely estimated and represents the overall ability level in Equation (23). In the DIF model Equation (26), γ_{00} now represents the mean ability level for the reference group. γ_{i0} is the item difficulty in Equation (23). In the DIF model Equation (26), γ_{i0} now represents the item difficulty for the reference group, while γ_{i1} represents the group difference in item difficulty. The constant anchor item method assumes that the set of reference items are free of DIF; thereby, the selection of reference items is crucial.

2.2.3 Performance of HGLM in DIF detection

Dichotomous items Cheong & Kamata (2013) explored the performance of the EMD method and the constant anchor item method for DIF detection in the HGLM framework and compared the results to the well-researched MH procedure. For the EMD method, the item centering approach was used and constraints for the dummy-coded item indicators were designed to deviate the indicator from the reciprocal of the total number of items. For the constant anchor item method, person centering was used and constrains were 0 and 1. Six dichotomous items with

one reference item for 1000 examinees were simulated (*R*55/*F*500, where *R* represented the reference group and *F* the focal group). The ability difference between the reference and focal groups (i.e., impact) was set to .2 and indicated small impact. The authors manipulated the DIF patterns (constant and balanced), percentage of DIF items (1/3, 5/6, and 1), and whether the reference item was DIF-free. If the DIF pattern was constant, meaning DIF items were in favor or against one group, the sum of DIF for each item would not equal zero; then the assumption of the EMD method was violated. If the reference item exhibited DIF, then the assumption of the constant anchor item was violated.

The results showed that both methods performed well when their respective assumptions were met. When the pattern of DIF was asymmetric, especially when most items with DIF are biased in favor of or against one group (i.e., constant pattern of DIF), the EMD method showed a tendency to overestimate the magnitude of DIF for an item that is DIF-free. The method tended to underestimate the magnitude of DIF for an item with DIF. The mean group difference of ability estimates was biased as well. In addition, the EMD method produced similar DIF estimates to that of the MH procedure. When the reference item was not DIF-free, the constant anchor item method falsely identified items that were DIF-free and overestimated the magnitude of some items with DIF while underestimating the magnitude of other items with DIF. The constant anchor item method also severely overestimated the group mean ability difference parameter. The authors concluded that if the sum of DIF is approximately zero, or when there is a small amount of items that exhibit DIF so the total DIF magnitude is relatively small, the EMD method is recommended since it is more robust and the consequence of violating its assumption is relatively small. For the constant anchor item method, it is crucial that the reference items are DIF-free.

Chen et al. (2014) explored the performance of HGLM in DIF detection using the equal mean ability (EMA) method with rank-based strategy (Woods, 2009) and the constant anchor item method with the DIF-free-then-DIF strategy (DFTD; Wang, Shih, & Sun, 2012). For the EMA method, the mean ability for both groups was fixed to zero; *I* item indicators were used. The reference items were selected using the rank-based strategy proposed by Woods (2009). For the constant anchor item method, the difficulty of the anchor items were fixed to zero; *I* -1 item indicators were used. The reference items were selected by performing the constant anchor item method iteratively to choose the items with the smallest mean absolute DIF effect values. The authors manipulated the amount of impact (0 and .5), number of reference items (1 and 4), DIF patterns (constant and balanced), percentage of DIF items (0%, 20%, and 40%), and sample size (R500/F250, R250/250, and R250/F150).

The authors first compared the accuracy rate in selecting DIF-free items and found that the constant anchor item method with DFTD outperformed the EMA method with rank-based strategy across various conditions, and thus recommended the former. The EMA method performed poorly with the presence of impact when the DIF pattern was balanced; the authors concluded that it was because the ability and item parameters for the focal group were shifted under the equal group mean ability constraint; thus, the magnitude of DIF for the DIF items that favored the focal group were shifted towards the mean ability because of being incorrectly selected as the reference items. The authors then compared the Type I error rates and power rates of DIF assessment. For the constant anchor item method with DFTD, the presence of impact had little effect. Its power rates decreased when the sample size decreased, percentage of DIF increased, or DIF pattern was constant instead of balanced. The lowest power rate combination was when the sample size was small with high percentage of DIF items and constant DIF pattern, possibly due to the slight

inaccuracy in selecting the reference items. Four reference items produced a higher power rate but also slightly increased the Type I error rate. The EMA method with rank-based strategy performed well under the no-impact condition; however, with the presence of impact, it performed poorly with highly inflated Type I error; thereby, the authors recommended against it.

Polytomous items The performance of HGLM in DIF detection for Polytomous items is relatively sparse; thus more research is needed. Willams & Beretvas (2006) extended Kamata's (2011) dichotomous HGLM to polytomous items and compared the results to GMH. The threshold difference between response categories is a constant fixed effect across items and persons. The item discrimination parameter was not estimated, instead fixed at 1. The authors generated the data using the original Muraki's rating scale model and the constrained version with fixed item discrimination parameter *a*, using item parameters adapted from Koch (1983). Parameter recoveries of HGLM were compared with IRT. With the original rating scale model, IRT performed better in parameter recoveries; while with the constrained model, which is the correct model for HGLM, both HGLM and IRT performed similarly well.

Ryan (2008) expanded on the work of Williams and Beretvas (2006) by exploring the use of continuous grouping variable, and more simulation conditions. Both studies used a person centering with equal mean ability method to identify the model; neither simulated group difference in ability, i.e. no impact. Both studies found that HGLM performed similarly to GMH, with GMH showed higher power in certain small sample size conditions. For sample size larger than 1000, the two methods behaved similarly.

2.3 DIF Assessment with Rating Scale

In the field of psychological measurement and health studies, questionnaires are widely used. Traditionally, these instruments are developed and validated based on classical testing theory (CTT) (Andrich, 2011; Pesudovs, 2010). In recent years, IRT is becoming increasingly popular (Massof, 2011). Compared to simple summary of scores over response categories as in the CTT framework, a probabilistic model in the IRT framework, such as a rating scale model, has obvious advantages in evaluating the psychometric properties of a questionnaire with rating scales, which is commonly used in health studies (Andrich, 2011; De Ayala, 2013).

It has been increasingly common for researchers to use the rating scale model to validate the instruments in health studies. Massof (2005) evaluated the measurement properties of 4 visual functioning instruments with both Andrich's and Muraki's rating scale model. Estimates of model parameters, model fits, and measurement precision were compared. The author found that the two models produced linearly related parameter estimates, with Muraki's model produced a better overall fit, while Andrich's model a better average fit for person and item. Gothwal and colleges (2012) fitted Andrich's rating scale model to data of two forms of the visual functioning scales, and assessed the instruments' psychometric properties, such as measurement precision, dimensionality, DIF, and so forth. The authors concluded that the rating scale model fitted the data well and can be a useful tool in processing such data. Similar studies includes using rating scale model to validate visual functioning instruments (Dougherty & Bullimore, 2010; Massof, 2007; Stelmack et al., 2004 ;Velozo, Warren, Hicks, & Berger, 2013;), quality of life instruments (Denny, Marshall, Stevenson, Hart, & Chakravarthy, 2007; du Toit, Palagyi, Ramke, Brian, & Lamoureux, 2008; Williams, Brian, & Toit, 2012), and other health-related instruments

(Dougherty, Nichols, & Nichols, 2011; Eakman, 2012; González, Sierra, Martínez, Martínez-Molina, & Ponce, 2015; Rovner et al., 2011).

Although DIF assessment in health studies is becoming increasingly popular, DIF assessment under the rating scale model framework is relatively new and studies are lacking. Researchers have been using rating scale model to validate instruments and analyze empirical data (e.g. Wolle et al., 2011; Massof, Deremeik, Park, & Grover, 2007), but few discussed the application to DIF. Ahmadian and Massof (2008) fitted a rating scale model to a visual functioning instrument and examined DIF in low vision patients using the implemented procedure in RUMM2020 (Andrich, Lyne, Sheridan, & Luo, 2003). Visual measures were binned into arbitrary ranges and the observed mean of the bin was compared with the item characteristic curve. The authors found a result of 15 flagged items out of 48. Gothwal and colleges (2012) briefly discussed DIF in terms of difference in logits without statistical testing. Dye, Eakman, and Bolton (2013) fitted a rating scale model to a gait ability instrument to examine the psychometric properties of the instrument, while also discussed DIF using the logistic regression t-test implemented in WINSTEPS (WINSTEPS, 2009). These studies were conducted using empirical data, so it was difficulty to discuss type I error rate and power. It is necessary to have more studies to explore the behavior of DIF in rating scale model.

3.0 Method

The purpose of this study is to evaluate the performance of HGLM and its comparison to the GMH procedure and logistic regression when used to evaluate DIF for polytomous items. Two simulation studies were used for this purpose. For study 1, HGLM with DFTD strategies were evaluated for its accuracy in selecting anchor items. For study 2, HGLM with constant anchor item method was used for DIF detection, and the results were compared to GMH and polytomous logistic regression. Three research questions were addressed:

1. How accurately can HGLM select DIF-free items as anchor items for DIF analysis?
2. What is the Type I error rate for DIF detection using HGLM and how does it compare to using GMH and logistic regression?
3. What is the statistical power for DIF detection using HGLM and how does it compare to using GMH logistic regression?

In this chapter, the fixed factors and manipulated factors of the simulation study are discussed first, followed by evaluation criteria, an introduction of data generation and validation, and finally, a description of data analysis.

3.1 Fixed Factors

3.1.1 Scale length

Studies have shown that for a long test (> 20 items), scale length has little effect on DIF detection. A scale with less than 10 items is likely to be too short to produce sufficient power since observed score based on a short scale is less reliable and more likely not a sufficient proxy for the latent ability; and this mismatch is more serious when large trait differences between groups are present. However, Scott et. al (2009) found that a scale as short as 5 items could produce similar results as 20 items using polytomous logistic regression (PLR) on simulated health-related data. For a scale with more than 20 items, scale length has an insignificant impact on the performance of DIF assessment. For this reason, in this study the number of items was fixed at 20.

3.1.2 Item discrimination parameter α

Item discrimination parameter pertains to what has been referred to as the uniform and nonuniform DIF. For dichotomous items, a varying *a* parameter means a nonuniform DIF pattern; however, for polytmous items, nonuniform DIF is not necessarily only a function of *a*, but can occur without the inferencing of *a* as shown by Su and Wang (2005).

Although the item discrimination parameter *a* could vary for each score category, polytomous DIF studies commonly define the model or simulate data with a common item discrimination parameter. In the two studies which explored polytomous HGLM compared to GMH (Ryan, 2008; Williams & Beretvas. 2006), item discrimination parameter was held constant across all items.

Since complicated patterns of DIF can occur without involving item discrimination parameter *a*, in this study, the item discrimination parameter was held constant in data generation when manipulating DIF for reference group. Instead, the constant and unbalanced patterns of DIF were manipulated.

3.1.3 Model identification method

In HGLM, in order to identify the model, it is necessary to set constraints. There are three popular approaches for this purpose: equal mean ability method, equal mean item difficulty method, and constant anchor item method. Equal mean ability method assumes the mean ability for focal and reference groups is equal, which is an unlikely scenario in practice. In addition, numerous studies have shown that when there is impact, most DIF estimation methods perform poorly. Equal mean item difficulty method assumes the presence of DIF in a test to be balanced in size and direction. When this assumption is met, it performs well; however, when the pattern of DIF becomes complicated, equal mean item difficulty method can show an inflated Type I error, and in the meantime, underestimate the true size of DIF (Cheong & Kamata, 2013). Thus, constant anchor item method is recommended. However, the constant anchor item method requires the anchor items to be free of DIF and when this assumption is violated, performs very poorly. Based on these studies, the constant anchor item method was chosen for this study.

3.2 Manipulated Simulation Conditions

In this study, 6 independent variables were manipulated for study 1(Table 2): 1. number of anchor items (1, 4); 2. impact size (0, 1); 3. sample size (R400/F100, R250/F250. R800/F200, and R500/F500); 4. percentage of DIF (0%, 20%, and 40%); 5. magnitude of DIF (.2, .6); 6. DIF patterns (constant, balanced, and unbalanced). In total there are 208 conditions. For study 2, the same 6 independent variables were manipulated, with 2 extra levels of sample size (R4000/F1000, R2500/F2500) added to the original sample size of 4 levels. For study 2, in total there are 312 conditions.

3.2.1 Anchor items

As Cheong and Kamata (2013) demonstrated, the consequence of using a contaminated anchor item is serious; thereby the anchor items must be carefully selected. Two approaches have been proposed in literature for selecting anchor items. Woods (2009) proposed a rank-based method based on the all-other-items method, in which all the other items except the studied item are used as reference. This method was originally proposed for IRT with the likelihood ratio test, but can be generalized to HGLM as well. The other approach is to use an iterative method (Shin & Wang, 2009), which involves the following steps:

1. Use Item 1 as anchor and test all the other items for DIF using HGLM, and estimate DIF index for each studied item.
2. Use the next item as anchor and test all the other items for DIF using HGLM, and estimate DIF index for each studied item.
3. Repeat Step 2.

4. Compute an absolute value of DIF index for each item over all iterations, and then choose the item(s) with the smallest index as the anchor item(s).

Chen et al. (2014) compared the accuracy of these two methods and found that iterative constant anchor item method with DFTD outperformed the equal mean ability method with rank-based strategy across various conditions. The equal mean ability method performed poorly with the presence of impact when the DIF pattern was balanced. The authors concluded that it was because the ability and item parameters for the focal group were shifted under the equal group mean ability constraint. Thus, the magnitude of DIF for the DIF items that favored the focal group were shifted towards the mean ability as a result of being incorrectly selected as the reference items. The authors thus recommended constant anchor item method with DFTD method; as a result this is the anchor selecting method for this study.

The appropriate number of anchor items was investigated in many studies; generally speaking, a larger number of anchor items is associated with higher power of DIF detection. This effect is more prominent when the number increases from 1 to 4, but less so when the number increases to 50 (Thissen et al., 1988). Wang and Yeh (2003) explored 1-, 4-, and 10- items as anchor for IRT likelihood ratio test and found that 1 anchor item could give satisfying results, although 4- and 10- anchor items can produce even higher power. Shih and Wang (2009) found similar results using the MIMIC method, and concluded that 4 anchor items were enough to produce sufficient power. Woods (2009) discussed that using a 1- item anchor can minimize DIF contamination, which is crucial considering the consequences of contaminated anchor items are serious. However, for a small sample, 1-item anchor may not produce enough power; in addition, 1-item anchor may not be a sufficient estimation of the matching variable. Given these results, in this study, 1 and 4 anchor items were compared.

3.2.2 Latent trait parameter difference between groups and impact (0, 1)

The mean latent trait difference between the focal and reference groups, or "impact", has significant influence on the performance of DIF detection. When impact is present, the matching variable may not be a sufficient index of the latent proficiency; thereby may cause inflated Type I error. Many studies have found that with large impact (mean difference ≥ 1 on standard normal distribution), many parametric and nonparametric methods show a tendency to have inflated Type I error rates, particularly when in combination with a high percentage of DIF items, a shorter scale, and smaller magnitude of DIF. However, some studies found that when the size of impact is moderate (mean difference = .5 on standard normal distribution), group mean trait difference shows little effect (Kristjansson et. al., 2005).

In the few studies that examined the effect of nonnormality of ability distribution on polytomous DIF detection, some found that nonnormality in ability distribution causes more difficulty in estimation (Moyer, 2013). Some found it has little effect on the performance of DIF detection when impact size is moderate (Kristjansson et. al., 2005).

Given these results, the latent trait parameter θ for the reference group were simulated from a standard normal distribution N (0, 1), while for the focal group θ were simulated from either a standard normal distribution N (0, 1) or N (-1, 1). This represents a medium-large impact between groups, which is also reportedly a common value of impact between focal and reference groups (Donoghue, Holland & Thayer, 1993).

3.2.3 Sample size and sample size ratio

Various studies have shown that a larger sample size is usually related to higher power; however, in some conditions it may inflate Type I error. The parametric methods tend to have larger sample size requirements than the nonparametric methods; to acquire a decent power rate, some studies suggest a total sample size of 2000 is necessary, while a sample size smaller than 1000 typically does not produce sufficient power (Elosua & Wells, 2013; French & Miller, 1996; Hildago & Gómez, 2006). In education setting, most sample size for DIF detection studies with polytomous responses is between 300-1000, with the smallest total sample size considered is 250 (Ryan, 2008). For health studies, the guidelines are lacking; Scott and colleges (2009) recommended a sample size of at least 200 participants per group for ordinal logistic regression to achieve satisfying power.

The effect of unequal number of participants for the focal and reference groups is not well understood. Some studies found unbalanced sample size ratio causes a decrease of power (Kristjansson et. al, 2005; Ryan, 2008). This result is not surprising; an unequal sample size means smaller subjects in the focal group, and consequentially fewer data at each level of the matching variable, resulting in less reliable matching. Researchers have considered sample size ratio as extreme as R20:F1 (Woods & Grimms, 2011). Typically, the sample size ratio is between R1:F1 and R4:F1.

Given these previous findings, in this study 4 combinations of sample size were considered in study 1: R400/F100, R250/F250, R800/F200, and R500/F500. R400/F100 represents a relatively small and unbalanced reference to focal group ratio, which will provide meaningful guidelines for practitioners to use regarding minimum sample size requirement; while R500/F500 is a more ideal condition. In order to make the findings more comparable to previous studies such

as Williams and Beretevas (2006), and to see how the methods would perform under large sample size, 2 extra large sample size combinations were included for study 2: R4000/F1000 and R2500/R2500.

3.2.4 Percentage of DIF items (0%, 20%, and 40%)

A higher percentage of DIF can cause difficulty in selecting anchor items, and affect the accuracy of the matching variable, thus result in inflated Type I error and reduced power. Some researchers argue that it is not the percentage of DIF items but the magnitude of overall DIF for the test, which is a function of the percentage of DIF items and the DIF patterns, that is causing the inflation (Su & Wang, 2005; Wang & Su, 2004; Wang & Yeh, 2003). Most studies on polytomous responses explored DIF contamination at 5% to 30% (e.g. Fidalgo & Bartram, 2010; Flowers, Oshima, & Nambury, 1999; Gomez-Benito et. al., 2013; Hidalgo & Gomez, 2006; Meade, Lautenschlager, & Johnson, 2007; Penfield, 2007; Penfield & Algina, 2003; Wang & Su, 2004), while some studies explored contamination as high as 66% (Woods & Grimm, 2011). In this study, 3 conditions were manipulated: 0%, 20%, and 40%.

3.2.5 Magnitude of DIF (.2, .6)

Studies have shown that DIF of a small size can be hard to detect, while increasing the magnitude of DIF is typically related to the increase of power. This effect is more prominent when combined with a large sample size. When the magnitude of DIF is small (.1 or .2), power is very poor; for constant DIF, increase the magnitude of DIF (> .4) can increase the power rate. However, there is also a slight increase of Type I error rate (Elosua & Wells, 2013; Hildago & Gómez, 2006;

Scott, 2009). Typically, the magnitude of DIF is between (.2, .8) for simulated DIF studies. In health studies, it is common to use guidelines from education settings (for example, the simulation study conducted by Scott et. al, 2009). Magnitude of DIF smaller than .2 is common in empirical health studies (for example, Dorans& Kulick, 2006; Terluin, Smits, Brouwers, & de Vet, 2016).

In this study, 2 levels of magnitude were studied. A magnitude of .2 represents a relatively small DIF effect, while .6 represents a medium-large DIF effect.

3.2.6 DIF patterns (constant, balanced, unbalanced)

For dichotomous items, a constant pattern refers to when DIF items are in favor or against one group constantly across items; while balanced DIF refers to when some DIF favor the focal group and others favor the reference group, resulting the magnitude of DIF to be balanced across item categories. In polytomous items, DIF can take on more complex patterns since the patterns can also be exhibited in response categories. For an item with unbalanced DIF within categories, it is possible for DIF to only exist in the lower categories or only in the higher categories. Studies have shown that balanced DIF is harder to detect and can decrease power rate. When DIF is only present in the highest or lowest response category, power can be as poor as below 50% (Fidalgo & Bartram, 2010; Su & Wang, 2005). Although the effect size of DIF remained the same (measured by the unsigned area measure), Fidalgo and Bartram (2010) found that the high-unbalanced pattern seemed to produce worse results than low-unbalanced pattern in terms of Type I error and power rate.

Typically, DIF studies manipulate DIF patterns by manipulating category boundary parameter. As previously demonstrated, for an item *i* with *t* score categories, the category boundary

parameter in graded response model is partitioned into two parameters in Muraki's rating scale model: the location parameter b_i and the category threshold c_t.

This study manipulated 3 DIF patterns: constant, balanced, and high-unbalanced similar to Fidalgo and Bartram (2010). For an item that exhibits DIF, the item difficulty parameter b_{it} for the focal group were changed accordingly.

$$\text{Constant DIF: } b_{it}F = b_{it}R + s,\ t = 1, 2 \tag{39}$$

$$\text{Balanced DIF: } b_{i1}F = b_{i1}R + s;\ \ b_{i2}F = b_{i2}R - s$$

$$\text{High-unbalanced DIF: } b_{i2}F = b_{i2}R + s$$

where s represents the magnitudes of DIF, t represents the tth response category. These DIF patterns are common in the literature (e.g. Chen et al., 2014; Su & Wang, 2005).

Note that since $b_{it} = b_i - c_t$, DIF can exhibit at b_i or c_t. When DIF is exhibited at b_i, since b_i is constant across response categories within an item, the only DIF pattern possible is the constant pattern.

$$b_{i1}F = b_iF - c_2 = (b_iR + s) - c_2 = b_{i1}R + s \tag{40}$$

$$b_{i2}F = b_iF - c_3 = (b_iR + s) - c_3 = b_{i2}R + s$$

When DIF is exhibited at c_t rather than b_i, DIF can exhibit more complicated patterns.

$$\text{Constant DIF: } b_{i1}F = b_iF - c_2 = b_iR - (c_2 + s) = b_{i1}R - s \tag{41}$$

$$b_{i2}F = b_iF - c_3 = b_iR - (c_3 + s) = b_{i2}R - s$$

$$\text{Balanced DIF: } b_{i1}F = b_iF - c_2 = b_iR - (c_2 + s) = b_{i1}R - s \tag{42}$$

$$b_{i2}F = b_iF - c_3 = b_iR - (c_3 - s) = b_{i2}R + s$$

$$\text{High-unbalanced DIF: } b_{i1}F = b_iF - c_2 = b_iR - c_2 = b_{i1}R \tag{43}$$

$$b_{i2}F = b_iF - c_3 = b_iR - (c_3 + s) = b_{i2}R - s$$

Since Equation (40) and Equation (41) showed the same constant pattern except on different direction, these two patterns are combined into one condition. This simulation study manipulated 3 DIF patterns as defined in Equation (39).

Table 2 Specifications of Simulation Conditions

1. Number of anchor items 1.1. 1 1.2. 4
2. Impact 2.1. 0 2.2. 1
3. Sample size 3.1. R400/F100 3.2. R250/F250 3.3. R800/F200 3.4. R500/F500 3.5. R4000/F1000 3.6. R2500/F2500
4. Percentage of DIF 4.1. 0% 4.2. 20% 4.3. 40%
5. Magnitude of DIF 5.1. .2

5.2. .6
6. DIF pattern 6.1. Constant 6.2. Balanced 6.3. Unbalanced

3.3 Evaluation Criteria

3.3.1 Accuracy of selecting DIF-free items

After fitting HLGM on the data, iterative constant anchor item method was applied in order to select DIF-free items as anchor to identify the model and connect the metric scales between groups. The accuracy rate in selecting such DIF-free items were evaluated; for example, for the 4-anchor condition, the accuracy rate were .25, 50, .75, and 1, respectively, if 1, 2, 3, or 4 selected items are indeed DIF-free.

3.3.2 Type I error rate

Type I error rate is defined as the percentage of the mis-identification of a DIF-free item as a DIF item over the number of replications. When using a level of α=.05, type I error rate should be around .05.

3.3.3 Statistical power

Statistical power is calculated by the proportion of the times a DIF item is correctly identified over the number of replications. Statistical power, as well as type I error rate, are meaningful tools for practitioners.

3.4 Data Generation and Validation

3.4.1 Data generation

Trait parameter θ for the focal and reference group examinees were simulated from a standard normal distribution. For the reference group θ were simulated from a standard normal distribution N (0, 1); while for the focal group θ were simulated from either a standard normal distribution N (0, 1) (no impact) or a normal distribution N (-1, 1) (impact present).

Item responses were generated using a constrained form of Muraki's rating scale model (Muraki, 1990; Williams & Beretvas, 2006), as showed in Equations (31)-(33). Since there was no simulation study conducted in the health research area using HGLM, item parameters used in this study were modified from Willams and Beretvas (2006). The *a* parameter was constrained to 1 in order to obtain a constant threshold difference adopted by the HGLM. The location parameter b_i was generated from a uniform distribution [-1, 1] while the two threshold parameters c_1 and c_2 were fixed to .5 and -.5, respectively. The two category difficulty parameters b_{i1} and b_{i2} were generated as $b_{i1} = b_i - c_1$ and $b_{i2} = b_i - c_2$. The location parameter b_i for the first item is set to zero in order to identify the HGLM. For the condition of 20% items with DIF, the last 4 items were

manipulated to exhibit DIF by applying Equations (39). For the condition of 40% items with DIF, the last 8 items were manipulated to exhibit DIF.

Item responses were generated using the IRTGEN program (Whittaker, Fitzpatrick, Williams, & Dodd, 2003) for SAS, a collection of SAS macros for generating dichotomous and polytomous IRT data. Data generation was performed using SAS 9.4.

Table 3 Item Parameters for Data Generation for the Reference Group

Item	a_i	b_{i1}	b_{i2}
1	1	-0.50	0.50
2	1	0.01	1.01
3	1	-0.02	0.98
4	1	-0.90	0.10
5	1	-0.91	0.09
6	1	0.49	1.49
7	1	-1.08	-0.08
8	1	-0.91	0.09
9	1	-0.37	0.63
10	1	-1.07	-0.07
11	1	-0.40	0.60
12	1	-1.23	-0.23
13	1	-0.04	0.96
14	1	-0.51	0.49
15	1	-0.93	0.07
16	1	0.38	1.38
17	1	0.09	1.09
18	1	-0.65	0.35
19	1	0.44	1.44
20	1	-1.30	-0.30

3.4.2 Data validation

To examine the adequacy of generated data, data validation was performed on a few randomly selected datasets with 100 replications. First, generated trait parameter θ were checked by examining the descriptive statistics such as means and variances (Table 4). The simulated participants' responses were fitted to a graded response model with fixed *a* parameter equaled to1 to check for discrepancies between the estimated parameters for focal and reference group using MULTILOG. Table 5 presents an example of estimated parameters for focal and reference groups with 4 items exhibiting constant DIF with a magnitude of .6.

Table 4 Estimated θ for the no impact and impact groups

θ	mean	sd	skewness
No impact	.006	.1.002	.00
Impact	-1.004	1.003	-.002

Table 5 Generated item parameters for the focal and reference groups when impact = 0

	Estimated parameters for the reference group		Estimated parameters for the focal group		Discrepancies of parameter estimates between focal and reference groups		Discrepancies between generating and estimated reference group parameters		Discrepancies between generating and estimated focal group parameters	
item	b_{i1}	b_{i2}	b_{i1}	b_{i2}	db_{i1}	db_{i2}	db_{i1}	db_{i2}	db_{i1}	db_{i2}
1	-0.48	0.62	-0.55	0.42	0.07	0.20	-0.02	-0.12	0.05	0.08
2	-0.68	0.38	-0.36	0.54	-0.32	-0.16	0.15	0.09	-0.17	-0.07
3	0.55	1.57	0.47	1.50	0.08	0.06	-0.13	-0.15	-0.05	-0.08
4	-0.69	0.25	-0.76	0.10	0.07	0.16	0.00	0.06	0.07	0.21
5	0.10	1.19	0.21	1.31	-0.11	-0.12	0.21	0.12	0.10	0.00
6	0.41	1.30	0.29	1.26	0.11	0.04	-0.02	0.09	0.10	0.13
7	-1.52	-0.42	-1.46	-0.43	-0.06	0.01	0.09	0.00	0.03	0.01
8	0.19	1.36	0.18	1.04	0.02	0.32	0.05	-0.12	0.06	0.20
9	-1.45	-0.41	-1.20	-0.27	-0.25	-0.13	0.18	0.15	-0.06	0.01
10	0.30	1.45	0.19	1.18	0.11	0.27	0.02	-0.12	0.13	0.15
11	-0.03	1.03	0.01	1.07	-0.03	-0.05	0.12	0.07	0.08	0.02
12	-0.67	0.29	-0.81	0.14	0.14	0.15	-0.11	-0.07	0.03	0.08
13	-0.94	0.05	-1.09	-0.22	0.15	0.26	-0.11	-0.10	0.03	0.16
14	0.49	1.40	0.42	1.65	0.07	-0.25	-0.03	0.06	0.03	-0.19
15	-1.26	-0.26	-1.24	-0.21	-0.02	-0.05	-0.06	-0.07	-0.09	-0.12
16	-1.09	-0.13	-1.16	-0.15	0.07	0.01	0.04	0.08	0.11	0.09
*17**	0.06	1.01	0.68	1.63	-0.62	-0.62	0.07	0.11	-0.55	-0.50
*18**	-1.47	-0.62	-0.98	0.11	-0.49	-0.73	0.01	0.16	-0.48	-0.58
*19**	-1.41	-0.40	-0.81	0.30	-0.60	-0.71	0.12	0.10	-0.48	-0.60
*20**	-1.45	-0.32	-0.94	0.01	-0.51	-0.33	-0.02	-0.15	-0.52	-0.47

Note: * indicates items with DIF for the focal group. The pattern of DIF is constant with s=.6.

3.5 Data Analysis

3.5.1 Estimation methods

Simulated examinee responses were analyzed by fitting a HLGM model using PROC GLIMMIX in SAS 9.4.

There are various methods for estimating high dimension HGLM likelihood functions. Quasi-likelihood methods, such as MQL and PQL, linearize the estimation function for approximation, while integral approximation methods, such as GHQ, AGQ, and Laplace, approximate the true likelihood function using numeric integration. Bayesian methods, such as MCMC and MCEM, can also be used. Each of these methods presents its advantages and shortcomings. Laplace approximation performs well in general, balancing both computational intensity and estimation precision, thus recommended by many researchers (Capanu, et al., 2013; Kim et al., 2013). However, Laplace has a tendency to produce larger standard errors (Diaz, 2007; Joe, 2008; Schoeneberger, 2016); thereby, researchers suggest practitioners to use with caution when the estimation of standard errors is important to the study, such as for a Wald test. In addition, Laplace has also been observed to show inaccuracies in estimation of random effect variance (Browne & Draper, 2006; Goldstein & Rasbash, 1996; Cho, Rabe-Hesketh, 2011; Cho et al., 2012; Schoeneberger, 2016). If the size of random effect variance is small, PQL is recommended since it produces more accurate estimation with decent sample size; moreover, as a linearization method, it is easy to compute and implement (Schoeneberger, 2016). If the size of random effect variance is large, PQL should be avoided as it produces large downward bias, while Laplace is recommended (Capanu et al., 2013; Diaz, 2007; Kim et al., 2013; Pinheiro & Chao, 2006). If computational burden and implementation difficulty can be managed, AGQ is an accurate

approximation method, provided given a decent number of quadrature points, usually larger than 5 is sufficient (Capanu, et al., 2013; Joe, 2008). However, for a complex model, the computation is likely to be intense. Bayesian methods, such as MCMC and AIP with adaptive quadrature, are also viable options (Cho et al., 2012; Jiao et al., 2013); however, computational intensity is a serious concern for this type of method; in addition, they are not as easy to implement.

Given the complexity of the model, computational intensity is likely to be of concern for this study, thereby supporting the use of PQL or Laplace. Considering PQL's many disadvantages, such as a tendency of producing downward bias toward zero and inconsistent estimates and a lack of proper model fit statistics, Laplace method was used in this study.

3.5.2 Generalized Mantel-Haenszel and polytomous logistic regression

As comparisons to HGLM, GMH and logistic regression were also be calculated for this study. GMH is not commonly used in health studies, but is one of the most popular methods used in education testing. Dorans and Kulick (2006) have applied the Mantel-Haenszel and standardized procedures to the Mini-Mental State Examination and demonstrated the potential of these types of DIF detection methods on health-related data. These types of methods are well-studied in education testing and have established guidelines for practitioners to use. Therefore, it is useful to compare the performance of HGLM to GMH, which is particularly popular in education testing. Unlike GMH, logistic regression is very commonly used in health studies to evaluate DIF; thus, it is necessary to compare the results of HGLM to logistic regression as well. SAS 9.4 was used for the analysis.

4.0 Results

This section presents the results for simulation study 1 and 2. The section is organized in 4 sections. The first 3 sections correspond to each of the 3 research questions. Section 1 presents findings from study 1, answering the first research question: how accurately can HGLM select DIF-free items as anchor items for DIF analysis? Section 2 presents findings from on the Type I error rates from study 2, answering the second research question: what is the Type I error rate for DIF detection using HGLM, and how does it compare to using GMH and logistic regression? Section 3 presents findings on power from study 2, answering the third research question: what is the power rate for DIF detection using HGLM, and how does it compare to using GMH and logistic regression? Section 4 presents a summary of the results for study 2.

4.1 Results of Study 1

For study 1, datasets were generated for each of the 208 simulation conditions. Each item from each dataset was then fitted with a HGLM model with pseudo-likelihood with residual method. The convergence criterion for PROC GLIMMIX in SAS 9.4 was set to .001. One hundred replications were performed considering the long estimation time. Estimation of all replications converged.

Once the model converged, accuracy of selecting DIF-free items was calculated. In the HGLM model, DIF was identified as the interaction terms between item and groups. Interaction parameter estimations of all 20 items were ranked, and items with the smallest parameter

estimations were selected as anchor items. The selection was considered successful if the selected anchor item avoided any items with DIF. The percentage of accurate selection were computed over 100 replications for each condition.

4.1.1 Accuracy rates of using HGLM to select DIF free items

For conditions with 20% DIF items, the results were presented in Table 6. For conditions with 40% DIF items, the results were presented in Table 7.

Table 6 Accuracy (%) of Selecting DIF-Free Items as Anchor with 20% DIF Items

			Impact = 0		Impact = 1	
Sample size	Magnitude	DIF patterns	1-item anchor	4-item anchor	1-item anchor	4-item anchor
R400/F100	.2	Constant	81	85	82	84.25
		Balanced	82	82	81	83.75
		Unbalanced	80	81.5	82	77.75
	.6	Constant	95	96.5	96	97
		Balanced	88	86.25	85	83.25
		Unbalanced	84	82.75	85	77.75
R250/F250	.2	Constant	89	80.75	88	85.25
		Balanced	82.5	83	79	81
		Unbalanced	80	80.75	73	77
	.6	Constant	99	99.25	98	98.5

		Balanced	86	86.5	86	84.25
		Unbalanced	86	85.75	82	84
R800/F200	.2	Constant	90	87	88	88
		Balanced	84	84.75	80	82.25
		Unbalanced	84	81.5	84	81
	.6	Constant	100	99.5	100	99.75
		Balanced	91	87	86	84.25
		Unbalanced	87	87	86	84
R500/F200	.2	Constant	93	88	93	88.5
		Balanced	85	83.25	85	82.75
		Unbalanced	86	81.75	83	81.75
	.6	Constant	100	100	99	99.75
		Balanced	86	86	85	84
		Unbalanced	98	95.75	89	89.5

Table 7 Accuracy (%) of Selecting DIF-Free Items as Anchor with 40% DIF Items

			Impact = 0		Impact = 1	
Sample size	Magnitude	DIF patterns	1-item anchor	4-item anchor	1-item anchor	4-item anchor
R400/F100	.2	Constant	55	60.25	57	61.25
		Balanced	55	59.5	62	57.25
		Unbalanced	53	56.5	52	57
	.6	Constant	83	78.25	78	75
		Balanced	68	66.75	64	58.5
		Unbalanced	56	62.25	53	58.25
R250/F250	.2	Constant	54	58.25	67	62
		Balanced	53	54.5	56	59.5
		Unbalanced	53	60.25	59	58.5
	.6	Constant	84	83.25	83	77.25
		Balanced	63.5	61.25	56	57.5
		Unbalanced	57.5	57.75	54	58.75
R800/F200	.2	Constant	60	62.5	60	55.5
		Balanced	62	60.75	51	58.75
		Unbalanced	64.5	59.5	56	61
	.6	Constant	90	85.25	84	84.75
		Balanced	69	62.25	67	61.25
		Unbalanced	61	61	57	57
R500/F200	.2	Constant	64.5	62.75	65	63

		Balanced	49	60.25	58	60.5
		Unbalanced	70	58.5	55	59
	.6	Constant	89	89.75	90	91.5
		Balanced	60	64.75	61	57.75
		Unbalanced	65	68.25	62	60.5

4.1.2 ANOVA results of study 1

There were in total 6 conditions manipulated in Study 1, the number of anchor items (1, 4), the presence of impact (0, 1), the sample size and ratio (R400/F100, R250/F250, R800/F200, R500/F500), the percentage of DIF items (20%, 40%), the magnitude of DIF (.2, .6), and the pattern of DIF (constant, balanced, high unbalanced). Mean and standard deviation for each of the condition were presented at Appendix A (Table A1).

The results of a mixed Analysis of Variance (ANOVA) with 1 within-subject factor and 5 between-subject factors were presented in Appendix A (Table A2). The number of anchor items was considered as a repeated measure and thus treated as a within-subject factor. There was only 1 significant 3-way interaction with $\rho < 0.5$ and partial eta squared (η^2) > .01 (Figure 1). Means and standard deviations were reported in Table 8. All factors considered in this study appeared to have significant influence on the accuracy rate in selecting anchor items except the number of anchor items (1-item anchor vs. 4-item anchor). The percentage of DIF items, sample size and sample size ratio, magnitude of DIF, DIF patterns and the presence of impact all had significant effect on the accuracy rate.

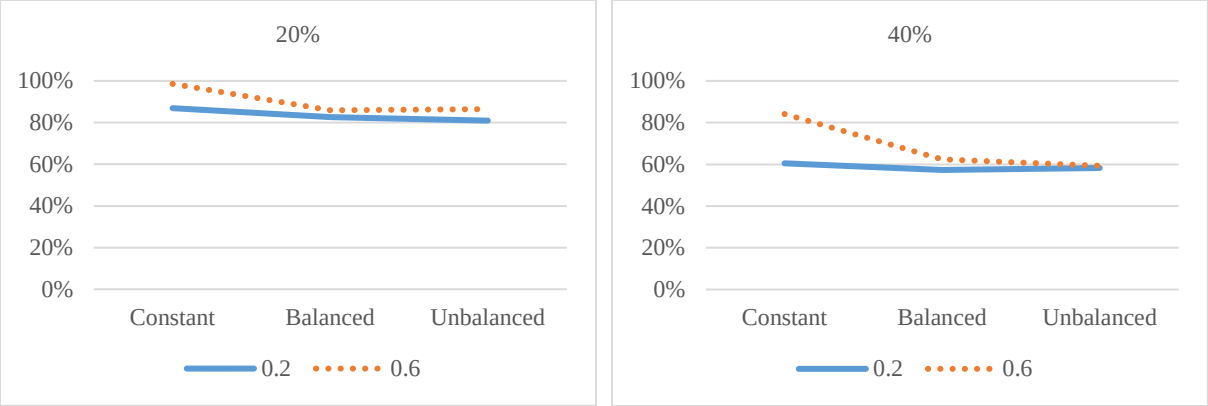


Figure 1 Three-way Interaction of Accracy among Pattern, Percentage, and Magnitude of DIF

Table 8 Mean and Standard Deviation of Accuracy for Pattern, Percentage and Magnitude of DIF

% of DIF	Magnitude	Constant Pattern		Balanced Pattern		Unbalanced Pattern	
		Mean	SD	Mean	SD	Mean	SD
20%	0.2	86.92	3.70	82.58	1.75	80.94	3.08
	0.6	98.58	1.61	85.91	1.82	86.47	4.94
		Mean	SD	Mean	SD	Mean	SD
40%	500	60.50	3.76	57.31	3.91	58.30	4.52
	1000	84.13	5.08	62.41	3.97	59.33	3.93

Figure 1 showed 3-way interaction among pattern, percentage, and magnitude of DIF. The left panel represented the mean accuracy rates of 3 DIF patterns for a magnitude size of 0.2 and 0.6 for the 20% DIF condition, while the right panel represented the mean accuracy rates of 3 DIF patterns for the magnitude size of 0.2 and 0.6 for the 40% DIF condition. The percentage of DIF significantly lowered the accuracy rates of selecting DIF-free items, this effect was more prominent when the magnitude of DIF was small. The combination of constant pattern and a large magnitude of DIF seemed to be more robust against the increasing of DIF-items.

Four 2-way interactions were significant in predicting the accuracy rates: DIF pattern by sample size and ratio, DIF pattern by percentage of DIF, DIF pattern by magnitude of DIF, and percentage of DIF by magnitude of DIF. Simple main effects were analyzed for these factors, results presented in Appendix Table A3-A6. Plots of interactions effects were presented in Appendix Figure A1-A4. For each DIF pattern, there were significant differences on the accuracy rates among sample size and ratio, percentage of DIF items, and magnitude of DIF, with only exception: for balanced DIF there was no significant differences found among different sample sizes. Different levels of percentage of DIF items had significant effect on the accuracy rates between 2 levels of magnitude of DIF. The means and standard deviations for the interactions were presented in Appendix Table A7-A10.

Percentage of DIF items (0%, 20%, and 40%) The percentage of DIF items appeared to be the most significant factor affecting the accuracy of selecting DIF-free items. With 20% DIF items, the mean accuracy rate is 86.90%, while with 40% of items exhibit DIF, the accuracy rate dropped down to range of 50% to 60%, with the mean accuracy rate dropped to 63.66% (Appendix Table A1). These results suggested that the HGLM method was having difficulties in selecting DIF-free items when many items were exhibiting DIF.

The percentage of DIF had significant interaction effects with the pattern of DIF and the magnitude of DIF. When the DIF pattern was constant, HGLM picked out DIF-free items about 93% of the time when there were 20% items with DIF. But the successful rate dropped to 72% when there were 40% items with DIF. When the pattern of DIF was not constant, the accuracy rates were about 84% and 59% (Appendix Table A8). When the magnitude of DIF was small, HGLM picked out DIF-free items about 83% of the time with 20% items exhibiting DIF, and 58% when there were 40% items with DIF (Appendix Table A9). These results suggested that even with conditions less favorable, the percentage of DIF items in a scale was crucial; if most of the items were DIF-free, HGLM method could succeed in selecting DIF-free items as anchors more than 80% of the time.

Magnitude of DIF The magnitude of DIF was also a very influential factor in the accuracy rate of selecting DIF-free items as anchors. When the magnitude of DIF equaled to .2, the mean accuracy rate was 71%, while with the magnitude of DIF equaled to .6, the mean accuracy rate was 79%. These results suggested that a smaller size of DIF would make detection of items with DIF more difficult.

The magnitude of DIF had significant interaction effects with the pattern of DIF and the percentage of DIF items. When the pattern was constant, a smaller magnitude of DIF of .2 would result in a 74% accuracy rate; under the condition of 40% DIF, the accuracy rate would future drop down to only about 60% (Appendix Table A3). But a larger magnitude of DIF of .6 would result in an accuracy rate over 90%. When the pattern of DIF was not constant, the effect of DIF magnitude was smaller, but larger DIF still produced higher accuracy rates (Appendix Table A10).

DIF patterns The pattern of DIF also had significant influence on the accuracy of selecting DIF-free items. When the pattern of DIF was constant, the mean accuracy rate was the highest at

82.53%, while the mean accuracy rate was lower for balanced pattern at 72.05% and high-unbalanced pattern at 71.26%. The constant pattern consistently out preformed the non-constant pattern, i.e. the balanced and un-balanced patterns, indicated that HGLM was sensitive to the pattern of DIF, and a non-constant pattern would reduce its performance in selecting DIF-free items as anchor. In addition, the constant DIF pattern seemed especially sensitive to the magnitude of DIF, as it showed a considerable drop of accuracy when the magnitude of DIF was small, from 91% to 74% (Appendix Table A9).

Sample size and sample size ratio The 4-level of sample size and sample size ratio also significantly affected the accuracy of selecting DIF-free items. With the least favorable condition a Rt400/F100, the accuracy rate was 73.45%, while with the most favorable condition at R500/F500, the accuracy rate was raised to 77.67%. These results were consistent with numerous literatures that suggested a larger sample size would produce more ideal analysis results. However, in this study, although the effect of sample size and sample size ratio was still statistically significant, the differences were less dramatic than some of the factors discussed.

A future analysis of contrast decomposed the sample size and ratio factor into sample size and sample size ratio. The results showed that although sample size had a significant influence on HGLM's accuracy rates, sample size ratio did not. These results indicated that unequal sample size ratio was less of a concern in using HGLM to select DIF-free items. However, even with a sample size of 1000, the accuracy rate was still less than 80%, suggested that even a sample size of 1000 might still not be sufficient in producing ideal accuracy rates.

Impact The presence of impact also had significant influence on accuracy. When there was no impact among the groups, the mean accuracy rate was 76.05% while with the presence of impact the mean accuracy rate was 74.51%. Chen et.al (2014) found that with dichotomous responses

impact did not make a significant difference, whereas, this study found that with polytomous data, impact did have some influence on the accuracy of selecting the DIF-free items, although not as influential as some other factors in the study.

Number of anchor items The different numbers of anchor items produced very similar results. Chen et.al (2014) found that when fewer DIF-free items need to be selected, the accuracy rates increased; however, this study found that the number of anchor items needed to be selected did not make a significant difference in the accuracy rate.

In general, the HGLM method performed well when, the percentage of DIF items is small (<20%), the magnitude of DIF was large, the pattern of DIF was constant, and the sample size was large (>1000). The percentage of DIF items appeared to be very influential; when the percentage is 40%, HGLM was often unable to correctly identify DIF free items; when combined with small size of DIF, HGLM was only accurate for about half the times. The influential factors were consistent with the findings by Chen et.al (2014), except that, Chen et al. (2014) showed that with dichotomous item responses, the accuracy of selecting DIF-free items was high for HGLM with iterative DFTD strategy, ranging from 90% to 100%. Table 6 and 7 showed that with polytomous responses, the accuracies were generally lower.

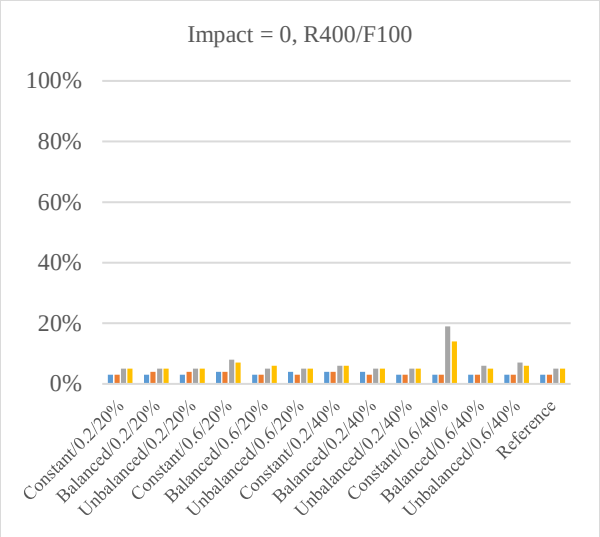
4.2 Results of Study 2: Type I Error

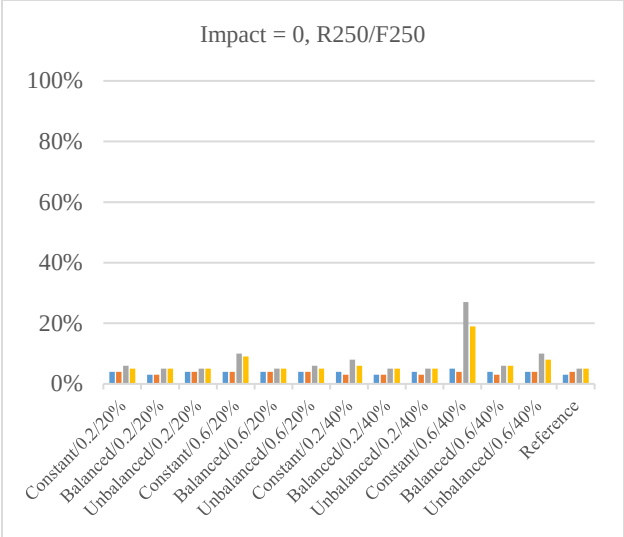
This section presents the results for the second research question: what is the Type I error rate for DIF detection using HGLM, and how does it compare to using GMH and logistic regression? In order to answer this question, datasets were generated for each of the 312 simulation conditions. Each item from each dataset was then fitted with a HGLM model using a 1-item anchor

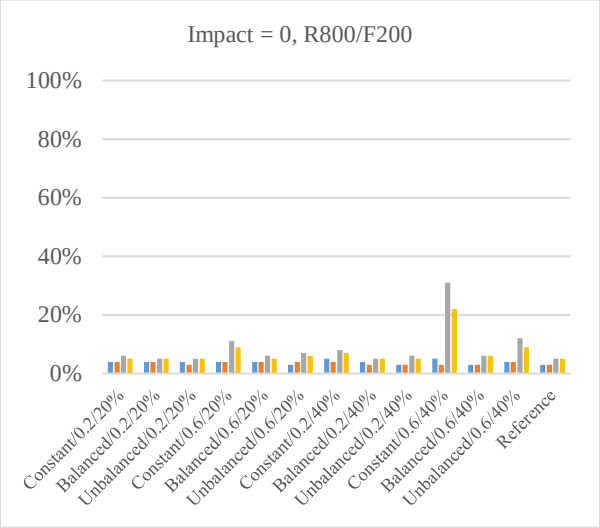
or a 4-item anchor. The estimation method was the maximum likelihood method with Laplace approximation. The same dataset was then fitted with GMH and logistic regression methods to identify DIF items. The Type I error and power rates from the 4 methods were then estimated. Unlike study 1 with 100 replications, for study 2 500 replications were performed in order to increase estimation accuracy of Type I error and power rates. Estimation of all replications converged.

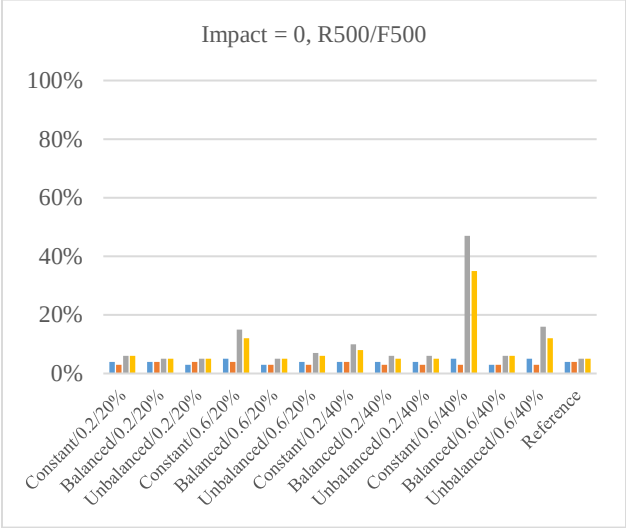
4.2.1 Results of Type I error rates

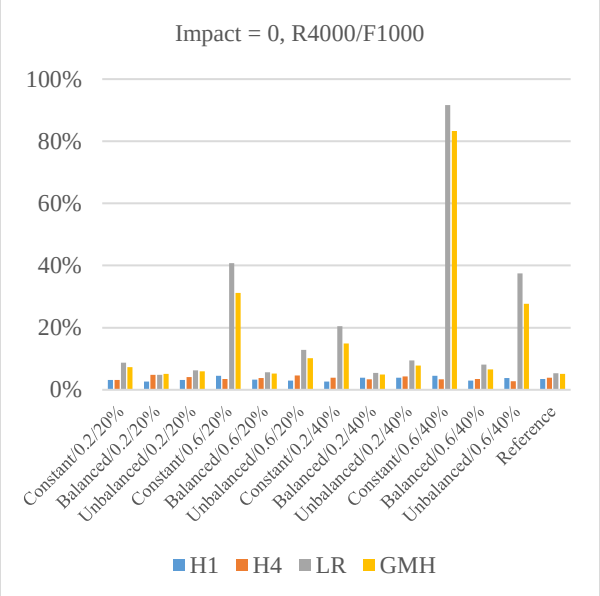
Type I error rate was calculated for each DIF-free item by computing the percentage of times the DIF-free item was identified as a DIF item over the total number of replications. The mean Type I error rates for all the DIF-free items for all conditions were presented in Figure 2. The first 4 columns represented the mean Type I error rate for the HGLM with a 1-item anchor, HGLM with a 4-item anchor, the logistic regression, and the GMH methods, respectively, for constant DIF pattern with a magnitude of 0.2 and a total of 20% DIF items. The full results of mean Type I error rate for all the conditions were presented in Appendix B (Table B11-B12). The mean and standard deviation for each condition were presented in Appendix B (Table B13).

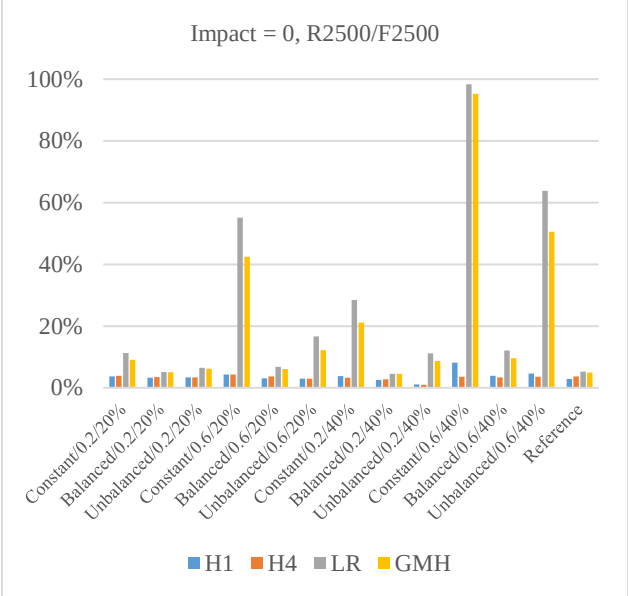

Impact = 0, R400/F100
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%
Reference


Impact = 0, R250/F250
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%
Reference


Impact = 0, R800/F200
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%
Reference


Impact = 0, R500/F500
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%
Reference


Impact = 0, R4000/F1000
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%
Reference
H1
H4
LR
GMH


Impact = 0, R2500/F2500
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%
Reference
H1
H4
LR
GMH

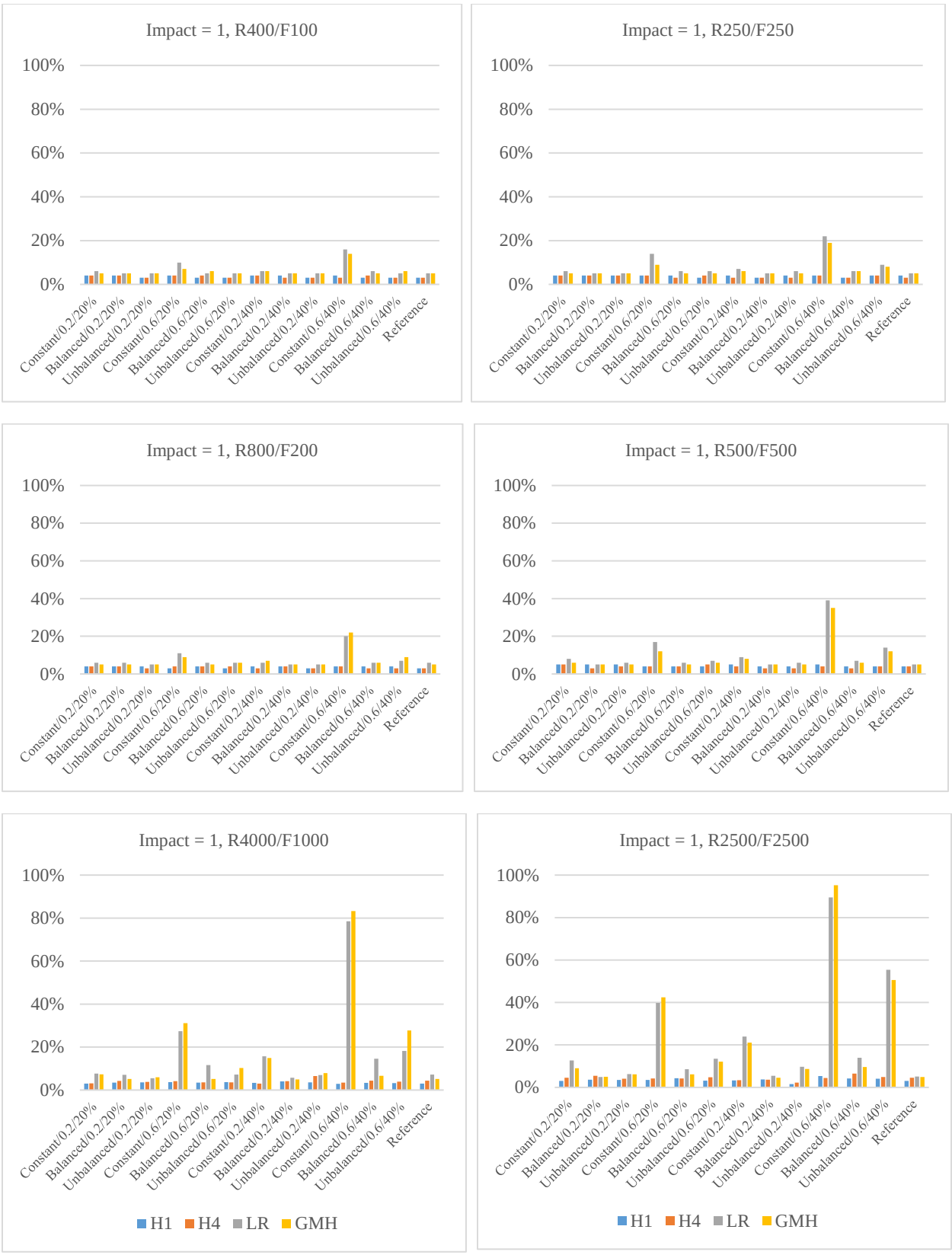


Figure 2 Type I error Rates for All Conditions

The Type I error rates of the two HGLM methods were consistently lower than that of the logistic regression and the GMH methods. The two HGLM methods exhibited decent control across the conditions, with an average lower than 4%. The number of anchor item did not seem to produce a significant difference. Whereas the control of logistic regression and the GMH fluctuated among the conditions, with an average over 10%. When the sample size was not large, and there was only a small amount of DIF present, the two behaved relatively better; however, when conditions turned less favorable, the logistic regression and the GMH methods could lose control entirely.

4.2.2 ANOVA results of Type I error

There were in total 7 conditions being manipulated for the simulation study, the presence of impact (0, 1), the sample size (500, 1000, 5000), the sample size ratio (1:1, 4:1), the percentage of DIF items (20%, 40%), the magnitude of DIF (.2, .6), the pattern of DIF (constant, balanced, high unbalanced), and the 4 different DIF detection methods (HGLM with a 1-item anchor, HGLM with a 4-item anchor, the logistic regression, the GMH).

A 7-way mixed ANOVA was conducted and results were presented in Appendix B (Table B14). Most of the higher order interactions were significant at $\rho < 0.5$; however, since higher order interaction effects could be difficult and confusing to interpret, only 3-way interactions were considered for the final ANOVA. In addition, since this study focus on the HGLM methods performance comparing to the other methods, only 3-way interactions involving the 4 DIF detection methods were included. For the HGLM methods, different level of impact and sample size ratio produced similar Type I error rates, about 4%, for logistic regression and GMH, the error were similar at about 13%. In addition, none the two factors' main effect analysis was significant.

Thereby only 3-way interactions among 4 DIF detection methods, sample size, magnitude, pattern, and the percentage of DIF with a significant $\rho < 0.5$ and partial $\eta^2 > 0.01$ were included. Figure 3-8 showed results from the 6 3-way interactions. The means and standard deviations were reported in Table 15-18.

Figure 3 showed 3-way interaction among the 4 methods, sample size and DIF patterns. The left panel represented the Type I error rates of the 4 methods with 3 different sample sizes for the constant DIF pattern, while the middle and right panels represented the balanced and unbalanced DIF patterns, respectively.

The two HGLM methods showed consistently good control of Type I error at around 4% and behaved similarly. The logistic regression and GMH methods showed good control as well when the total sample size was small (500). However, when the sample size increased, Type I error rate increased significantly, especially for the constant pattern (Table 9). When DIF pattern was constant, the Type I error rate was inflated even when the sample size was small, around 10%. When the sample size increased to 5000, Type I error rates were around 40% for the logistic regression and the GMH methods. When the pattern was balanced, meaning DIF was in different directions within an item's response categories, there was no inflation of Type I error rates even when the sample sizes were large, suggested that the LR and GMH methods were sensitive to the sample size only when the absolute value of DIF within the items were substantial.

Figure 4 showed the 3-way interaction among the 4 methods, sample size and the percentage of DIF items. The left panel represented the Type I error rates of the 4 methods with 3 different sample sizes for the 20% DIF items condition, while right panel represented the 40% DIF items condition.

The two HGLM methods again behaved rather similarly and showed good control of Type I error rate overall. When 20% items exhibit DIF, the logistic regression and GMH were almost as good as the two HGLM methods at around 6%, with a slight inflation to over 10% when the sample size increased to 5000. However, when the percentage of DIF items increased to 40%, the increase of sample size resulted in a significant inflation of Type I error rate for the two methods. When the sample size was 5000, the Type I error rates of both methods increased to 30% for the logistic regression method and 28% for the GMH method, suggested that both were sensitive to the percentage of DIF items when the sample size was large.

Figure 5 showed the 3-way interaction among methods, sample size and the magnitude of DIF. The left panel represented the Type I error rates of the 4 methods for the 3 different sample sizes when the magnitude of DIF was 0.2, while the right panel showed the results for when the magnitude of DIF was 0.6.

Figure 5 showed similar findings as presented in Figure 4. The HGLM methods exhibited good control overall, while the logistic regression and the GMH's control of Type I error decreased when the sample size increased to 5000, and especially so for the magnitude of 0.6.

The above three analysis showed that the logistic regression and the GMH methods were sensitive to the non-balanced DIF pattern, larger sample size, magnitude and higher percentage of DIF items. When the sample size was 5000, magnitude was 0.6 with 40% of items exhibit constant DIF, the two methods showed a complete lack of control over Type I error that were around 90% (Appendix Table B11, B12).

Figure 6 showed the 3-way interaction among methods, DIF patterns and the percentage of DIF. The left panel represented the Type I error rates of the 4 methods for the 3 different DIF

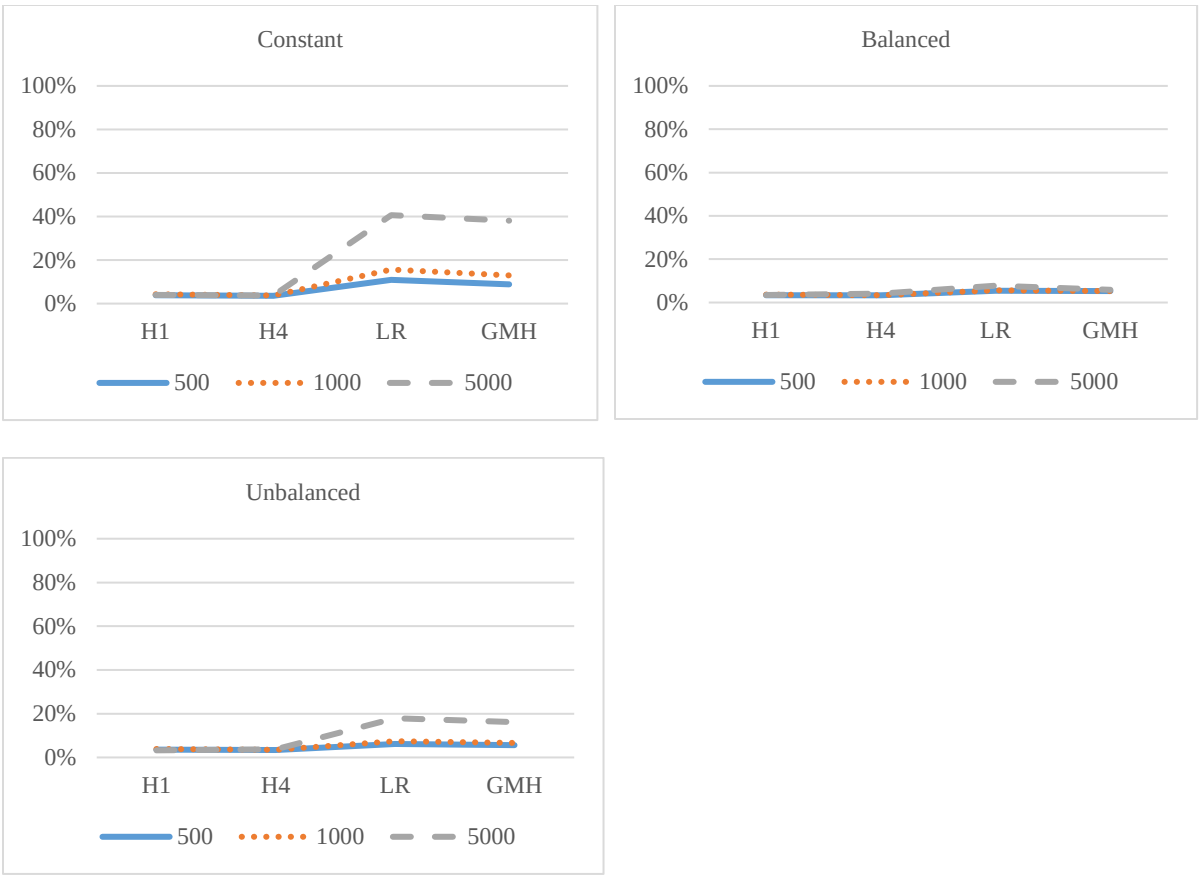
patterns with 20% of DIF items, while the right panel showed the results for the condition with 40% DIF items.

Figure 7 showed the 3-way interaction among methods, DIF patterns and the magnitude of DIF. The left panel represented the Type I error rates of the 4 methods for the 3 different DIF patterns with a magnitude of 0.2, while the right panel showed the results for the condition with the magnitude of 0.6.

Figure 6 and Figure 7 produced similar results. The two HGLM methods, HGLM with a 1-item anchor and HGLM with a 4-item anchor, behaved similarly and showed good control across conditions. The logistic regression and GMH methods showed a significant inflation of Type I error when detecting constant DIF with a larger magnitude or when there were a larger number of items exhibiting DIF; under such conditions about a third of the times the two methods would flag a DIF-free item as exhibiting DIF.

Figure 8 showed the 3-way interaction among methods, DIF patterns and the percentage of DIF. The left panel showed the results of 4 methods for 2 different DIF magnitudes when there were 20% items with DIF. The right panel showed the results with 40% DIF.

Consistent with previous analysis, the two HGLM method showed good control under these conditions. With 20% DIF, there was a slight elevation of Type I error rate when the magnitude was larger for the logistic regression and the GMH methods. With 40% DIF, the Type I error rates for the logistic regression and the GMH were 26% and 23%, respectively (Table 14), meaning there was a quarter of the items flagged by the two methods were in fact DIF-free items.



Note: H1=HGLM with 1-item anchor, H4= HGLM with 4-item anchor, LR=polytomous logistic regression, GMH = Generalized Mantel-Haenszel.

Figure 3 Three-Way Interaction of Type I Error among Methods, Sample Size and DIF Pattern

Table 9 Mean and Standard Deviation of Type I Error Rates for Method, Sample Size and DIF Patterns

Pattern	Sample Size	HGLM1		HGLM4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Constant	500	0.04	0.004	0.04	0.003	0.11	0.07	0.09	0.05
	1000	0.04	0.01	0.04	0.005	0.16	0.13	0.13	0.10
	5000	0.04	0.01	0.04	0.005	0.41	0.32	0.38	0.33
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Balanced	500	0.03	0.003	0.03	0.003	0.05	0.00	0.05	0.003
	1000	0.04	0.004	0.03	0.005	0.06	0.01	0.05	0.005
	5000	0.03	0.01	0.04	0.01	0.08	0.03	0.06	0.02
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Unbalanced	500	0.04	0.004	0.03	0.003	0.06	0.02	0.06	0.01
	1000	0.04	0.01	0.04	0.004	0.07	0.03	0.07	0.03
	5000	0.03	0.01	0.04	0.01	0.18	0.18	0.16	0.15

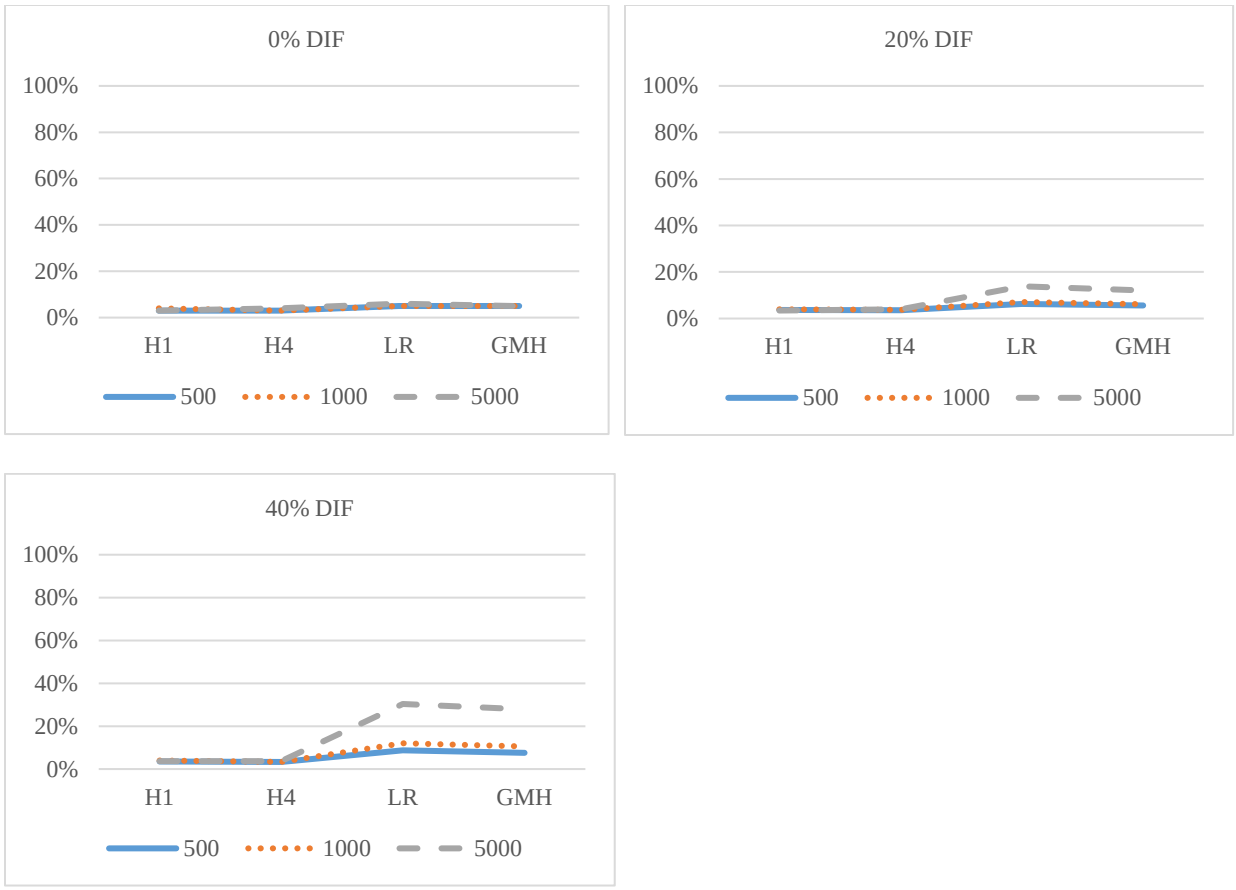


Figure 4 Three-Way Interaction of Type I Error among Methods, Sample Size and Percentage of DIF

Table 10 Mean and Standard Deviation of Type I Error Rates for Method, Sample Size and % of DIF

% of DIF	Sample Size	HGLM 1		HGLM 4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0%	500	0.03	0.003	0.03	0.004	0.05	0.003	0.05	0.001
	1000	0.04	0.005	0.03	0.005	0.05	0.002	0.05	0.001
	5000	0.03	0.003	0.04	0.004	0.06	0.01	0.05	0.001
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
20%	500	0.04	0.004	0.04	0.003	0.06	0.02	0.06	0.01
	1000	0.04	0.005	0.04	0.004	0.07	0.03	0.06	0.02
	5000	0.03	0.005	0.04	0.01	0.14	0.13	0.12	0.12
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
40%	500	0.04	0.005	0.03	0.003	0.09	0.06	0.08	0.04
	1000	0.04	0.01	0.03	0.005	0.12	0.11	0.11	0.09
	5000	0.04	0.01	0.04	0.01	0.30	0.31	0.28	0.31

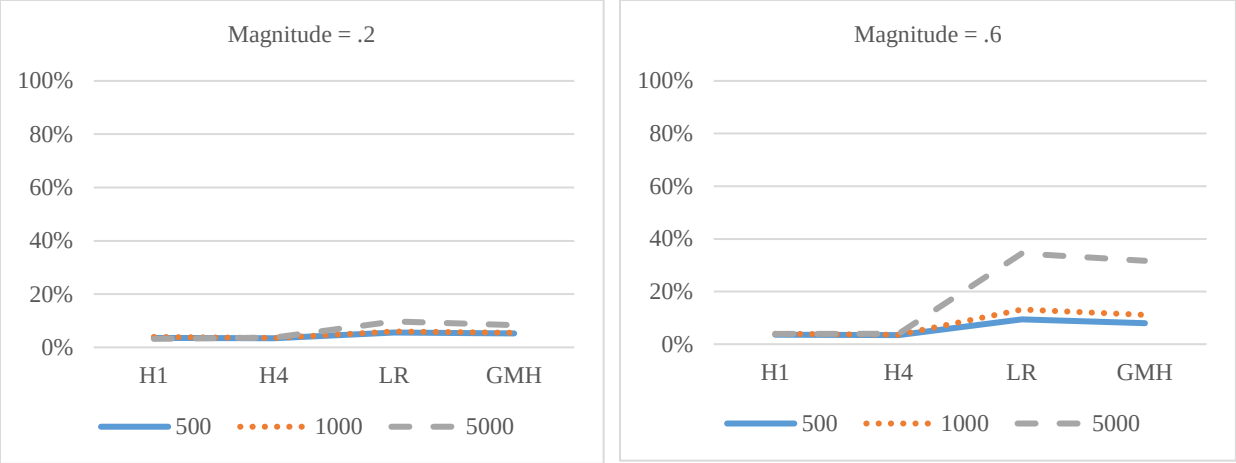


Figure 5 Three-Way Interaction of Type I Error among Methods, Sample Size and Magnitude of DIF

Table 11 Mean and Standard Deviation of Type I Error Rates for Method, Sample Size and Manitude of DIF

% of DIF	Sample Size	HGLM 1		HGLM4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.2	500	0.04	0.004	0.03	0.003	0.06	0.01	0.05	0.004
	1000	0.04	0.005	0.04	0.005	0.06	0.01	0.05	0.01
	5000	0.03	0.01	0.04	0.01	0.10	0.06	0.08	0.05
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.6	500	0.04	0.004	0.03	0.003	0.09	0.06	0.08	0.04
	1000	0.04	0.01	0.04	0.005	0.13	0.11	0.11	0.09
	5000	0.04	0.01	0.04	0.01	0.34	0.30	0.32	0.30

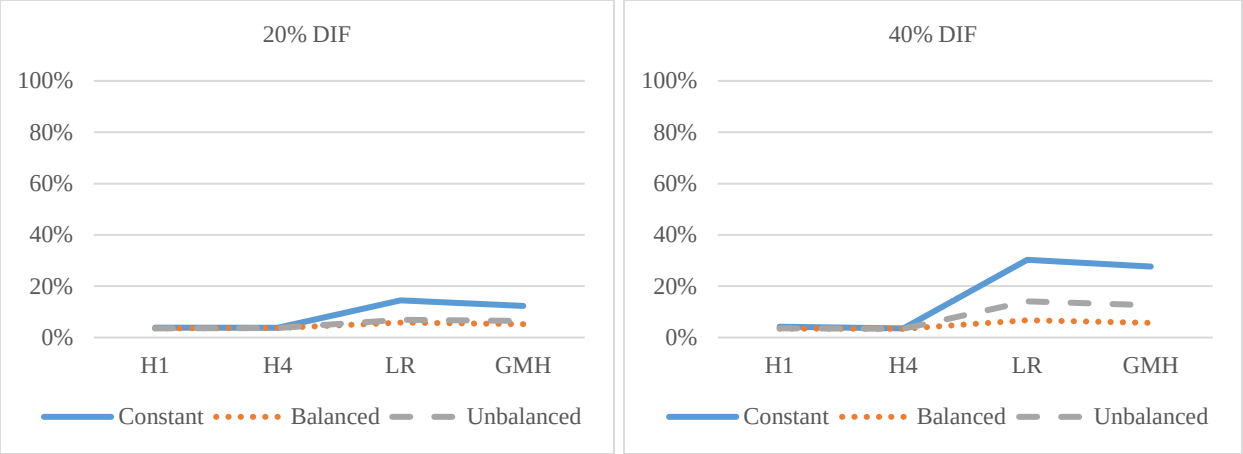


Figure 6 Three-Way Interaction of Type I Error among Methods, DIF pattern and Percentage of DIF

Table 12 Mean and Standard Deviation of Type I Error Rates for Method, DIF Pattern and Percentage of DIF

% of DIF	Sample Size	HGLM 1		HGLM 4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
20%	Constant	0.04	0.01	0.04	0.004	0.14	0.13	0.12	0.12
	Balanced	0.04	0.005	0.04	0.01	0.06	0.02	0.05	0.003
	Unbalanced	0.04	0.004	0.04	0.00	0.07	0.03	0.06	0.02
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
40%	Constant	0.04	0.01	0.04	0.004	0.30	0.29	0.28	0.29
	Balanced	0.03	0.004	0.03	0.01	0.07	0.03	0.06	0.01
	Unbalanced	0.04	0.01	0.03	0.01	0.14	0.16	0.13	0.13

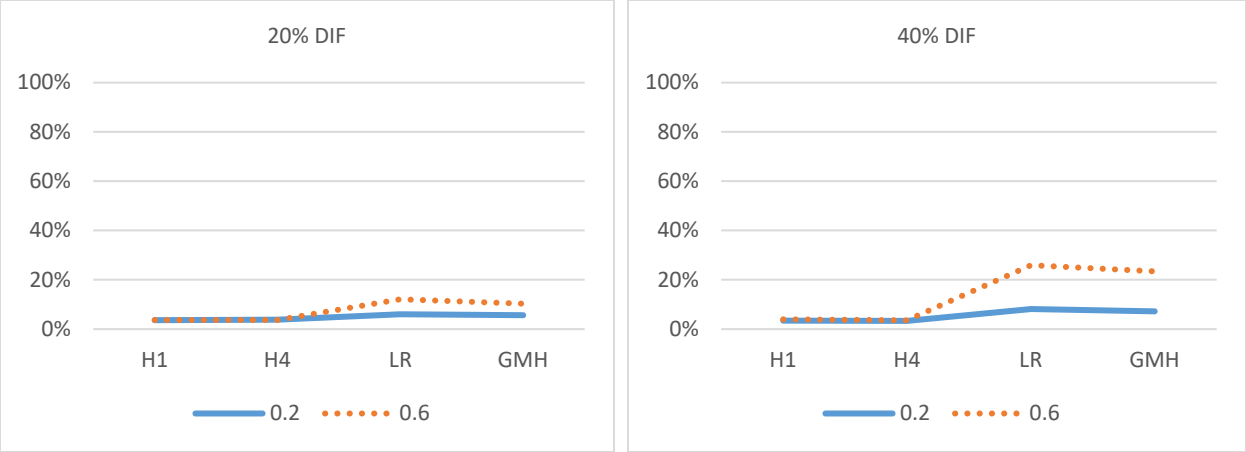


Figure 7 Three-Way Interaction of Type I Error among Methods, Magnitude and Percentage of DIF

Table 13 Mean and Standard Deviation of Type I Error Rates for Method, Magnitude and Percentage of DIF

% of DIF	Sample Size	HGLM 1		HGLM 4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
20%	0.2	0.04	0.005	0.04	0.01	0.06	0.02	0.06	0.01
	0.6	0.04	0.01	0.04	0.00	0.12	0.11	0.10	0.10
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
40%	0.2	0.03	0.01	0.03	0.01	0.08	0.06	0.07	0.04
	0.6	0.04	0.01	0.04	0.01	0.26	0.27	0.23	0.27

Figure 8 Three-Way Interaction of Type I Error among Methods, DIF Pattern and Magnitude of DIF

Table 14 Mean and Standard Deviation of Type I Error Rates for Method, DIF Pattern and Magnitude of DIF

% of DIF	Sample Size	HGLM 1		HGLM 4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.2	Constant	0.04	0.01	0.04	0.005	0.10	0.06	0.08	0.05
	Balanced	0.04	0.004	0.04	0.01	0.05	0.005	0.05	0.002
	Unbalanced	0.03	0.01	0.03	0.01	0.06	0.02	0.06	0.01
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.6	Constant	0.04	0.01	0.04	0.004	0.35	0.28	0.32	0.29
	Balanced	0.03	0.004	0.04	0.01	0.07	0.03	0.06	0.01
	Unbalanced	0.04	0.01	0.04	0.01	0.15	0.15	0.13	0.13

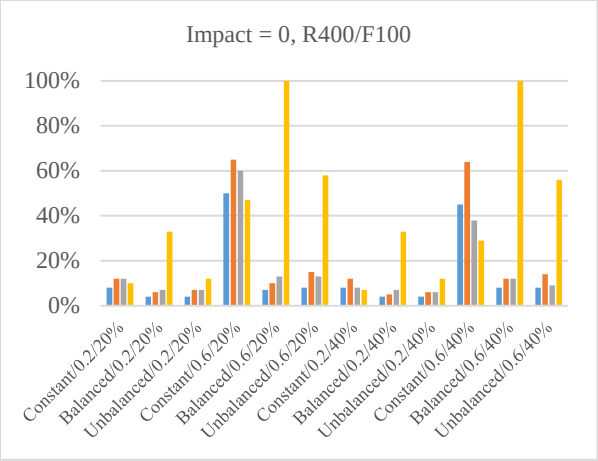
4.3 Results of Study 2: Power

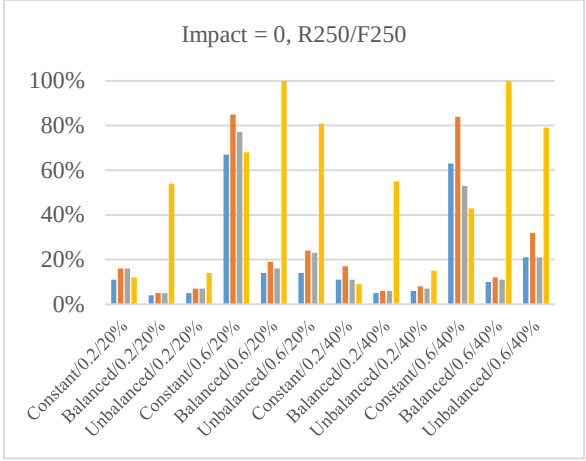
This section presents the results for the third research question: what is the statistical power for DIF detection using HGLM, and how does it compare to using GMH and logistic regression?

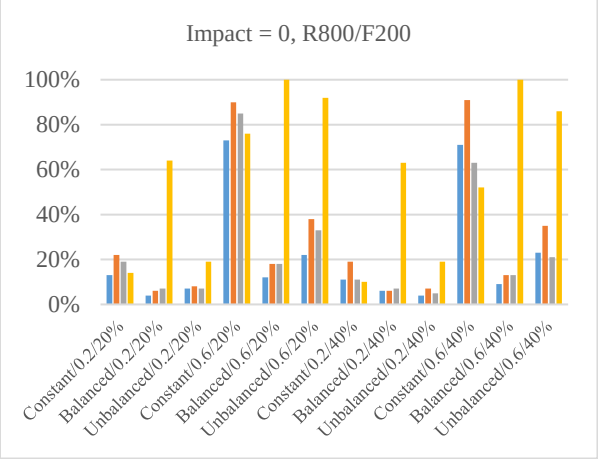
4.3.1 Results of Power

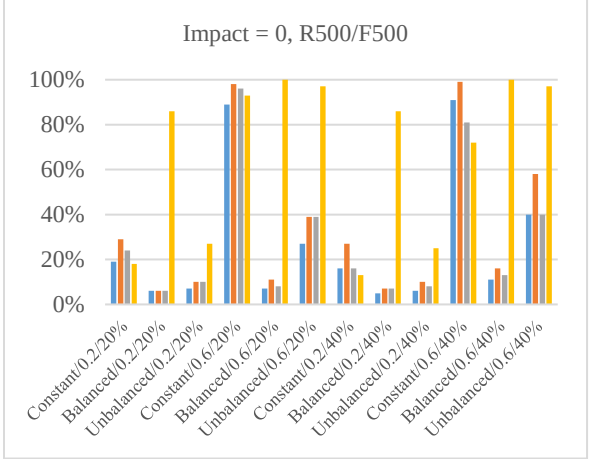
After the generated dataset were examined by the 4 DIF detection methods for DIF, statistical power was calculated along with the Type I error rate for each DIF item by computing the percentage of times the DIF item is correctly identified over the total number of replications. The mean power rates for all the DIF items for all conditions were presented in Figure 9. The average power rate for all the items with DIF were presented in Appendix B (Table B15, B16). The mean and standard deviation for each condition were presented in Appendix B (Table B17).

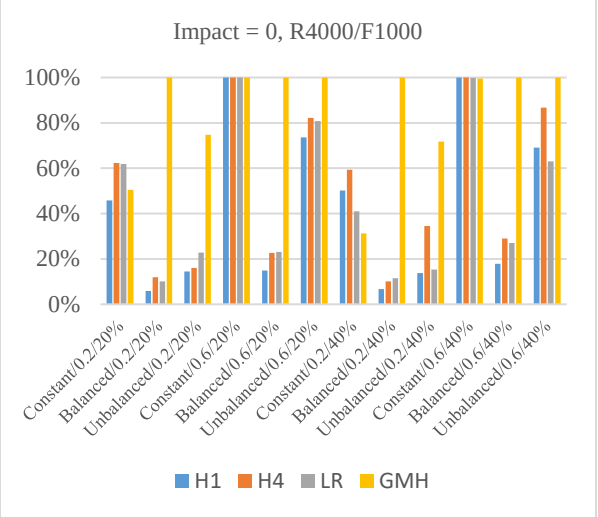
The mean power for HGLM with 1-item anchor was 28%, while for HGLM with 4-item anchor 36%. The two HGLM showed good control over Type I error rates consistently. However, with power the two methods behaved vastly different under various conditions. HGLM with 4-item anchor were significantly more powerful than the HGLM with 1-item anchor, but was less powerful than the GMH method, which showed a 66% mean power. Under various conditions, the GMH significantly outperformed the other three in terms of power.


Impact = 0, R400/F100
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%


Impact = 0, R250/F250
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%


Impact = 0, R800/F200
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%


Impact = 0, R500/F500
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%


Impact = 0, R4000/F1000
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%
H1
H4
LR
GMH

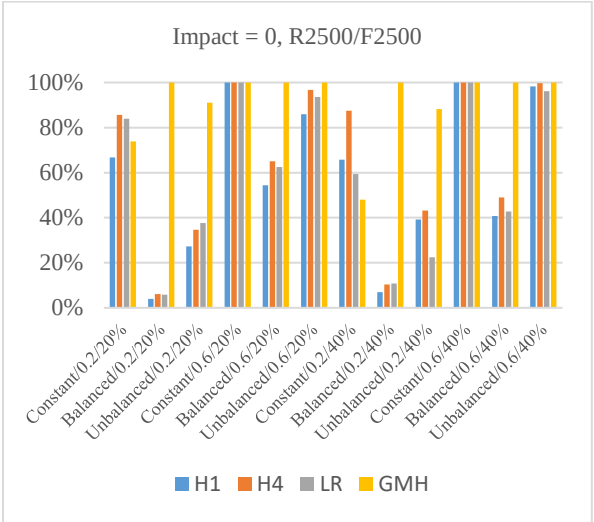

Impact = 0, R2500/F2500
100%
80%
60%
40%
20%
0%
Constant/0.2/20%
Balanced/0.2/20%
Unbalanced/0.2/20%
Constant/0.6/20%
Balanced/0.6/20%
Unbalanced/0.6/20%
Constant/0.2/40%
Balanced/0.2/40%
Unbalanced/0.2/40%
Constant/0.6/40%
Balanced/0.6/40%
Unbalanced/0.6/40%
H1
H4
LR
GMH

Figure 9 Power for All Conditions

4.3.2 ANOVA results of Power

A 7-way mixed ANOVA were conducted and results presented in Appendix B (Table B17). None of the high order interactions was significant, so only 3-way interactions were included in the ANOVA model. Only the significant interactions with $\rho < 0.5$ and partial $\eta^2 > 0.01$ were presented. The results for the 3-way interactions were presented in Figure 10 -16, and the mean and standard deviation presented in Table 15-18.

Figure 10 showed the interaction among the 4 DIF-detection methods, the 3 levels of sample size, and 3 different DIF patterns. For all the patterns and sample sizes, the HGLM with 1-item anchor behaved similarly to the logistic regression method. The constant pattern produced much better results for the two HGLM methods, with 4-item anchor outperformed the 1-item anchor. When sample size was as large as 5000, the power rate for the two were 79% and 88%, respectively (Table 15). When the sample size was small, the power rates were much lower, but HGLM with 4-item anchor still outperformed the others.

However, for the balanced and unbalanced patterns, the two HGLM methods were less powerful than the GMH method, which, although were having difficulties dealing with the constant pattern, showed a much significant improvement. The two HGLM methods' power rates were typically less than 10% for the smaller sample sizes, especially for the balanced pattern. Even when the sample size was large, HGLM with 4-item anchor was still only about successful half the time at best for detecting items with DIF, while the GMH could correctly identify items with DIF over 90% of the times with the unbalanced pattern.

Figure 11 showed the 3-way interaction among the 4 methods, sample size and 2 levels of DIF magnitude. When the sample size was small, all the methods were having difficulties dealing with the smaller magnitude of DIF, with the GMH seemed to perform better than the others. The

two HGLM methods were both poorly behaved with around 10% of power, although the HGLM with 4-item anchor was slightly better (Table 16). When the sample size was large at 5000, the performance improved, with HGLM with 4-item anchor given a 41% power. The GMH, however, had a power of 77% with a sample size of 5000 and a magnitude of 0.2.

When the magnitude increased to 0.6, the performance for all the methods improved, especially when sample size was large. HGLM with 1-item anchor was behaving similarly to the logistic regression, less powerful than the HGLM with 4-item anchor, which was still not as powerful as the GMH. When the sample size and the magnitude were both large, the GMH could reach near 100% of power.

Figure 12 showed the 3-way interaction of 4 methods, 3 DIF patterns, and 2 level of percentage of DIF items. The 2 HGLM methods behaved similarly, while HGLM with 4-item anchor slightly better. For the constant pattern and balanced pattern, the percentage of DIF produced little differences between the two HGLM methods. The constant pattern produced a power rate of 52% for HGLM with 1-item anchor and about 65% for HGLM with 4-item anchor, however, the unbalanced pattern reduced the power rates to 11% and 16%, respectively (Table 17). With the unbalanced pattern, the increased percentage of DIF produced higher power rates for the two HGLM methods, this somewhat surprising results suggested that a larger the number of items with DIF would help the HGLM methods to identify DIF items with unbalanced patterns more easily, although the increased the power rates were still less than a third times for even the HGLM with 4-item anchor.

The logistic regression and GMH methods, unlike the two HGLM methods, showed slight decrease in power when the percentage of DIF increased from 20% to 40%, which was consistent with the findings of previous studies. For the GMH, the constant pattern reduced its performance,

but with non-constant DIF patterns, the GMH was more powerful than the other three, especially with the balanced pattern. This suggested that the GMH was more sensitive to more complicated DIF patterns, although less powerful with constant pattern.

Figure 13 showed the 3-way interaction of 4 methods, 3 DIF patterns, and 2 level of magnitude of DIF. The magnitude of DIF made a significant difference for all the methods, especially with the constant pattern, the HGLM with 4-item anchor showed an 89% power with a constant DIF size of 0.6, with HGLM with 1-item anchor a 78% power. When the magnitude was small and DIF pattern non-constant, the power rates were close to 10% for both HGLM methods. The GMH showed a similar behavior of not handling the constant DIF pattern very well, but outperformed the others when the pattern was non-constant.

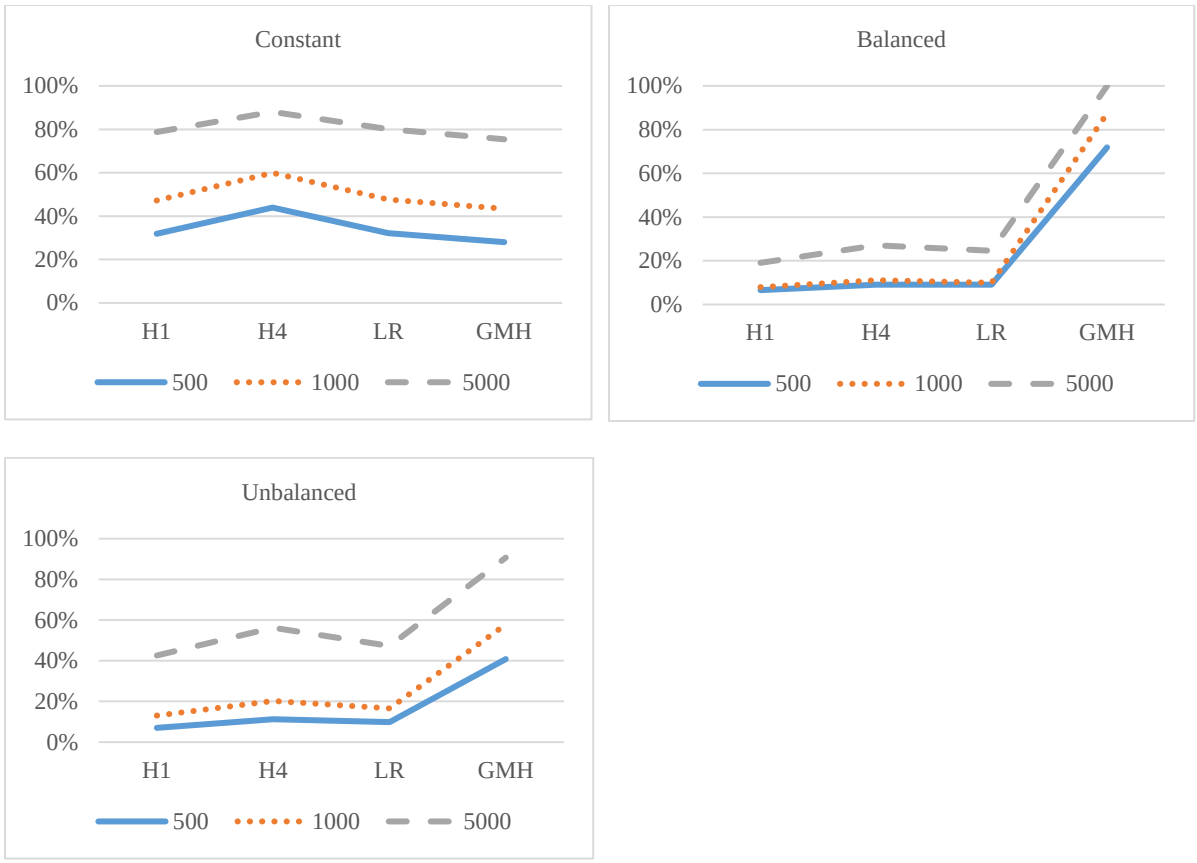


Figure 10 Three-way Interaction of Power oor the Method, Sample Size, and DIF Pattern

Table 15 Mean and Standard Deviation of Power for Method, Sample Size and DIF Patterns

Pattern	Sample Size	HGLM1		HGLM4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Constant	500	0.32	0.24	0.44	0.32	0.32	0.24	0.28	0.22
	1000	0.47	0.34	0.60	0.35	0.48	0.33	0.43	0.32
	5000	0.79	0.23	0.88	0.16	0.80	0.23	0.75	0.28
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Balanced	500	0.07	0.03	0.09	0.03	0.09	0.03	0.72	0.30
	1000	0.08	0.02	0.11	0.04	0.10	0.03	0.87	0.15
	5000	0.19	0.14	0.27	0.16	0.25	0.16	1.00	0.00
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Unbalanced	500	0.07	0.05	0.11	0.07	0.10	0.06	0.41	0.30
	1000	0.13	0.10	0.20	0.16	0.17	0.12	0.58	0.37
	5000	0.43	0.30	0.56	0.30	0.47	0.29	0.91	0.11

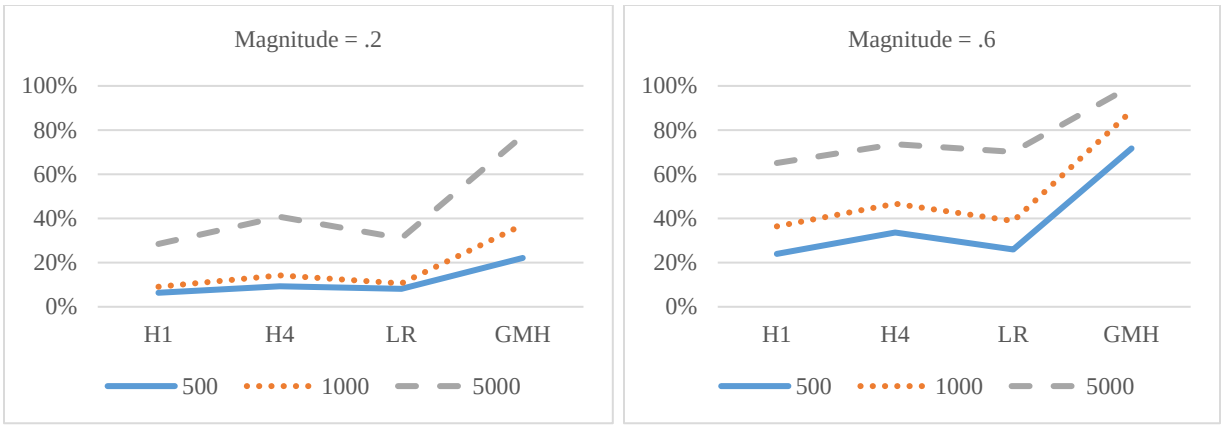


Figure 11 Three-way Interaction of Power for the Method, Sample Size, and Magnitude of DIF

Table 16 Mean and Standard Deviation of Power for Method, Sample Size and Manitude of DIF

Magnitude	Sample Size	HGLM1		HGLM4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.2	500	0.06	0.03	0.09	0.04	0.08	0.03	0.22	0.17
	1000	0.09	0.05	0.14	0.10	0.11	0.06	0.37	0.28
	5000	0.28	0.23	0.41	0.28	0.31	0.24	0.77	0.23
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.6	500	0.24	0.23	0.34	0.30	0.26	0.22	0.72	0.25
	1000	0.36	0.32	0.47	0.35	0.39	0.30	0.89	0.15
	5000	0.65	0.33	0.74	0.29	0.70	0.28	1.00	0.00

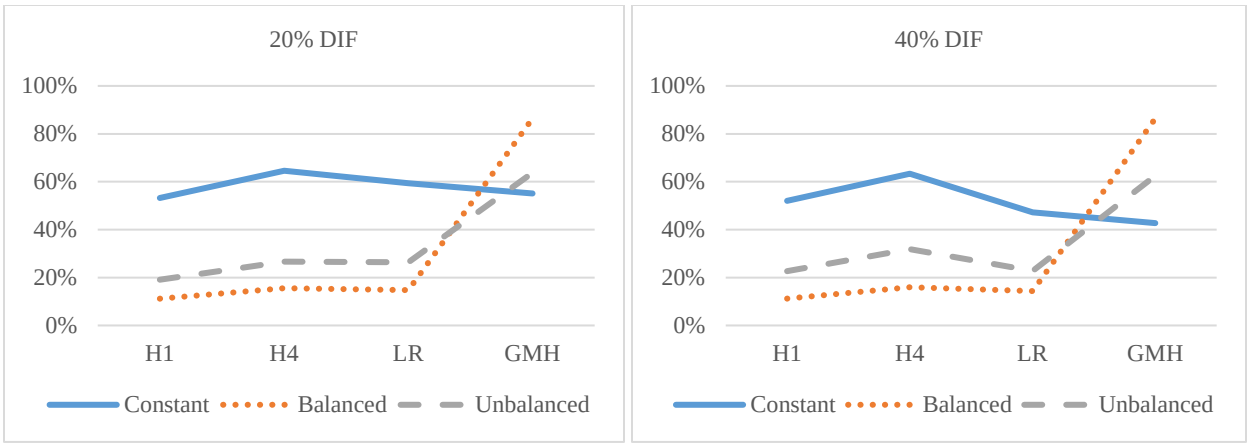


Figure 12 Three-Way Interaction Of Power For The Method, DIF Pattern, and Percentage Of DIF

Table 17 Mean and Standard Deviation of Power for Method, DIF Pattern and Percentage of DIF

% of DIF	Pattern	HGLM 1		HGLM 4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
20%	Constant	0.53	0.34	0.65	0.34	0.59	0.34	0.55	0.34
	Balanced	0.11	0.11	0.16	0.13	0.15	0.14	0.86	0.23
	Unbalanced	0.19	0.22	0.27	0.26	0.26	0.25	0.64	0.35
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
40%	Constant	0.52	0.34	0.63	0.34	0.47	0.32	0.43	0.33
	Balanced	0.11	0.09	0.16	0.12	0.14	0.10	0.86	0.23
	Unbalanced	0.23	0.26	0.32	0.29	0.23	0.24	0.63	0.35

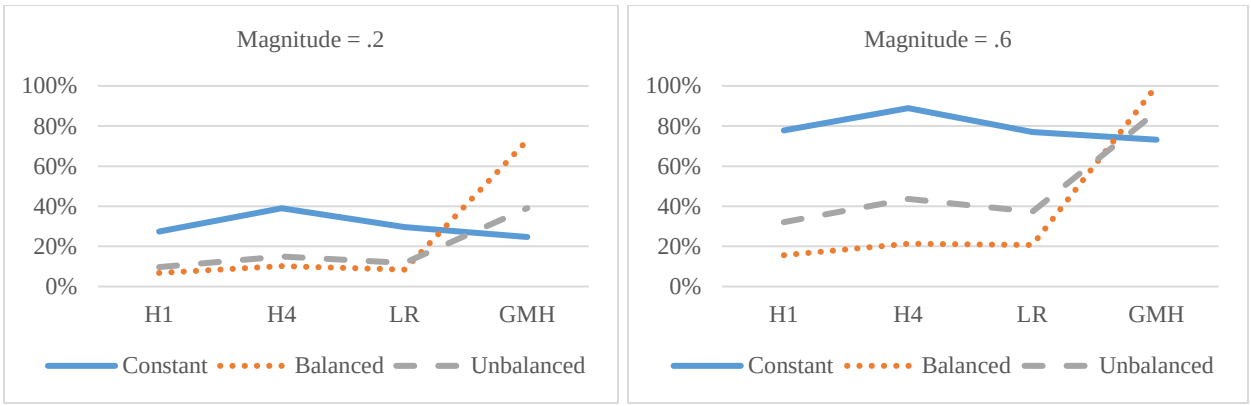


Figure 13 Three-way Interaction of Power for the Method, DIF Pattern, and Magnitude of DIF

Table 18 Mean and Standard Deviation of Power for Method, DIF Pattern, and Magnitude of DIF

Magnitude	Pattern	HGLM 1		HGLM 4		LR		GMH	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.2	Constant	0.27	0.23	0.39	0.29	0.30	0.24	0.25	0.21
	Balanced	0.07	0.03	0.10	0.06	0.08	0.03	0.73	0.25
	Unbalanced	0.10	0.09	0.15	0.12	0.12	0.09	0.39	0.31
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.6	Constant	0.78	0.21	0.89	0.14	0.77	0.23	0.73	0.25
	Balanced	0.16	0.12	0.21	0.15	0.21	0.14	1.00	0.00
	Unbalanced	0.32	0.29	0.44	0.31	0.37	0.29	0.87	0.16

4.4 Summary of the Results for Study 2

To summarize, HGLM with 1- or 4-item anchor performed similarly good with Type I error control, but HGLM with 4-item anchor had better power than HGLM with 1-item anchor. The polytomous logistic regression and the GMH methods showed similar performance on Type I error, but the GMH was much more powerful. since type I error and power are closely related, it is necessary to examine them together. And since HGLM with 4-item anchor outperformed HGLM with 1-item anchor, while GMH outperformed the logistic regression, HGLM with 4-item anchor and GMH results were shown. Figure 14-17 showed the comparison of Type I error and power for HGLM with 4-item anchor and GMH.

Figure 14 presented the Type I error and power rates for HGLM with 4-item anchor and GMH for the 3 levels of sample size and 3 different DIF patterns. For the constant pattern, the power steadily increased along with sample size for the HGLM, while the Type I error was consistently under control. The GMH was less powerful with significant inflation of Type I error when sample size was large. For the balanced pattern, the GMH was more powerful than the HGLM, with controlled Type I error rates. For the unbalanced pattern, the GMH was more powerful; when the sample size was small it showed good error control as well. But when the sample size was large as 5000, the error rate was also elevated. These results seemed to confirm previous findings that the GMH, as a nonparametric method, might be more suited for small sample conditions.

Figure 15 presented the Type I error and power rates for HGLM with 4-item anchor and GMH for the 3 levels of sample size and 2 levels of magnitude of DIF. Smaller magnitude of DIF reduced the power for both methods, while the GMH seemed to be more sensitive. When magnitude of DIF increased, power increased as well, but for the GMH, Type I error rates was

also inflated. These results seemed to suggest that GMH was better suited for smaller magnitude conditions.

Figure 16 presented the Type I error and power rates for HGLM and GMH for the 3 DIF patterns and 2 levels of percentage of DIF. The main effect of the percentage of DIF was small for power, but with notable interaction effect with DIF patterns. For HGLM, the percentage of DIF had little effect on Type I error and power rates. For GMH, higher percentage of DIF items would decrease power and inflated error for the constant pattern. The balanced pattern did not seem to be affected. For the unbalanced pattern, power was not affected, but higher percentage of DIF inflated the Type I error rates.

Figure 17 presented the Type I error and power rates for HGLM and GMH for the 3 DIF patterns and 2 levels of magnitude of DIF. When the magnitude was large, HGLM behaved well with constant DIF, while GMH was good with balanced DIF. For the unbalanced pattern, HGLM was less powerful but with less error, while GMH was more powerful but with more error. When the magnitude was small, the power rates were much lower for both methods. The GMH was more powerful with non-constant DIF, while the GMH was more powerful with constant DIF.

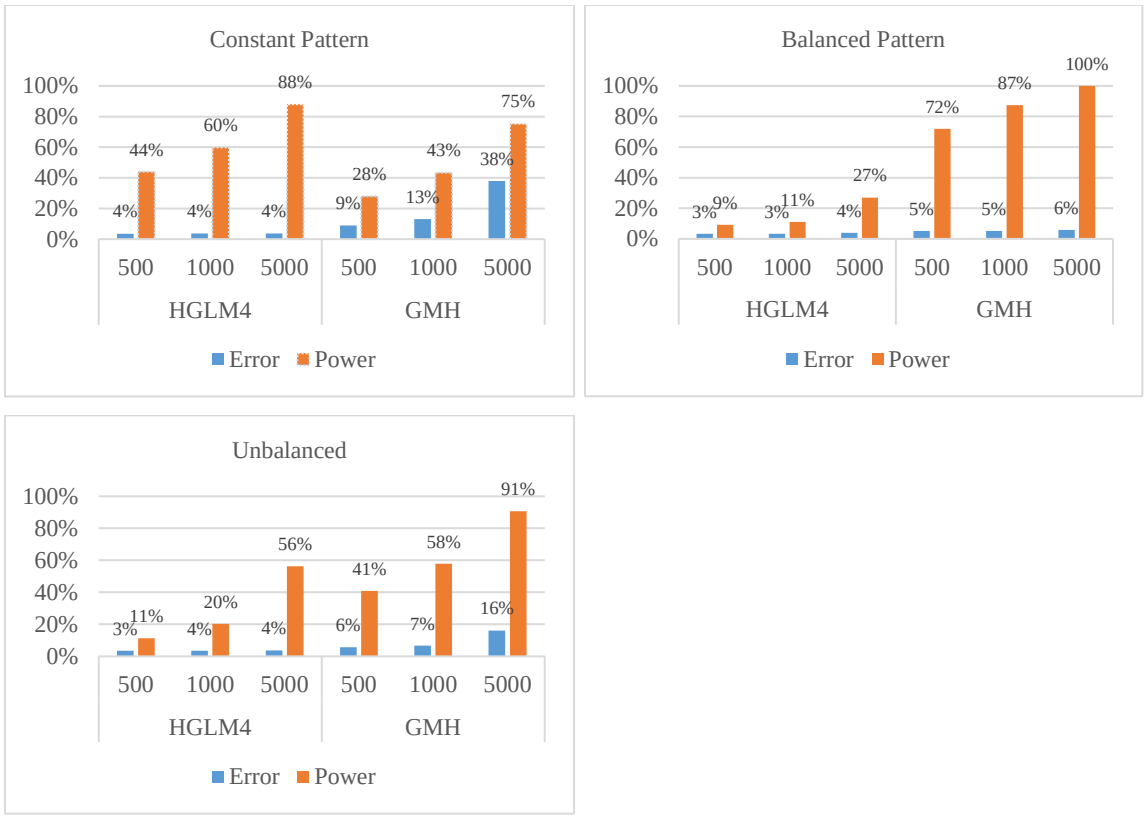


Figure 14 Type I Error and Power Rates for Sample Size and DIF Pattern for HGLM with 4-item Anchor and GMH

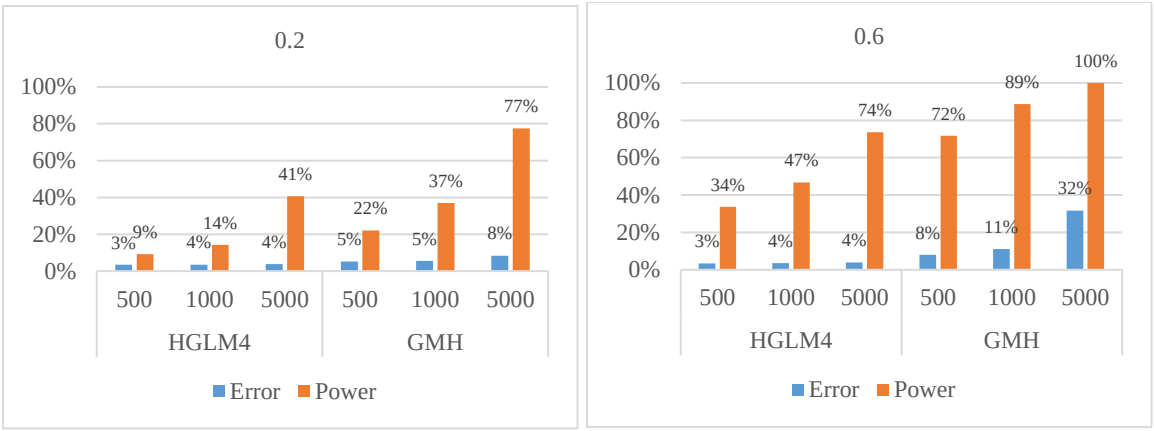


Figure 15 Type I Error and Power Rates for Sample Size and Magnitude of DIF for HGLM with 4-item Anchor and GMH

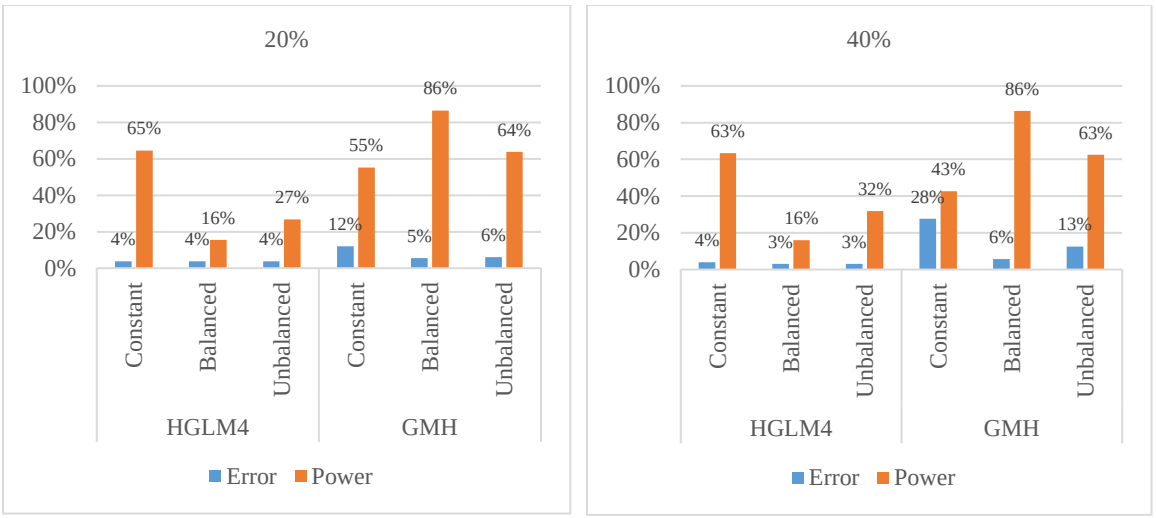


Figure 16 Type I Error and Power Rates for DIF Pattern and Percentage of DIF for HGLM with 4-item Anchor and GMH

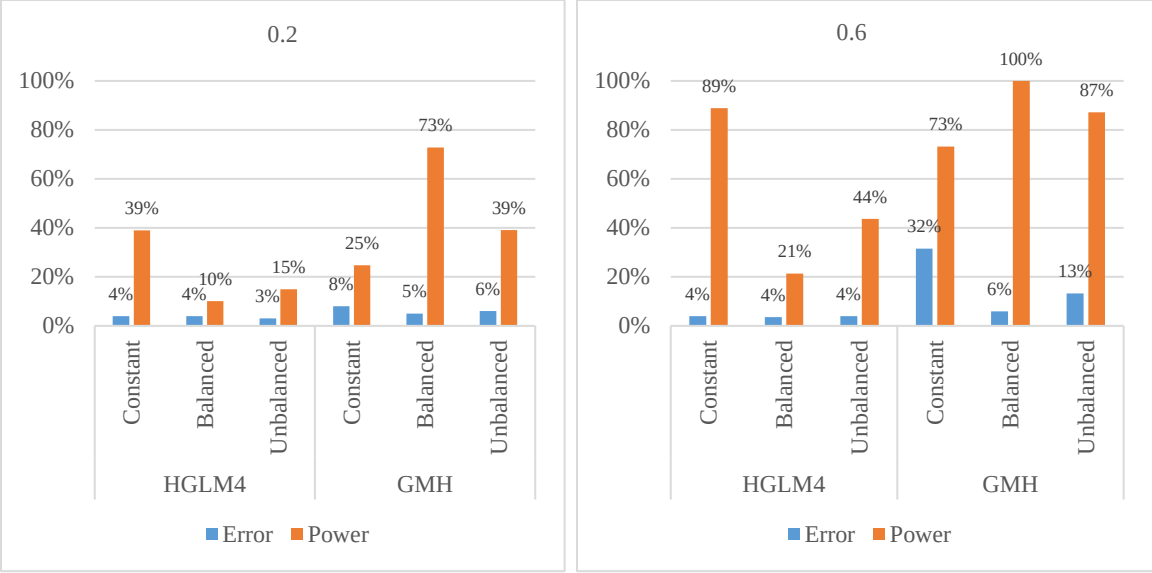


Figure 17 Type I Error and Power Rates for DIF Pattern and Magnitude of DIF for HGLM with 4-item Anchor and GMH

5.0 Discussion

This study aimed to examine the performance of using HGLM as a DIF detection model to identify polytomous response items that exhibit DIF, and how it compared to other well-established methods such as the logistic regression and the GMH methods. This chapter presents a summary of the study's major findings and its practical implications to practitioners, followed by a discussion of limitation and directions for future research. The chapter is organized as follows: first, a discussion of major findings and implications was presented. The answer to three research questions were discussed first, then the factors that influence the DIF detection performance were discussed separately, followed by a conclusion. In the second section, the limitation of the study was discussed, followed by a prediction of future research directions.

5.1 Major Findings and Implications

5.1.1 Answers to research questions

This study provides answers to three research questions in terms of using HGLM, logistic regression and the GMH methods for DIF detection. Findings are partly consistent with those from previous studies, while providing additional information for the use of HGLM as a DIF detection method.

Research question 1: How accurately can HGLM select DIF-free items as anchor items for DIF analysis?

This study found that the average rate of using HGLM with interactive constant item method yield an average accuracy of 75% across all conditions. This method had been proved to be highly accurate with mostly over 98% in dichotomous items under similar conditions (Chen et al., 2014), however, with polytomous it was not as accurate. Consistent with previous findings, this method was robust against the number of anchor items, suggesting that using this method to select 1 DIF-free item or 4 would not impact its performance.

The accuracy rate was most significantly influenced by the percentage of items with DIF, the magnitude of DIF, the pattern of DIF, and the sample size. When there was a small number of items that exhibited DIF in a scale, the iterative HGLM method could be mostly accurate at over 80%, but when there was a larger number of items with DIF, its performance significantly worsened. In addition, the larger magnitude of DIF and a larger total sample size appeared to help with the performance and increased the accuracy rates. The HGLM method was also sensitive to the pattern of DIF, with the constant pattern resulted in a much higher accuracy rate of over 80%, while balanced and unbalanced pattern yielded significantly less accurate results.

These results suggested that the ideal condition for using HGLM method to select DIF-free items would be a sample size of at least 1000, with a scale that has less than 20% of items exhibiting significant size of DIF, and the pattern of DIF was constant. The presence of differences in group abilities would negatively affect the performance, but it was not as influential as the previous mentioned factors. When the conditions were less ideal, the performance of the iterative HGLM method was less desirable, and for the practitioners, the implication was that it would be wise to use some other methods to cross check the items for DIF. Since previous studies had shown that anchor items that exhibited DIF would result in biased estimations in identifying DIF items, the anchor items should be carefully selected to ensure they were indeed DIF-free (Cheong & Kamata,

2013). The iterative HGLM method showed nearly perfect accuracy with dichotomous data. However, since the HGLM method was not as accurate with polytomous data, researchers should consider other reliable methods to double check the items selected by the HGLM method.

Research question 2: What is the Type I error rate for DIF detection using HGLM, and how does it compare to using GMH and logistic regression?

The HGLM with 1-item anchor and 4-item anchor both showed decent control over Type I error rates under all conditions, which was consistent with previous studies (Chen et al., 2014; Wang, 2004). For both methods, nearly all Type I error rates were below 5%. The average error rates for the HGLM with 1-item anchor was 3.7%, and for the HGLM with 4-item anchor was 3.61%. Consistent with the findings of Chen et al. (2014), these were rather conservative. The two models performed quite similarly to each other, suggested that for the purpose of controlling misidentification of DIF-free items as DIF, using 1-item anchor would be just as effective as the 4-item anchor.

The polytomous logistic regression and the GMH methods, on the other hand, showed general higher Type I error rates than the HGLM methods, as well as significant inflation of Type I error for certain conditions. The GMH method had been shown to have a general higher Type I error rates than the HGLM methods, but did not show a severe inflation (Ryan, 2008; Willams & Beretvas, 2006). However, these studies did not consider a scale with DIF items as much as 40%. The higher percentage of item with DIF seemed to relate to the strong inflation of error rates for the GMH as well as the logistic regression methods. The polytomous logistic regression method performed generally worse than the GMH, with an average error rate of 13.05% over 11.65%, with a similar pattern of behavior, which seemed to be consistent with previous findings (Kristjansson

et al., 2005). For certain conditions, the GMH and the logistic regression showed significant inflation of Type I error rates, the risk factors being the constant DIF pattern, larger number of items exhibiting DIF, larger magnitude of DIF, and larger sample sizes. The number of items exhibiting DIF appeared to be especially influential.

The inflated Type I error rates suggested that caution must be taken when using these two methods to identify DIF items for polytomous response items, since inflated Type I error rate might render the power of such testing meaningless. In addition, mistakenly marking DIF-free items as DIF is not desirable for practical use, as it could create unnecessary revising or modifying, thus increasing workload and disturbing the original scales.

Research question 3: What is the statistical power for DIF detection using HGLM, and how does it compare to using GMH and logistic regression?

The average power rates for the HGLM with 1-item anchor and 4-item anchor were 28% and 36%, respectively, which were somewhat consistent with previous findings (Ryan, 2008; Willams & Beretvas, 2006). In general, the 4-item anchor method out-performed the 1-item anchor method. Increasing sample size would increase the power rates, as many researchers had suggested, however, it was only combined with the constant pattern and a larger magnitude of DIF to achieve power rates over 90%. When sample size was 500, almost all the power rate would not exceed 50%.

The polytomous logistic regression method had a similar performance as the HGLM with 1-item anchor. The GMH method, however, behaved quite differently than the other three and in many conditions had very high power. Specifically, when the magnitude of DIF was large, the GMH displayed near perfect power with a balanced DIF pattern under all conditions. When the

sample size was also large, the GMH had high power at unbalanced DIF pattern as well. It appeared that the GMH was sensitive to the non-constant DIF patterns, while the HGLM methods performed poorly for such patterns. However, when sample size and the magnitude of DIF was large, the GMH showed inflated Type I error rates under the unbalanced pattern, so the power of the GMH on the unbalanced DIF pattern should be interpreted carefully. For the constant pattern, the GMH was generally not very powerful, and it showed such highly inflated Type I error rates, that the power, no matter how high, would be meaningless.

These results suggested that sufficient power was difficult to achieve when the sample size was small. Smaller magnitude of DIF also made the detection of DIF items difficult, thereby resulting in very low power rates. With smaller sample size, the GMH was a little more powerful than the other three; however, the GMH was also associated with higher level of Type I error rates than the HGLM methods. The two HGLM methods performed similarly at controlling for Type I error rate, but the HGLM with 4-item anchor had better power, so the 4-item anchor should be favored over the 1-item anchor. The polytomous logistic regression method had similar pattern of power as the HGLM with 1-item anchor, but had much higher inflation of Type I error rates, hence it should be a less favorable choice.

The HGLM methods had clear advantages for the constant pattern, which is the condition when all DIF were favoring one group. Under the constant pattern, practitioners should avoid the GMH method since its very much prone to Type I error rates. Thus, the HGLM with 4-item anchor method should be preferred when DIF pattern is constant. With a balanced pattern, the GMH method should be favored. With unbalanced DIF, the GMH were powerful but also more prone to Type I error, while the HGLM with 4-item anchor was less powerful but less prone to

misidentification. As a result, caution should be taken with selecting appropriate DIF detection methods.

It is worth noting that the Type I error and power correlate with each other, so when type I error rate is high, the high power is related to the elevated Type I error rates. In this situation, discussing power is meaningless. Thus, caution should be taken when interpreting the polytomous logistic regression and GMH results, especially under risky conditions. It is sensible to make sure that the power is actually meaningful, and not the results of inflated Type I error, before proceeding with interpretation.

5.1.2 Summary of major findings

This study manipulated several factors: the number of anchor items, the presence of impact, sample size and ratio, the percentage of DIF items, the magnitude of DIF, and the pattern of DIF. The effect of these factors for the performance of the HGLM methods as well as the polytomous logistic regression and the GMH methods were discuss separately.

Anchor items This study examined the performance of HGLM with 1-item anchor and 4-item anchor. Several studies had shown that an anchor with a larger number of items can produce higher power for DIF detection (Shin & Wang, 2009; Wang & Yeh, 2003), although a 1-item anchor could be enough in giving sufficient power. This study confirmed that a higher power rate was indeed associated with the 4-item anchor. The HGLM with 1-item anchor and 4-item anchor showed similarly good control over Type I error rates, but the 4-item anchor consistently outperformed the 1-item anchor in terms of power. These results seem to favor the HGLM with 4-item anchor over the 1-item anchor.

However, although the 4-item anchor is more powerful, the 1-item anchor can minimize the risk of contamination by items with DIF (Woods, 2009), and since the consequences of such contamination is severe (Cheong & Kamata, 2013), this should be taken into consideration when it comes to the anchor selection. As this study showed, the accuracy rates of selecting DIF-free items using the HGLM methods for polytomous responses items were rarely perfect, on average only about 75%, thus raising the challenge of selecting a pure anchor without DIF.

Latent trait parameter difference between groups The latent trait difference between the focal and reference groups, known as impact, was set at 0 and 1 for this study. When impact equaled to 0, there was no group difference existed. When the impact was 1, there was medium to large group differences on the latent trait parameter between the reference and focal group. The effect of impact on polytomous DIF is not well studied; Kristjansson et. al. (2005) found an impact size of 0.5 has little effect on the performance of DIF. On dichotomous items, Chen et al. (2014) found that impact size of 1 has little effect on the HGLM constant anchor method. Consistent with previous findings, this study found that impact did not significantly affect the performance of Type I error rates nor power. These results suggest that the HGLM with constant anchor item method might be quite robust against the presence of medium-large impact.

However, this study also found out that although the HGLM with constant anchor item method was quite accurate with dichotomous items, with polytomous response items, the existence of impact lowered the accuracy rate of selecting DIF-free items. These findings suggest that a medium-large impact should be taken into consideration when using HGLM with constant anchor item method to identify DIF-free items in polytomous response items.

Sample size and sample size ratio This study examined 6 sample size and ratio combinations (R400/F100, R250/F250, R800/F200, R500/F500, R4000/F1000, R2500/F2500) to study its influence on Type I error rate and power. Many studies suggested that a larger sample size may also inflate Type I error rates; however, this tendency was not found for the two HGLM methods. Even when the sample size was as large as 5000, the Type I error rates for the two HGLM methods were consistently below 5%. Consistent with findings from previous studies, a larger sample size was associated with larger power. The HGLM with 4-item anchor was more powerful overall than the 1-item anchor; however, neither was particularly powerful when sample size was 1000 and below. Willams and Beretvas (2006) found similar results with a sample size of 2000 and speculated that larger sample sizes would generate more power. However, this study showed that a higher power rate of over 80% would be difficult to achieve even when sample size was increased to 5000; the pattern and magnitude of DIF also had to be ideal. When the DIF pattern was constant and the magnitude of DIF was large, the HGLM with 4-item anchor was very powerful under larger sample size, and thus should be favored.

When sample size was small, the GMH method could generally produce much higher power than the other 3 methods. While it was associated with a slight elevation of Type I error rates, overall the GMH could be tentatively recommended for small samples. When the sample size increased, the GMH still tend to be more powerful than the other three; however, it was often associated with inflated Type I error rate as high as 95%. With such high error rates, the power, no matter how high, would be meaningless.

Percentage of DIF items The percentage of DIF was set to 0%, 20%, and 40% for this study. For the two HGLM methods, the percentage of DIF had little effect on the control of Type I error nor

power on average. Although, the two methods showed a slight increase in power for the unbalanced pattern when percentage of DIF items increased. However, the percentage of DIF had significant effect on the accuracy of selecting DIF-free items. It appeared that a large number of items exhibiting DIF would increase the difficulty for the HGLM method to pick out DIF-free items.

The logistic regression and GMH methods showed an increasing in Type I error rates and decreasing in power when the percentage of DIF increased, which was consistent with previous studies (Hidalgo & Gómez, 2006; Kristjansson et al., 2005). When the percentage of DIF was large, the two methods showed particularly large inflation of Type I error rates when sample size and magnitude of DIF were large. In addition, the constant pattern also resulted in inflated Type I error rate when the percentage was large. These results suggest that when there are a large number of items exhibiting DIF, caution should be taken when applying the logistic regression or GMH methods as DIF detection methods.

Magnitude of DIF The magnitude of DIF was set to 0.2 and 0.6 for this study, representing a small and a medium-large magnitude of DIF. This study found that a smaller magnitude of DIF would result in more difficulty finding items with DIF, thus reduce power, which was consistent with findings from previous studies on dichotomous data (Elosua & Wells, 2013; Hildago & Gómez, 2006; Scott, 2009). When the magnitude of DIF was small, the GMH appeared to be most powerful out of the 4 methods, although it was still less than 50%. Since in health research the magnitude of DIF is commonly small, this brings challenges in DIF evaluation. When the magnitude was large, the power for all 4 methods increased, however, for the logistic regression and GMH methods, the Type I error rates were inflated as well.

DIF patterns The 3 pattern of DIF considered in this study were constant, balanced, and high-unbalanced. For the constant pattern, all the DIF were in favor of the focal group. For the balanced pattern, some of the DIF favor the reference group and some favor the focal group, resulting the DIF to be balanced across item categories. For the high-unbalanced pattern, DIF were only present at the highest category within an item. Previous studies had shown that balanced and unbalanced DIF was especially difficult to detect, which was confirmed by this study.

For the constant pattern, the HGLM with 4-item anchor method was the most powerful; this was consistent with the findings by Cheong and Kamata (2013). The GMH showed significant inflation of Type I error, thereby should be avoid when DIF were consistently favoring one group over another. For the balanced DIF, the GMH was most powerful, and the Type I error rates were well under control, although on some conditions were still slightly elevated. Since GMH is designed to measure group differences across the entire distribution of response categories, it is expected it would be more sensitive to the balanced pattern which favors different groups within the item categories. For the unbalanced DIF pattern, the GMH was more powerful than the HGLM methods, but also more likely to have inflated Type I error.

5.1.3 Conclusion

This study extended the work of Chen et al. (2014), exploring the performance of the HGLM model using the constant anchor item method to identify the model and to select the DIF-free items for the polytomous response items, instead of dichotomous items. In addition, this study extended the work of Willams and Beretvas (2006) and Ryan (2008) by using the constant anchor method instead of the equal mean ability method to set the model, and by comparing the results to the logistic regression method, as well as the GMH method. This study also included the presence

of ability differences between the reference and focus groups so the effect of impact could be studied.

The study found out that compared to dichotomous items, the accuracy rate of HGLM methods in selecting DIF-free items was generally lower in polytomous item. While for dichotomous item the accuracy was near perfect as found in Chen et al. (2004), the accuracy for polytomous items were only about 75%. Since constant anchor method only performs well when the anchor items are free of DIF, practitioners should be very careful when using the iterative HGLM methods to select DIF-free items to serve as anchor items.

Overall, the HGLM with 1-item anchor and 4-item anchor methods both have decent control over Type I error rates. However, the HGLM with 4-item anchor method is more powerful than the 1-item anchor method, so if possible, the 4-item anchor method should be favored. The polytomous logistic regression method has similar power rates as the HGLM with 1-item anchor method, but higher Type I error rates, thus is not recommended over the HGLM with 4-item anchor method.

The GMH method is overall more powerful, but also prone to Type I error rates. Since high power without decently controlled Type I error rates is meaningless, the results of GMH should be used with caution. The GMH can handle balanced DIF pattern much better than the constant pattern. If all the DIF favor one group over the other, resulting in a constant DIF pattern, the GMH is not very powerful and prone to significantly inflated Type I error, thus is not recommended. Instead the HGLM with 4-item anchor is recommended. If the DIF with one item favors different groups on different response categories, resulting in a balanced DIF pattern, the GMH is recommended, since it is much more powerful, and the Type I error rate is more controlled. For the unbalanced DIF pattern, it is hard to give a general recommendation to balance the Type I error

and power rates. Typically, the GMH is more powerful when sample size and the magnitude of DIF is small, and under these conditions the inflation of Type I error is not too severe, so GMH could be considered. When the sample size and magnitude of DIF is large, the GMH tends to have highly inflated Type I error that renders the power meaningless. Whereas under these conditions, the power of HGLM with 4-item anchor increases, with the Type I error rate is still under control. Thus, the HGLM with 4-item anchor method could be considered. Cheong and Kamata (2013) suggested that the when the GMH type method is detecting a large number of DIF items, researchers should consider repeat the analysis with the HGLM with constant anchor item method. The findings from this study support this suggestion and recommend researchers and practitioners to re-check the analysis if a non-parametric method such as the GMH is flagging a large number of items for DIF, possibly with a method such as the HGLM with constant anchor item.

5.2 Limitations and Future Research

This study uses simulated data to explore the performance of HGLM methods as well as the logistic regression and GMH methods under various conditions for polytomous items. Although simulation studies have many advantages, it cannot fully substitute empirical data. The simulation factors were all taken from literature and meant to be close to mimic conditions present in real data situations. However, this study and its findings are restricted to these simulated conditions. The generalizability of findings is not guaranteed for real life datasets, as the real data may be much more complex than simulated data.

Another clear limitation of this study is that only polytomous items with 3 categories were considered. Items with more categories were commonly used in psychology and health studies, so

it’s necessary to study DIF for polytomous items with higher number of categories. In addition, higher numbers of categories within items can result in much more complex DIF patterns, which may complicate the performance of DIF evaluation. Future studies are needed to investigate polytomous items with more than 3 response categories.

Additionally, this study compared the HGLM method to only two other methods: the polytomous logistic regression and the GMH method. The GMH is a non-parametric method widely used in education settings while the logistic regression is a parametric method very commonly used in health studies and both are well-studied. However, another popular set of DIF detection methods based on the IRT likelihood ratio test was not discussed at all in this study. IRT model as DIF detection methods has several advantages and is becoming increasingly popular. In addition, although the HGLM method has been shown to mathematically comparable to certain IRT model, the two’s actual performance and comparability for DIF evaluation has not been explored. For these reasons, it is necessary for future research to study and compare the HGLM and IRT methods.

Furthermore, the HGLM method is capable of identifying and investigating the source of DIF simultaneously by adding covariates into the model. The HGLM method can also handle more levels in the model to account for common clusters, such as schools and cohorts. This is a clear advantage of the HGLM method, yet its performance on exploring potential sources of DIF goes unexplored. It is necessary for future researchers to explore this feature of the HGLM method, and establish guidelines for practitioners to reference in empirical studies.

Appendix A Detailed Results for Study 1

Appendix Table 1 Means and Standard Deviations for the Accuracy of Selecting DIF Items

Source	N	Mean	SD
Number of Anchor Items			
Anchor=1	96	75.39	15.01
Anchor=4	96	75.17	13.89
Impact			
Impact=0	96	76.05	14.50
Impact=1	96	74.51	14.38
Sample size and ratio			
R400/F100	96	73.45	13.69
R250/F250	96	73.82	14.55
R800/F200	96	76.19	14.44
R500/F500	96	77.67	15.02
Percentage of DIF			
20%	96	86.90	6.42
40%	96	63.66	10.20

Magnitude			
.2	96	71.09	13.08
.6	96	79.47	14.54
DIF Pattern			
constant	96	82.53	14.41
Balance	96	72.05	12.83
High-Unbalanced	96	71.25	13.34

Appendix Table 2 ANOVA Results for the Accuracy of Selecting DIF Items

Source	DF	SS	MS	F	p-value	η_p^2
impact	1	115.24	115.24	11.40	0.001	0.09
samplesize	3	576.70	192.23	19.01	<.001	0.32
pdif	1	25917.95	25917.95	2562.93	<.001	0.95
magnitude	1	3368.84	3368.84	333.13	<.001	0.74
pattern	2	5067.93	2533.97	250.58	<.001	0.81
pdif*magnitude*pattern	2	552.78	276.39	552.78	<.001	0.31
samplesize*pattern	6	332.91	55.48	332.91	<.001	0.20
pdif*pattern	2	190.15	95.08	190.15	<.001	0.14
magnitude*pattern	2	2066.20	1033.10	2066.20	<.001	0.63
pdif*magnitude	1	113.70	113.70	113.70	0.001	0.08
residuals	74	748.33	10.11			

MANOVA test for Anchor				
Statistics	Value	F Value	DF	p-value
Wilks' Lambda	0.99	0.38	1	0.54
Pillai's Trace	0.01	0.38	1	0.54
Hotelling-Lawley Trace	0.01	0.38	1	0.54
Roy's Greatest Root	0.01	0.38	1	0.54

Appendix Table 3 Simple Comparison for the Accuracy of Selecting DIF Items, DIF Pattern by Sample Size

Source	DF	F	p-value
Constant	3	16.53	<.001
Balance	3	2.46	0.07
High-Unbalanced	3	17.00	<.001

Appendix Table 4 Simple Comparison for the Accuracy of Selecting DIF Items, DIF Pattern by Percentage of DIF Items

Source	DF	F	p-value
Constant	1	793.14	<.001
Balance	1	1128.91	0.07
High-Unbalanced	1	1176.43	<.001

Appendix Table 5 Simple Comparison for the Accuracy of Selecting DIF Items, DIF Pattern by Magnitude of DIF Items

Source	DF	F	p-value
Constant	1	590.91	<.001
Balance	1	33.67	<.001
High-Unbalanced	1	20.44	<.001

Appendix Table 6 Simple Comparison for the Accuracy of Selecting DIF Items, Percentage of DIF by Magnitude of DIF Items

Source	DF	F	p-value
20%	1	133.20	<.001
40%	1	280.10	<.001

Appendix Table 7 Means And Standard Deviations for The Accuracy of Selecting DIF Items, DIF Pattern by Sample Size

Pattern	Sample size	N	Mean	SD
Constant	R400/F100	16	79.03	14.17
	R250/F250	16	81.66	14.63
	R800/F200	16	83.39	15.35
	R500/F500	16	86.05	13.91
Balanced	R400/F100	16	72.64	12.16
	R250/F250	16	70.59	13.70
	R800/F200	16	73.20	12.84
	R500/F500	16	71.77	13.66
Un-Balanced	R400/F100	16	68.67	13.41
	R250/F250	16	69.20	12.74
	R800/F200	16	71.97	12.99
	R500/F500	16	75.19	14.42

Appendix Table 8 Means and Standard Deviations for the Accuracy of Selecting DIF Items, DIF Pattern by Percentage of DIF Items

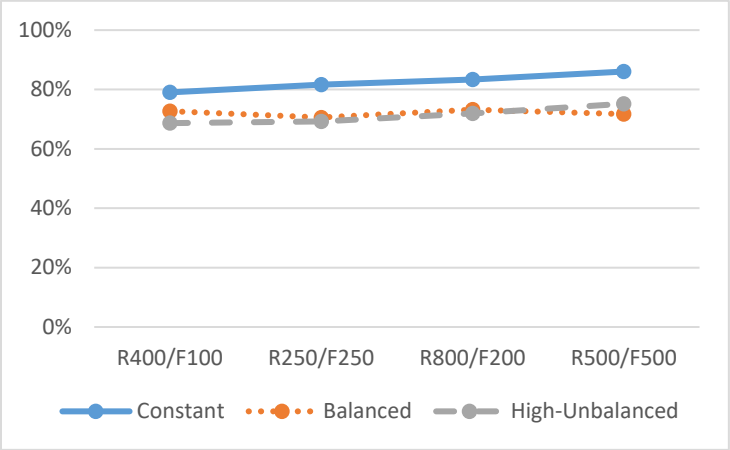
Pattern	Percentage of DIF	N	Mean	SD
Constant	20%	32	92.75	6.55
	40%	32	72.31	12.78
Balanced	20%	32	84.24	2.44
	40%	32	59.86	4.66
High-Unbalanced	20%	32	83.70	4.93
	40%	32	58.81	4.20

Appendix Table 9 Means and Standard Deviations for the Accuracy of Selecting DIF Items, DIF Pattern by Magnitude of DIF Items

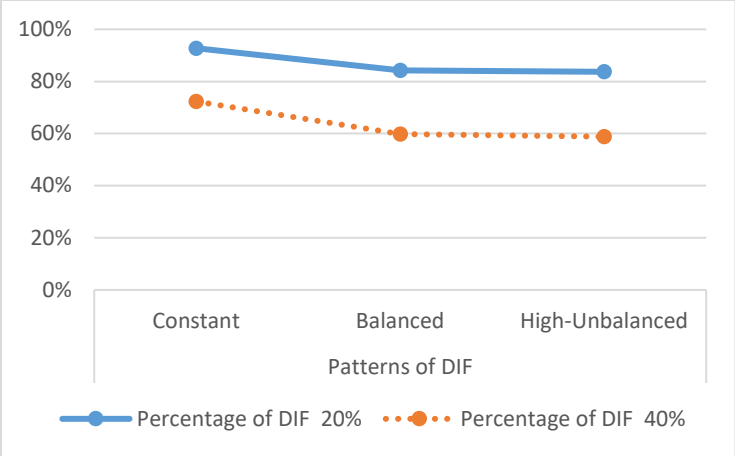
Pattern	Magnitude of DIF	N	Mean	SD
Constant	0.2	32	73.71	13.92
	0.6	32	91.35	8.22
Balanced	0.2	32	69.95	13.18
	0.6	32	74.16	12.32
High-Unbalanced	0.2	32	69.62	12.11
	0.6	32	72.90	14.47

Appendix Table 10 Means and Standard Deviations for the Accuracy of Selecting DIF Items, Percentage of DIF Items by Magnitude of DIF

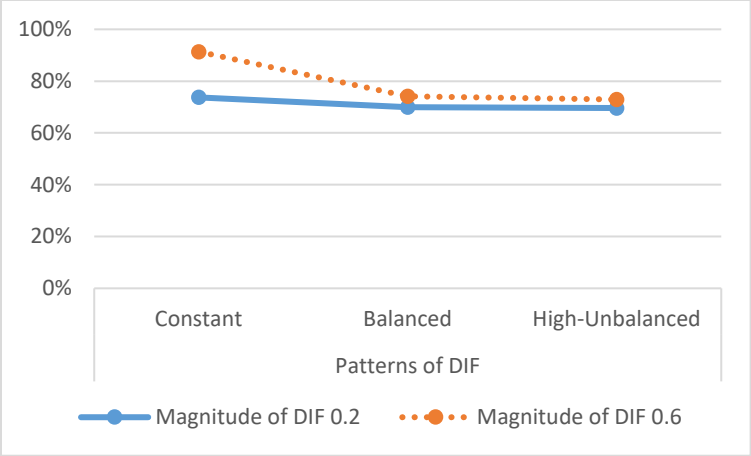
% of DIF	Magnitude	N	Mean	SD
20%	0.02	48	83.48	3.86
	0.06	48	90.32	6.68
40%	0.02	48	58.70	4.21
	0.06	48	68.62	11.94



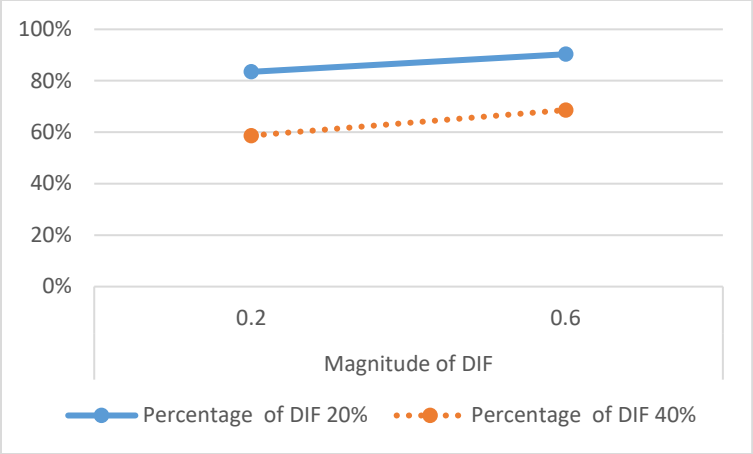
Appendix Figure 1 Two-way Interaction of Accuracy between DIF Pattern and Sample Size



Appendix Figure 2 Two-way Interaction of Accuracy between DIF Pattern and Percentage of DIF Items



Appendix Figure 3 Two-way Interaction of Accuracy between DIF Pattern and Magnitude of DIF Items



Appendix Figure 4 Two-way Interaction of Accuracy between Percentage and Magnitude of DIF

Appendix B Detailed Results for Study 2

Appendix B.1 Results for Type I Error

Appendix Table 11 Mean Type I Error Rates (%), without Impact

Sample Size	DIF%	Magnitude	DIF Patterns	HGLM1	HGLM4	LR	GMH
R400/F100	20%	0.2	Constant	3	3	5	5
			Balanced	3	4	5	5
			Unbalanced	3	4	5	5
		0.6	Constant	4	4	8	7
			Balanced	3	3	5	6
			Unbalanced	4	3	5	5
	40%	0.2	Constant	4	4	6	6
			Balanced	4	3	5	5
			Unbalanced	3	3	5	5
		0.6	Constant	3	3	19	14
			Balanced	3	3	6	5
			Unbalanced	3	3	7	6
	0	0	Reference	3	3	5	5
R250/F250	20%	0.2	Constant	4	4	6	5
			Balanced	3	3	5	5
			Unbalanced	4	4	5	5

		0.6	Constant	4	4	10	9
			Balanced	4	4	5	5
			Unbalanced	4	4	6	5
	40%	0.2	Constant	4	3	8	6
			Balanced	3	3	5	5
			Unbalanced	4	3	5	5
		0.6	Constant	5	4	27	19
			Balanced	4	3	6	6
			Unbalanced	4	4	10	8
	0	0	Reference	3	4	5	5
R800/F200	20%	0.2	Constant	4	4	6	5
			Balanced	4	4	5	5
			Unbalanced	4	3	5	5
		0.6	Constant	4	4	11	9
			Balanced	4	4	6	5
			Unbalanced	3	4	7	6
	40%	0.2	Constant	5	4	8	7
			Balanced	4	3	5	5
			Unbalanced	3	3	6	5
		0.6	Constant	5	3	31	22
			Balanced	3	3	6	6
			Unbalanced	4	4	12	9
	0	0	Reference	3	3	5	5

R500/F500	20%	0.2	Constant	4	3	6	6
			Balanced	4	4	5	5
			Unbalanced	3	4	5	5
		0.6	Constant	5	4	15	12
			Balanced	3	3	5	5
			Unbalanced	4	3	7	6
	40%	0.2	Constant	4	4	10	8
			Balanced	4	3	6	5
			Unbalanced	4	3	6	5
		0.6	Constant	5	3	47	35
			Balanced	3	3	6	6
			Unbalanced	5	3	16	12
	0	0	Reference	4	4	5	5
R4000/F1000	20%	0.2	Constant	3	3	9	7
			Balanced	3	5	5	5
			High unbalanced	3	4	6	6
		0.6	Constant	5	4	41	31
			Balanced	3	4	6	5
			High unbalanced	3	5	13	10
	40%	0.2	Constant	3	4	20	15
			Balanced	4	3	5	5

			High unbalanced	4	4	9	8
		0.6	Constant	5	3	92	83
			Balanced	3	3	8	7
			High unbalanced	4	3	37	28
	0%	0	reference	4	4	5	5
R2500/F2500	20%	0.2	Constant	4	4	11	9
			Balanced	3	3	5	5
			High unbalanced	3	3	6	6
		0.6	Constant	4	4	55	42
		0.6	Balanced	3	4	7	6
		0.6	High unbalanced	3	3	17	12
	40%	0.2	Constant	4	3	28	21
			Balanced	3	3	5	5
			High unbalanced	1	1	11	9
		0.6	Constant	8	4	98	95
			Balanced	4	3	12	10
			High unbalanced	5	4	64	51
	0%	0	reference	3	4	5	5

Appendix Table 12 Mean Type I Error Rates (%), with Impact = 1

Sample Size	DIF%	Magnitude	DIF Patterns	HGLM1	HGLM4	LR	GMH
R400/F100	20%	0.2	Constant	4	4	6	5
			Balanced	4	4	5	5
			Unbalanced	3	3	5	5
		0.6	Constant	4	4	10	7
			Balanced	3	4	5	6
			Unbalanced	3	3	5	5
	40%	0.2	Constant	4	4	6	6
			Balanced	4	3	5	5
			Unbalanced	3	3	5	5
		0.6	Constant	4	3	16	14
			Balanced	3	4	6	5
			Unbalanced	3	3	5	6
	0	0	Reference	3	3	5	5
R250/F250	20%	0.2	Constant	4	4	6	5
			Balanced	4	4	5	5
			Unbalanced	4	4	5	5
		0.6	Constant	4	4	14	9
			Balanced	4	3	6	5
			Unbalanced	3	4	6	5

	40%	0.2	Constant	4	3	7	6
			Balanced	3	3	5	5
			Unbalanced	4	3	6	5
		0.6	Constant	4	4	22	19
			Balanced	3	3	6	6
			Unbalanced	4	4	9	8
	0	0	Reference	4	3	5	5
R800/F200	20%	0.2	Constant	4	4	6	5
			Balanced	4	4	6	5
			Unbalanced	4	3	5	5
		0.6	Constant	3	4	11	9
			Balanced	4	4	6	5
			Unbalanced	3	4	6	6
	40%	0.2	Constant	4	3	6	7
			Balanced	4	4	5	5
			Unbalanced	3	3	5	5
		0.6	Constant	4	4	20	22
			Balanced	4	3	6	6
			Unbalanced	4	3	7	9
	0	0	Reference	3	3	6	5
R500/F500	20%	0.2	Constant	5	5	8	6
			Balanced	5	3	5	5
			Unbalanced	5	4	6	5

		0.6	Constant	4	4	17	12
			Balanced	4	4	6	5
			Unbalanced	4	5	7	6
	40%	0.2	Constant	5	4	9	8
			Balanced	4	3	5	5
			Unbalanced	4	3	6	5
		0.6	Constant	5	4	39	35
			Balanced	4	3	7	6
			Unbalanced	4	4	14	12
	0	0	Reference	4	4	5	5
R4000/F1000	20%	0.2	Constant	3	3	8	7
			Balanced	3	4	7	5
			High unbalanced	4	4	6	6
		0.6	Constant	4	4	27	31
			Balanced	3	4	12	5
			High unbalanced	4	4	7	10
	40%	0.2	Constant	3	3	16	15
			Balanced	4	4	6	5
			High unbalanced	4	7	7	8
		0.6	Constant	3	4	78	83

			Balanced	3	4	15	7
			High unbalanced	3	4	18	28
	0%	0	reference	3	4	7	5
R2500/F2500	20%	0.2	Constant	3	5	13	9
			Balanced	4	5	5	5
			High unbalanced	4	4	6	6
		0.6	Constant	4	4	40	42
		0.6	Balanced	4	4	9	6
		0.6	High unbalanced	3	5	13	12
	40%	0.2	Constant	3	3	24	21
			Balanced	4	4	5	5
			High unbalanced	2	2	10	9
		0.6	Constant	5	4	89	95
			Balanced	4	6	14	10
			High unbalanced	4	5	55	51
	0%	0	reference	3	5	5	5

Appendix Table 13 Means and Standard Feviations for the Type I Error Rates

Source	N	Mean	SD
Impact			
Impact = 0	288	8.15	12.92
Impact =1	288	7.86	11.68
Sample size			
500	192	5.28	3.32
1000	192	6.34	6.08
5000	192	12.38	19.44
Sample size ratio			
1:1	288	8.78	13.72
4:1	288	7.22	10.67
Percentage of DIF			
20%	288	6.12	6.16
40%	288	9.89	16.07
Magnitude			
.2	288	5.15	3.10
.6	288	10.86	16.65

DIF Pattern			
constant	192	12.51	18.74
Balance	192	4.72	1.75
High-Unbalanced	192	6.78	8.29
Method			
HGLM1	144	3.70	0.72
HGLM4	144	3.61	0.65
Logistic Regression	144	13.05	16.71
GMH	144	11.65	15.84

Appendix Table 14 ANOVA Results for Type I Error Rates

ANOVA results	DF	SS	MS	F	p-value	η_p^2
Impact	1	0.001	0.001	0.27	0.604	0.002
Sample size	2	0.56	0.28	63.82	<.001	0.51
Sample size ratio	1	0.04	0.04	7.97	0.006	0.06
% of DIF	1	0.20	0.20	46.31	<.001	0.28
magnitude	1	0.47	0.47	106.42	<.001	0.47
pattern	2	0.63	0.31	70.96	<.001	0.54
Residuals	122	0.54	0.004			
MANOVA results for interactions	Num DF	Den DF	Wilks' Lambda	F	p-value	η_p^2
Method*sample size*pattern	12	137.78	0.56	6.54	<.001	0.41
Method*sample size*% of DIF	6	240	0.79	5.07	<.001	0.19
Method*sample size*magnitude	6	240	0.63	10.27	<.001	0.34
Method*% of DIF *pattern	6	240	0.79	5.06	<.001	0.19
Method*% of DIF *magnitude	3	240	0.81	9.32	<.001	0.17
Method *magnitude *pattern	6	240	0.61	11.13	<.001	0.36

MANOVA test for Method					
Statistics	Value	F Value	Num DF	Den DF	p-value

Wilks' Lambda	0.28	100.39	3	120	<.001
Pillai's Trace	0.71	100.39	3	120	<.001
Hotelling-Lawley Trace	2.51	100.39	3	120	<.001
Roy's Greatest Root	2.51	100.39	3	120	<.001

Appendix B.2 Results for Power

Appendix Table 15 Mean Power Rates (%), without Impact

Sample Size	DIF%	Magnitude	DIF Patterns	HGLM1	HGLM4	LR	GMH
R400/F100	20%	0.2	Constant	8	12	12	10
			Balanced	4	6	7	33
			Unbalanced	4	7	7	12
		0.6	Constant	50	65	60	47
			Balanced	7	10	13	100
			Unbalanced	8	15	13	58
	40%	0.2	Constant	8	12	8	7
			Balanced	4	5	7	33
			Unbalanced	4	6	6	12
		0.6	Constant	45	64	38	29
			Balanced	8	12	12	100
			Unbalanced	8	14	9	56
R250/F250	20%	0.2	Constant	11	16	16	12
			Balanced	4	5	5	54
			Unbalanced	5	7	7	14
		0.6	Constant	67	85	77	68
			Balanced	14	19	16	100
			Unbalanced	14	24	23	81

	40%	0.2	Constant	11	17	11	9
			Balanced	5	6	6	55
			Unbalanced	6	8	7	15
		0.6	Constant	63	84	53	43
			Balanced	10	12	11	100
			Unbalanced	21	32	21	79
R800/F200	20%	0.2	Constant	13	22	19	14
			Balanced	4	6	7	64
			Unbalanced	7	8	7	19
		0.6	Constant	73	90	85	76
			Balanced	12	18	18	100
			Unbalanced	22	38	33	92
	40%	0.2	Constant	11	19	11	10
			Balanced	6	6	7	63
			Unbalanced	4	7	5	19
		0.6	Constant	71	91	63	52
			Balanced	9	13	13	100
			Unbalanced	23	35	21	86
R500/F500	20%	0.2	Constant	19	29	24	18
			Balanced	6	6	6	86
			Unbalanced	7	10	10	27
		0.6	Constant	89	98	96	93
			Balanced	7	11	8	100

			Unbalanced	27	39	39	97
	40%	0.2	Constant	16	27	16	13
			Balanced	5	7	7	86
			Unbalanced	6	10	8	25
		0.6	Constant	91	99	81	72
			Balanced	11	16	13	100
			Unbalanced	40	58	40	97
R4000/F1000	20%	0.2	Constant	46	62	62	50
			Balanced	6	12	10	100
			High unbalanced	14	16	23	75
		0.6	Constant	100	100	100	100
			Balanced	15	23	23	100
			High unbalanced	74	82	81	100
	40%	0.2	Constant	50	59	41	31
			Balanced	7	10	11	100
			High unbalanced	14	35	15	72
		0.6	Constant	100	100	100	100
			Balanced	18	29	27	100
			High unbalanced	69	87	63	100

R2500/F2500	20%	0.2	Constant	67	86	84	74
			Balanced	4	6	6	100
			High unbalanced	27	35	38	91
		0.6	Constant	100	100	100	100
		0.6	Balanced	54	65	63	100
		0.6	High unbalanced	86	97	94	100
	40%	0.2	Constant	66	88	59	48
			Balanced	7	10	11	100
			High unbalanced	39	43	22	88
		0.6	Constant	100	100	100	100
			Balanced	41	49	43	100
			High unbalanced	98	100	96	100

Appendix Table 16 Mean Power Rates (%), with Impact = 1

Sample Size	DIF%	Magnitude	DIF Patterns	HGLM1	HGLM4	LR	GMH
R400/F100	20%	0.2	Constant	9	14	13	10
			Balanced	5	7	7	33
			Unbalanced	4	5	6	12
		0.6	Constant	41	58	49	47
			Balanced	6	9	10	100
			Unbalanced	4	8	7	58
	40%	0.2	Constant	8	12	9	7
			Balanced	6	8	8	33
			Unbalanced	4	7	5	12
		0.6	Constant	42	63	33	29
			Balanced	6	9	10	100
			Unbalanced	5	9	6	56
R250/F250	20%	0.2	Constant	11	15	12	12
			Balanced	6	9	8	54
			Unbalanced	4	7	6	14
		0.6	Constant	65	88	69	68
			Balanced	8	10	11	100
			Unbalanced	6	11	15	81
	40%	0.2	Constant	11	16	10	9
			Balanced	5	8	6	55

			Unbalanced	5	7	6	15
		0.6	Constant	60	82	45	43
			Balanced	8	10	9	100
			Unbalanced	10	15	14	79
R800/F200	20%	0.2	Constant	14	22	17	14
			Balanced	8	9	10	64
			Unbalanced	4	5	6	19
		0.6	Constant	69	86	78	76
			Balanced	10	13	14	100
			Unbalanced	9	17	16	92
	40%	0.2	Constant	11	19	14	10
			Balanced	6	10	8	63
			Unbalanced	4	6	5	19
		0.6	Constant	64	84	57	52
			Balanced	8	11	13	100
			Unbalanced	11	17	13	86
R500/F500	20%	0.2	Constant	21	41	22	18
			Balanced	7	10	7	86
			Unbalanced	6	7	8	27
		0.6	Constant	88	99	92	93
			Balanced	10	17	8	100
			Unbalanced	13	26	25	97
	40%	0.2	Constant	18	34	15	13

			Balanced	7	11	8	86
			Unbalanced	6	8	7	25
		0.6	Constant	88	99	71	72
			Balanced	11	14	12	100
			Unbalanced	19	32	24	97
R4000/F1000	20%	0.2	Constant	46	67	61	50
			Balanced	11	13	16	100
			High unbalanced	9	17	19	75
		0.6	Constant	100	100	100	100
			Balanced	28	35	40	100
			High unbalanced	39	57	56	100
	40%	0.2	Constant	47	62	43	31
			Balanced	13	23	15	100
			High unbalanced	10	30	14	72
		0.6	Constant	100	100	99	100
			Balanced	25	39	39	100
			High unbalanced	38	65	43	100
R2500/F2500	20%	0.2	Constant	71	94	78	74
			Balanced	12	29	6	100

			High unbalanced	16	31	28	91
		0.6	Constant	100	100	100	100
		0.6	Balanced	19	27	36	100
		0.6	High unbalanced	47	74	70	100
	40%	0.2	Constant	67	90	54	48
			Balanced	14	23	13	100
			High unbalanced	22	36	17	88
		0.6	Constant	100	100	100	100
			Balanced	31	42	35	100
			High unbalanced	81	97	78	100

Appendix Table 17 Means and Standard Deviations for Power

Source	N	Mean	SD
Impact			
Impact = 0	288	.42	.36
Impact =1	288	.39	.35
Sample size			
500	192	.25	.27
1000	192	.35	.34
5000	192	.61	.34
Sample size ratio			
1:1	288	.44	.36
4:1	288	.36	.33
Percentage of DIF			
20%	288	.41	.36
40%	288	.39	.35
Magnitude			
.2	288	.25	.26
.6	288	.56	.36

DIF Pattern			
constant	192	.55	.34
Balance	192	.32	.35
High-Unbalanced	192	.34	.32
Method			
HGLM1	144	.28	.30
HGLM4	144	.36	.33
Logistic Regression	144	.31	.30
GMH	144	.66	.34

Appendix Table 18 ANOVA Results for Power

Source	DF	SS	MS	F	p-value	η_p^2
Impact	1	0.09	0.09	4.87	0.03	0.04
Sample size	2	12.99	6.50	357.66	<.001	0.85
Sample size ratio	1	0.90	0.90	49.44	<.001	0.28
% of DIF	1	0.05	0.05	2.96	0.09	0.02
magnitude	1	14.47	14.47	796.49	<.001	0.86
pattern	2	5.97	2.98	164.27	<.001	0.72
Residuals	125	2.27	0.02			
MANOVA results for interactions	Num DF	Den DF	Wilk's Lambda	F	p-value	η_p^2
Method*sample size*pattern	12	325.72	0.70	3.84	<.001	0.05
Method*sample size*magnitude	6	246	0.37	26.13	<.001	0.32
Method*% of DIF *Pattern	6	246	0.77	5.62	<.001	0.03
Method*Magnitude*Pattern	6	264	0.68	8.54	<.001	0.12

MANOVA test for Method					
Statistics	Value	F Value	Num DF	Den DF	p-value
Wilks' Lambda	0.05	758.51	3	123	<.001
Pillai's Trace	0.95	758.51	3	123	<.001
Hotelling-Lawley Trace	18.50	758.51	3	123	<.001
Roy's Greatest Root	18.50	758.51	3	123	<.001

Appendix C SAS Syntax Sample

Appendix C.1 Sample Syntax for Study 1

```
*HGLM 1-item anchor no impact;
%do j=1 %to 20;
ods output ParameterEstimates=fixed;
ods output convergencestatus=converge;
proc glimmix data=long order=data method=rspl maxopt=1000 PCONV=.001
ABSPCONV=.001 noclprint noitprint;
class subid r(reference="R&j") group;
model resp= r group r*group/dist=multinomial solution link=clogit;
random intercept/subject=subid;
run;
```

Appendix C.2 Sample Syntax for Study 2

```
*HGLM with no impact;
ods output ParameterEstimates=fixed1;
ods output convergencestatus=converge1;
proc glimmix data=long method=laplace maxopt=1000 noclprint noitprint  ;*1
item anchor;
class subid r(reference= first) group;
model resp = r group r*group/dist=multinomial solution link=clogit;
random intercept/subject=subid type=un g;
run;

ods output ParameterEstimates=fixed4;
ods output convergencestatus=converge4;
proc glimmix data=long method=laplace maxopt=1000 noclprint noitprint  ;*4
item anchor;
class subid r group;
model resp = r5-r20 group r5*group r6*group r7*group r8*group r9*group
r10*group
r11*group r12*group r13*group r14*group r15*group r15*group r16*group
r17*group r18*group r19*group r20*group
/dist=multinomial solution link=clogit;
random intercept/subject=subid type=un g;
run;

*GMH ********************;
data total; * no impact;
set rfd;
total=sum(of r1-r20); run;
proc rank data=total out=rank group=10;
var total;
ranks stratun;
run;
%do d=1 %to 20;
ods output cmh=gmhout&d; * no impact;
proc freq data=rank; *GMH;
tables stratun*group*r&d/CMH;
run;
%end;

*LOGISTIC ********************;
%do d=1 %to 20;
proc logistic data=total;*no impact;
model r&d=total group;
ods output parameterestimates=logout&d;
run;
%end;
```

Bibliography

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47-76.

Agresti, A. (2013). *Categorical data analysis* (Vol. 3rd). Hoboken, NJ: Wiley.

Ahmadian, L., & Massof, R. (2008). Impact of general health status on validity of visual impairment measurement. *Ophthalmic Epidemiology, 15*(5), 345-355.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics and Outcomes Research, 11*(5), 571-585.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). RUMM 2020 [Computer software]. Perth: RUMM Laboratory.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing Impact, DIF, and DFF in Accommodated Item Scores: A Comparison of Multilevel Measurement Model Parameterizations. *Educational and Psychological Measurement, 72*(5), 754-773.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing Differential Testlet Functioning From Differential Bundle Functioning Using the Multilevel Measurement Model. *Educational and Psychological Measurement, 72*(2), 200-223.

Bolt, D. M. (2002). A Monte Carlo Comparison of Parametric and Nonparametric Polytomous DIF Detection Methods. *Applied Measurement in Education, 15*(2), 113-141.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44*.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153-168.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage Publications.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for Polytomously Scored Items: An Adaptation of the SIBTEST Procedure. *Journal of Educational Measurement, 33*(3), 333-353.

Chen, J.-H., Chen, C.-T., & Shih, C.-L. (2014). Improving the control of Type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement, 38*(1), 18-36.

Cheng, Y., Shao, C., & Lathrop, Q. N. (2016). The Mediated MIMIC Model for Understanding the Underlying Mechanism of DIF. *Educational and Psychological Measurement, 76*(1), 43-63.

Cheong, Y. F., & Kamata, A. (2013). Centering, scale indeterminacy, and differential item functioning detection in hierarchical generalized linear and generalized linear mixed models. *Applied Measurement in Education, 26*(4), 233-252.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of Differential Item Functioning in the Graded Response Model. *Applied Psychological Measurement, 17*(4), 335-350.

De Ayala, R. J. (2013). *The theory and practice of item response theory*: Guilford Publications.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: a generalized linear and nonlinear approach*. New York: Springer.

DeMars, C. E. (2008). Polytomous Differential Item Functioning and Violations of Ordering of the Expected Latent Trait by the Raw Score. *Educational and Psychological Measurement, 68*(3), 379-396.

DeMars, C. E. (2010). Type I Error Inflation for Detecting DIF in the Presence of Impact. *Educational and Psychological Measurement, 70*(6), 961-972.

Denny, F., Marshall, A. H., Stevenson, M. R., Hart, P. M., & Chakravarthy, U. (2007). Rasch analysis of the daily living tasks dependent on vision (DLTV). *Invest Ophthalmol Vis Sci, 48*(5), 1976-1982.

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

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*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: an application of the standardization approach1. *ETS Research Report Series, 1983*(1), i-14.

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

Dorans, N. J., & Kulick, E. (2006). Differential Item Functioning on the Mini-Mental State Examination: An Application of the Mantel-Haenszel and Standardization Procedures. *Medical Care, 44*(11), S107-S114.

Dorans, N. J., & Schmitt, A. P. (1991). Constructed Response and Differential Item Functioning: A Pragmatic Approach1. *ETS Research Report Series, 1991*(2), i-49.

Dougherty, B. E., & Bullimore, M. A. (2010). Comparison of scoring approaches for the NEI VFQ-25 in low vision. *Optometry and vision science: official publication of the American Academy of Optometry, 87*(8), 543.

Dougherty, B. E., Nichols, J. J., & Nichols, K. K. (2011). Rasch analysis of the ocular surface disease index (OSDI). *Invest Ophthalmol Vis Sci, 52*(12), 8630-8635.

du Toit, R. n. e., Palagyi, A., Ramke, J., Brian, G., & Lamoureux, E. L. (2008). Development and validation of a vision-specific quality-of-life questionnaire for Timor-Leste. *Invest Ophthalmol Vis Sci, 49*(10), 4284-4289.

Dye, D. C., Eakman, A. M., & Bolton, K. M. (2013). Assessing the Validity of the Dynamic Gait Index in a Balance Disorders Clinic: An Application of Rasch Analysis. *Physical Therapy, 93*(6), 809-818.

Eakman, A. M. (2012). Measurement characteristics of the Engagement in Meaningful Activities Survey in an age-diverse sample. *American Journal of Occupational Therapy, 66*(2), e20-e29.

Elosua, P., & Wells, C. (2013). Detecting DIF in Polytomous Items Using MACS, IRT and Ordinal Logistic Regression. *Psicologica: International Journal of Methodology and Experimental Psychology, 34*(2), 327.

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

Fidalgo, Á. M., & Bartram, D. (2010). A Comparison Between Some Generalized Mantel-Haenszel Statistics for Detecting DIF in Data Simulated Under the Graded Response Model. *Applied Psychological Measurement, 34*(8), 600-606.

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling, 18*(2), 229-252.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A Description and Demonstration of the Polytomous-DFIT Framework. *Applied Psychological Measurement, 23*(4), 309-326.

French, A. W., & Miller, T. R. (1996). Logistic Regression and Its Use in Detecting Differential Item Functioning in Polytomous Items. *Journal of Educational Measurement, 33*(3), 315-332.

French, B. F., & Finch, W. H. (2010). Hierarchical Logistic Regression: Accounting for Multilevel Data in DIF Detection. *Journal of Educational Measurement, 47*(3), 299-317.

French, B. F., & Finch, W. H. (2013). Extensions of Mantel–Haenszel for Multilevel DIF Detection. *Educational and Psychological Measurement, 73*(4), 648-671.

French, B. F., & Finch, W. H. (2015). Transforming SIBTEST to Account for Multilevel Data Structures. *Journal of Educational Measurement, 52*(2), 159-180.

Fukuhara, H., & Paek, I. (2015). Exploring the Utility of Logistic Mixed Modeling Approaches to Simultaneously Investigate Item and Testlet DIF on Testlet-based Data. *Journal of applied measurement, 17*(1), 79-90.

González, A. V. B., Sierra, C. M. T., Martínez, A. B., Martínez-Molina, A., & Ponce, F. P. (2015). An in-depth psychometric analysis of the Connor-Davidson Resilience Scale: calibration with Rasch-Andrich model. *Health and Quality of Life Outcomes, 13*(1), 154.

Gothwal, V. K., Reddy, S. P., Sumalini, R., Bharani, S., & Bagga, D. K. (2012). National Eye Institute Visual Function Questionnaire or Indian Vision Function Questionnaire for Visually Impaired: A ConundrumAssessing Visual Functioning in Visually Impaired. *Invest Ophthalmol Vis Sci, 53*(8), 4730-4738.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*(3), 244-253.

Hidalgo, M. D., & Gómez, J. (2006). Nonuniform DIF Detection using Discriminant Logistic Analysis and Multinomial Logistic Regression: A comparison for polytomous items. *Quality & Quantity, 40*(5), 805-823.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of applied measurement, 6*(3), 311.

Jiao, H., & Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology, 68*(1), 65-83.

Jin, Y., Myers, N. D., & Ahn, S. (2014). Complex Versus Simple Modeling for DIF Detection: When the Intraclass Correlation Coefficient (ρ) of the Studied Item Is Less Than the ρ of the Total Score. *Educational and Psychological Measurement, 74*(1), 163-190.

Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement, 38*(1), 79-93.

Kim, S.-H., & Cohen, A. S. (1998). Detection of Differential Item Functioning Under the Graded Response Model With the Likelihood Ratio Test. *Applied Psychological Measurement, 22*(4), 345-355.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*(1), 15-32.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A Comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Items. *Educational and Psychological Measurement, 65*(6), 935-953.

Lai, J.-S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation and the Health Professions, 28*(3), 283-294.

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

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

Lundström, M., & Pesudovs, K. (2009). Catquest-9SF patient outcomes questionnaire. Nine-item short-form Rasch-scaled revision of the Catquest questionnaire. *Journal of Cataract and Refractive Surgery, 35*(3), 504-513.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *22*(4), 719-748.

Massof, R. W. (2005). Application of Stochastic Measurement Models to Visual Function Rating Scale Questionnaires. *Ophthalmic Epidemiology, 12*(2), 103-124.

Massof, R. W. (2007). An interval-scaled scoring algorithm for visual function questionnaires. *Optometry & Vision Science, 84*(8), E689-E705.

Massof, R. W. (2011). Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. *Ophthalmic Epidemiology, 18*(1), 1-19.

Massof, R. W., Deremeik, J. T., Park, W. L., & Grover, L. L. (2007). Self-reported importance and difficulty of driving in a low-vision clinic population. *Invest Ophthalmol Vis Sci, 48*(11), 4955-4962.

Massof, R. W., Hsu, C. T., Baker, F. H., Barnett, G. D., Park, W. L., Deremeik, J. T., . . . Epstein, C. (2005). Visual disability variables. II: The difficulty of tasks for a sample of low-vision patients. *Archives of physical medicine and rehabilitation, 86*(5), 954-967.

McHorney, C. A., & Fleishman, J. A. (2006). Assessing and understanding measurement equivalence in health outcome measures. *Medical Care, 44*.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo Examination of the Sensitivity of the Differential Functioning of Items and Tests Framework for Tests of Measurement Invariance With Likert Data. *Applied Psychological Measurement, 31*(5), 430-455.

Mellenbergh, G. J. (1982). Contingency Table Models for Assessing Item Bias. *Journal of Educational Statistics, 7*(2), 105-118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127-143.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289-311.

Miller, T. R., & Spray, J. A. (1993). Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items. *Journal of Educational Measurement, 30*(2), 107-122.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*(1), 59-71.

Paek, I., & Fukuhara, H. (2015). Estimating a DIF decomposition model using a random-weights linear logistic test model approach. *Behavior research methods, 47*(3), 890-901.

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement, 45*(3), 247-269.

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

Pesudovs, K. (2010). Item banking: a generational change in patient-reported outcome measurement. *Optometry & Vision Science, 87*(4), 285-293.

Pesudovs, K., Gothwal, V. K., Wright, T., & Lamoureux, E. L. (2010). Remediating serious flaws in the National Eye Institute Visual Function Questionnaire. *Journal of Cataract & Refractive Surgery, 36*(5), 718-732.

Potenza, M. T., & Dorans, N. J. (1995). DIF Assessment for Polytomously Scored Items: A Framework for Classification and Evaluation. *Applied Psychological Measurement, 19*(1), 23-37.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1): Sage.

Ravand, H. (2015). Item Response Theory Using Hierarchical Generalized Linear Models. *Practical Assessment, Research & Evaluation, 20*(7), 2.

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.

Rovner, B. W., Casten, R. J., Hegel, M. T., Massof, R. W., Leiby, B. E., & Tasman, W. S. (2011). Improving function in age-related macular degeneration: design and methods of a randomized clinical trial. *Contemporary clinical trials, 32*(2), 196-203.

Ryan, C. H. (2008). *Using hierarchical generalized linear modeling for detection of differential item functioning in a polytomous item response theory framework: An evaluation and comparison with generalized Mantel-Haenszel.* (Dissertation), Georgia State University.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., . . . Sprangers, M. A. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes, 8*(1), 81.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., . . . Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology, 62*(3), 288-295.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*(3), 184-199.

Somes, G. W. (1986). The Generalized Mantel-Haenszel Statistic. *The American Statistician, 40*(2), 106-108.

Stelmack, J. A., Szlyk, J. P., Stelmack, T. R., Demers-Turco, P., Williams, R. T., Moran, D. A., & Massof, R. W. (2004). Psychometric Properties of the Veterans Affairs Low-Vision Visual Functioning Questionnaire. *Invest Ophthalmol Vis Sci, 45*(11), 3919-3928.

Su, Y. H., & Wang, W. C. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and Logistic Discriminant Function Analysis Methods in Detecting Differential Item Functioning for Polytomous Items. *Applied Measurement in Education, 18*(4), 313-350.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of Differential Item Functioning (DIF) Using Hierarchical Logistic Regression Models. *Journal of Educational and Behavioral Statistics, 27*(1), 53-75.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care, 44*.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond Group-Mean Differences: The Concept of Item Bias. *Psychological Bulletin, 99*(1), 118-128.

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

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models. *Journal of Educational and Behavioral Statistics, 30*(4), 443-464.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-Classification Multilevel Logistic Models in Psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369-386.

Velozo, C. A., Warren, M., Hicks, E., & Berger, K. A. (2013). Generating Clinical Outputs for Self-Reports of Visual Functioning. *Optometry and Vision Science, 90*(8), 765-775.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education, 72*(3), 221-261.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-Free-Then-DIF Strategy for the Assessment of Differential Item Functioning. *Educational and Psychological Measurement, 72*(4), 687-708.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*(2), 113-144.

Wang, W.-C., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement, 65*(4), 549-576.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479-498.

Williams, S., Brian, G., & Toit, R. d. (2012). Measuring vision-specific quality of life among adults in Fiji. *Ophthalmic Epidemiology, 19*(6), 388-395.

WINSTEPS. (2009). [Computer software] (Version Version 3.70.0.3). Chicago, IL.

Wolle, M. A., Cassard, S. D., Gower, E. W., Munoz, B. E., Wang, J., Alemayehu, W., & West, S. K. (2011). Impact of Trichiasis surgery on physical functioning in Ethiopian patients: STAR trial. *American journal of ophthalmology, 151*(5), 850-857.

Wood, S. W. (2011). *Differential item functioning procedures for polytomous items when examinee sample sizes are small*. (Dissertation), The University of Iowa.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*(1), 42-57.

Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH Tests, and IRT-LR-DIF when the latent distribution is nonnormal for both Groups. *Applied Psychological Measurement, 35*(2), 145-164.

Woods, C. M., & Grimm, K. J. (2011). Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models. *Applied Psychological Measurement, 35*(5), 339-361.

Xie, C. (2014). *Cross-classified modeling of dual local item dependence.* (Dissertation), University of Maryland, College Park.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145.

Zwick, R., & Thayer, D. T. (1996). Evaluating the Magnitude of Differential Item Functioning in Polytomous Items. *Journal of Educational and Behavioral Statistics, 21*(3), 187-201.

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