EFFECTS OF DIFFERENTIAL ITEM FUNCTIONING COMMON ITEMS ON NONEQUIVALENT GROUPS DESIGN LINKING

A Dissertation

by

HUAN WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee, Myeongsun Yoon
Committee Members, Lei-Shih Chen
Wen Luo
Oi-Man Kwok
Head of Department, Shanna Hagan-Burke

December 2017

Major Subject: Educational Psychology

ABSTRACT

Under common-item nonequivalent groups design linking, the functionality of common items as bridge between two parallel forms entails strict content and statistical restrictions on these items. Despite representativeness to the whole test, and similar positions of item placement on each of the parallel forms, common items could present differential item functioning (DIF) effects between test takers of different forms, especially when the groups of test takers are not equivalent in ability profiles. DIF common items under such scenario should impair the adequacy of linking if they were used as linking items instead of being examined and then taken care of. However, the menace of DIF common items on linking has not been substantiated by research yet.

In the current study, I have reviewed the related literature in item response theory, equating, and differential item functioning with emphases on linking methods, forms of DIF, and DIF detection methods. Responding to the scarcity of research on DIF common-item effects on linking, a series of Monte Carlo simulation studies were conducted under common-item nonequivalent groups design linking, testing potential influential factors in empirical research, i.e., sample size, ratio of common items, ratio of DIF items, magnitude of DIF, form of DIF, and direction of DIF. Recovery of equating slope *A* and equating intercept *B,* and item discrimination *a* and item location *b* was evaluated using signed bias and root mean square error (RMSE).

My results show that generally as sample size went up, the bias and RMSE went down, an effect tended to level off at 1000 participants in each group. The number of DIF

common items and the magnitude of uniform DIF items were testified as more influential factors than number of common items and the direction of DIF. As the number of DIF common items increased, and/or the magnitude of uniform DIF increased, the bias and RMSE increased quickly. Bias and RMSE of equating intercept *B* was mostly related to the uniform DIF common items against or in favor of group 2 test takers, while bias and RMSE of equating slope *A* was mostly related to the nonuniform DIF common items against or in favor of group 2 test takers. Only *B* was seriously biased when having uniform DIF. Both *B* and *A* were seriously biased when having uniform and nonuniform DIF at the same time. Overall, the mean bias and mean RMSE of item discrimination *a*, and the mean bias and mean RMSE of item location *b* were small on most simulation conditions. Within common items, the mean bias and mean RMSE of item discrimination *a*, and the mean bias and mean RMSE of item location *b* were sensitive to simulation condition changes. Results were canvassed and limitations were pointed out at the end of this dissertation with recommendations for future research.

DEDICATION

To my husband Wen, my daughter Megan, and my mom and dad.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor and chair, Myeongsun Yoon, for her support and guidance throughout my graduate studies at Texas A&M University at College Station. Without her guidance, this dissertation would not be completed. I would also like to thank to my committee members, Oi-Man Kwok, Wen Luo, and Lei-Shih Chen, for their helpful comments on this project.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Myeongsun Yoon [advisor], Oi-Man Kwok, and Wen Luo of the Department of Educational Psychology and Professor Lei-Shih Chen of Department of Health and Kinesiology.

All work conducted for the dissertation was completed by the student independently.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

1. INTRODUCTION

Test equating is one of the constant challenges that exist in the process of assigning numbers (i.e., test scores) to individuals to represent their certain trait or characteristic and using test scores for decision-making. Almost all types of test scores are of concern, e.g., norm-referenced scores used for school placement and grade advancement, such as SAT, ACT, and GRE, and criterion-referenced scores used for licensing purpose, such as driver's license exam, medical licensing examination, and a bar examination in law. Testing practitioners, educational researchers, and policy makers should take cautions when reporting, analyzing, and making decisions with test scores coming from different sittings or forms among different groups of test takers. The ideal situations include 1) multiple groups of people taking exactly the same form of a test; and 2) the same group of people repeatedly taking different forms of the same test. For the first situation, the observed mean differences are indicators of true ability differences among groups. Test scores could be compared directly without transformation or adjustment. In the second, the observed mean differences originate as test forms are different in difficulty. Testing scores between two forms could be compared after adjusting for the mean difference. However, it is not uncommon that different groups of people take alternate forms of the same test for the same purpose on different dates or in different terms of the examination due to test security concerns. Under such circumstances, meaningful comparisons are not achievable without test equating.

Equating existed as a statistical procedure to adjust for reasonable amount of difference in testing difficulties between alternate forms that are constructed under the same construct, content, and statistical specifications (Kolen & Brennan, 2013). Without adequate equating, some test takers could be advantaged sitting in an easy form of a test, while others could be disadvantaged sitting in an alternate form of the same test that is difficult. Equating is leverage to fairness in psychological measurement. Another leverage to measurement fairness is the testing of measurement invariance, the process of identifying measurement bias and purifying the test items to build an unbiased instrument. Even though acknowledged as two aspects that could impair testing fairness when not handled properly, research in two areas are not going hand in hand. Seldom has associated the two areas at the same time to explore the effect of measurement noninvariance or differential item functioning (DIF) on linking/equating results.

Measurement invariance has long been tested within the multiple group confirmatory factor analysis framework for factorial invariance (Reise, Widaman, & Pugh, 1993). Within IRT framework, Likelihood ratio test, Wald statistics, Mantel-Haenszel statistics, Raju area statistics, differential functions of items and tests (DFIT), and SIBTEST have been used and studied in item bias detection (Embretson & Reise, 2000; Millsap & Everson, 1993; Roussos & Stout, 1996). The evaluative studies with simulations generally favored DFIT method. Item bias detection methods are applied with the assumption that parameter estimates on different forms are on the same scale. When groups of test takers are not randomly equivalent and separate estimation is performed, linking step is required to put parameter estimates from different forms on the same scale.

Hanson and Beguin (1999) found it was even beneficial to perform linking and parameter scaling with randomly equivalent groups. Commonly used parameter linking methods include mean/mean method, mean/sigma method, item characteristic curve method, test characteristic curve method, minimum chi-square method (Divgi, 1985; Haebara, 1980; Loyd & Hoover, 1980; Marco, 1977; Stocking & Lord, 1983). Evaluative studies have emerged and favored toward the test characteristic curve method. However, adequate linking and item bias detection are not separate procedures but dependent on each other. The isolation of linking methods to DIF detection methods will still result in inefficiency altogether. Therefore, several two-stage iterative procedures were proposed by researchers to take care of the two problems at the same time (Lord, 1980; Marco, 1977; Park, 1988; Park & Lautenschlager, 1990; Segall, 1983).

Past research was performed and recommendations were generated on how to do linking in the context of item bias detection, but few studies investigated effects of item bias on linking and equating results when alternate forms were used for testing. Or rather, most studies in literature were based on different groups taking exactly the same form of a test, and equating was not of a concern. Bowles (2016) had pointed out that measurement variance or DIF should be included as important topics in the test equating illustration considering the measurement invariance and test equating are closely related to each other in the process of detection and linking.

Limited number of studies in literature targeted measurement invariance or DIF and test equating at the same time. And they were not without limitations. Kim and Cohen (1992) compared three linking methods under both iterative and noniterative DIF

detection procedures and concluded that linking with test characteristic method was most accurate in flagging misbehaved items even when the sample size was small. However, the DIF effects on linking procedures were not the focus of the study. Chu and Kamata (2000; 2004) proposed and tested a multilevel IRT model that handles and controls DIF effects on equating. The focus was the performance of the multilevel IRT model when DIF presented compared with traditional single level IRT model. Still the DIF effects on linking and equating were not investigated and compared in details. Huggins (2014) has investigated the impact of DIF on the property of population invariance of equating and concluded that the population invariance property could be jeopardized when anchor items display DIF. Kabasakala and Kelecioğlub (2015) have investigated the effect of DIF items on equating under both traditional IRT and multilevel IRT models with varied magnitudes of DIF and different placements of DIF. However, the study was conducted under the common items equivalent groups design.

In addition to the fact that limited number of studies investigated DIF effects on linking and equating in details, another characteristic of existed studies was the origins of DIF. Among all studies mentioned in the previous paragraph, DIF happened between a focal and reference groups, e.g., subpopulations in test score reporting, gender, or other demographic variable. Since DIF was not related to the study design or forms of the test, both focal and reference groups still took the same set of items. I will scrutinize the effects of DIF on linking when DIF happens due to nonequivalent abilities in two groups. Under common items nonequivalent groups design and the nature of DIF I will explore, the bias is only possible within the common items on different forms. Two groups of test takers no

longer take the same items as previous studies. Examples of DIF due to group ability difference could be a grade based subject test, like a math or a reading comprehension test. Common items are placed on tests for adjacent grades, e.g., grade 3 and grade 4 sharing common items, and grade 4 and grade 5 sharing the same set or another set of common items, a situation that is not uncommon in educational assessment.

There are several reasons to hypothesize the relationship of DIF and linking under common items nonequivalent groups data collection design. First, common-items in equating designs are required to be miniature of the whole test especially when group ability difference presents (Cook & Petersen, 1987). When DIF presents, the representative of those common items should be impaired. Consequently, assumption of the equating design is violated and equating adequacy will be compromised. Second, the deduction of linking constant, A and B, no matter through moment methods or test characteristic curve methods, should be affected when DIF appears in the common items. In a more specific way, the DIF items will have different item parameters and the moments of the common items or the item characteristic curves will be affected directly. Since the characteristic curve methods used the raw information from each biased item directly, linking functions obtained from characteristic curve methods have been assumed to contain more bias. Third, the item bias could happen in a subtler way. Even though, the common items are well established in the previous test developing, linking, and equating procedures, the item parameters could change in the long run. That is the item might become dated to the current test takers. When equating the new form to the old form, the adequacy of the equating could be comprised due to dated and biased items. Also, within

the common-item nonequivalent groups design or when vertical equating is under concern, some of the common items could have different item performance due to the difference in test difficulty and the difference between group abilities.

Considering the current status of literature on DIF, liking, and equating studies, it is necessary to explore the DIF effects on linking and item parameter recovery in details. Given various reasons that DIF items could present within common items nonequivalent groups design due to group ability difference, it is also important to delineate the impact of DIF items on equating coefficients and item parameter recovery. I will address the related issues in this study.

2. LITERATURE REVIEW

2.1. Overview of Item Response Theory

The theoretical and applicable development of item response theory (IRT) followed two lines, one represented by Lord, Novick, and Birnbaum with the publishing of Statistical Theories of Mental Test Scores (Lord, Novick, & Birnbaum, 1968), and the other characterized by Rasch and his book Probabilistic Models for some Intelligence and Attainment Tests (Rasch, 1960). However, up until the 1970s and 1980s, the basics and skills of IRT were new or remain unknown to most of psychological practitioners (Embretson & Reise, 2000).

In general, the IRT model is a logistic function with bounded area between 0 and 1. Depending on the type of the data collected, binary or Likert type scale, IRT models can be divided into dichotomous IRT models and polytomous IRT models. Based on the dimensions of the data, IRT models can have unidimensional models and multidimensional models. In the current research, only unidimensional models are discussed.

2.1.1. Assumptions

Three assumptions are made when using unidimensional IRT modeling, i.e., unidimensionality, local independence, and form of item response function. Unidimensionality requires only one latent variable underlies the response to all items, or the test only measures one latent variable. The unidimensionality seems to restrict on test development stage. However, upon responding to the items developed to measure a single

latent variable, a lot of other factors are concerted to give the best answer. For example, in working out a solution toward a mathematic problem, the reading comprehension and the special imagination ability may also come into cooperation. In most cases, the unidimensionality is reduced to have a principal single factor underlying the response to items. Using collected data and fitting a unidimensional IRT model, unless the assumption of unidimensionality is severely violated, the model still provides useful information about item parameters and latent variable.

Local independence is satisfied if the joint probability of p number of item response pattern is equal to the multiplication product of the probability of each item response given a latent variable, which could be expressed as

$$P(X_1, X_2, \dots, X_p|\theta) = \prod_{j=1}^{p} P(X_j|\theta). \quad (1)$$

Local independence states that with the test targeted latent variable taking into account, the responses to any pair of items are not associated. Local independence does not indicate that a person's responses to items are not correlated, but are all accounted for by the latent variable to be measured. This is very similar to the idea in the factor analysis. If only one latent factor is behind all measured variables, the residual variance of the all measured variables are not correlated after accounting of the one common factor. The local independence confirms that responses to all items are determined by one latent variable, which echoes the assumption of unidimensionality. However, local independence can be obtained even when the data is multidimensional. As a necessary condition for unidimensionality, the local independence will hold when unidimensionality assumption is met.

The IRT modeling also requires the shape of item response function (IRF) or the item characteristic curve (ICC) to be a monotonically increasing function. As the latent variable θ is increasing, the probability of getting form the IRF is increasing, indicating the probability of passing an item is increasing, given the item parameters. A typical three parameters logistic model is written as

$$P\big((X_{ij}=1\big|\theta_i\big)=c_j+\big(1-c_j\big)\frac{exp\big[a_j\big(\theta_i-b_j\big)\big]}{1+exp\big[a_j\big(\theta_i-b_j\big)\big]}, \qquad (2)$$

Where X_{ij} is the item response of i_{th} person on j_{th} item, θ_i is the i_{th} person latent variable score, and a_j, b_j, c_j are the j_{th} item discrimination, difficulty, and pseudo-guessing parameter, respectively. If $\theta_1<\theta_2$, $P\big((X_{1j}=1\big|\theta_1\big)<P\big((X_{2j}=1\big|\theta_2\big)$. The shape of the response function is curvilinear with a bounded area between 0 and

2.1.2. Dichotomous IRT Models

When the collected data is binary type or scale that is binned to binary type for analysis, a set of dichotomous IRT models available for model fitting, including Rasch model, one parameter logistic model (1PL), two parameters logistic model (2PL), and three parameters logistic model (3PL).

Rasch model is a special case of 1PL model. The distance between person ability and the item location (difficulty) will predict the probability of item response. Rasch model is written as

$$P\big(X_{ij}=1\big|\theta_i\big)=\frac{exp\big(\theta_i-b_j\big)}{1+exp\big(\theta_i-b_j\big)}=\frac{1}{1+exp\big[-\big(\theta_i-b_j\big)\big]}. \qquad (3)$$

Person ability or latent variable score θ_i and item location b_j, also known as item difficulty are put on the same latent continuum, the range for which is $(-\infty, +\infty)$. Putting on z score scale, the range of latent variable score θ_i and item location b_j is roughly within $(-3, +3)$. The probability of passing or endorsing an item is changing with the distance between latent variable score and item location.

When $\theta_i = b_j$, P=0.5;

When $\theta_i > b_j$, P>0.5;

When $\theta_i < b_j$, P<0.5.

The item location b_j is estimated at the latent variable score that has the probability of 0.5 of passing/endorsing the item.

The general 1 PL model is written as having an additional constant item discrimination parameter a to the Rasch model, which is

$$P(X_{ij} = 1|\theta_i) = \frac{exp[a(\theta_i - b_j)]}{1 + exp[a(\theta_i - b_j)]} = \frac{1}{1 + exp[-a(\theta_i - b_j)]}. \qquad (4)$$

When $a = 1$, 1PL model is reduced to Rasch model. Using Rasch or 1PL model, each item is equally important in determining the item response in terms of distance between person ability and item location. The unweighted sum of item scores is sufficient statistics for estimating θ_i.

Two parameters logistic model allows the items differ in both location and discrimination. The expression of the 2PL model is written as

$$P(X_{ij} = 1|\theta_i) = \frac{exp[a_j(\theta_i - b_j)]}{1 + exp[a_j(\theta_i - b_j)]} = \frac{1}{1 + exp[-a_j(\theta_i - b_j)]}. \qquad (5)$$

The parameter a_j is added to represent the item discrimination of j_{th} item. According to the function, the distance between person ability and item location $(\theta_i - b_j)$ is weighted by a_j in determining the probability of passing/endorsing an item. Consequently, the weighted sum of item scores is the sufficient statistics for estimating θ_i. The item location b_j is still estimated at the point where the latent variable score having a probability of 0.5 of passing/endorsing an item. The range of a_j is within $(0, 2)$. Based on the shape and monotonicity of IRF, the item discrimination cannot be 0 or a negative value. If $a_j = 0$, the item response function will give a constant probability of 0.5 regardless of the distance between person ability and item location. The ICC will be a straight line at 0.5 probability across the x-axis continuum for person ability. If $a_j < 0$, the IRF will be monotonically decreasing as the person ability increases. The larger the item discrimination is, the larger the difference in probabilities of passing/endorsing an item with the same amount of distance between person ability and item location.

Three parameters logistic model admits that guessing is a factor influence the item response when the person ability is very low. By adding the guessing parameter to the 2PL model, the IRF of a 3PL model is expressed in equation 2. The probability of passing/endorsing an item is the sum of two probabilities, i.e., the probability of guessing and the probability without guessing. c_j is the pseudo-guessing parameter. While the real guesting parameter is difficult to know, the estimated value of c_j usually is smaller than the random guessing probability. The item location is not estimated at the latent variable score where the probability of passing/endorsing an item is 0.5 with guessing, but the

latent variable score where the probability of passing/endorsing an item is 0.5 without guessing. When $\theta_i = b_j$, the $P\big((X_{ij} = 1|\theta_i\big) = c_j + \big(1 - c_j\big) \times 0.5 = 0.5 + 0.5 \times c_j$. In 3PL model, the probability of passing/endorsing an item is larger than 0.5 when $\theta_i = b_j$.

2.1.3. Polytomous IRT Models

In addition to dichotomous data and binary IRT models, polytomous data are often used in psychology measurement, especially those measuring person's attitude, endorsement, and personality. An example is the Likert type scale which might has the number of 1 standing for strongly disagree, 2 for disagree, 3 for neutral, 4 for agree, and 5 for strongly agree. Binary IRT models can be fitted to those data with more than two categories if the data are binned into two categories according to certain cutoff values. However, upon reducing the data into a lower level, some of the information (variance) in the data is lost. Instead of stick to binary IRT models, researchers have developed IRT models to deal with polytomous data type. Samejima (1970) has proposed the graded response model, which is a more general form of 2PL model in terms of item locations. Other polytomous models include partial credit model, generalized partial credit model, rating scale model, and nominal response model (Andrich, 1978a, 1978b; Bock, 1972; Masters, 1982; Muraki, 1992, 1993, 1997). Categorical response function (CRF) is an important concept in understanding polytomous IRT model. Let's suppose an item has $m + 1$ response options. A specific category of item response is represented by k, and $k = 0,1,2 \ldots m$. The probability of passing/endorsing a particular response option k is defined as the categorical response probability, and the CRF is written as,

$$P\big(X_{ij} = k|\theta_i\big). \qquad (6)$$

Categorical response function in polytomous data is not monotonic in most cases. It is intuitive that as the person's ability is increasing, the probability of endorsing the lower response category decreases. For some intermediate response options, the probability of endorsing the item might be increasing as the person's ability increases, but the probability of endorsing the item could also decreases as the person's ability increases given the person's ability is higher enough.

Graded response model (GRM) gives the cumulative probability of passing/endorsing the k_{th} category or higher by modifying the item location parameter into threshold parameters. The general expression of GRM is written as,

$$P(X_{ij} \geq k|\theta_i) = \frac{exp[a_j(\theta_i - b_{jk})]}{1 + exp[a_j(\theta_i - b_{jk})]}, \quad (7)$$

where $k = 0, 1, 2, \ldots, m$ are $m + 1$ response options of item j. The categorical response function is the difference between the cumulative probabilities of two adjacent categories in GRM. The categorical response function is written as,

$$P(X_{ij} = k|\theta_i) = P(X_{ij} \geq k|\theta_i) - P(X_{ij} \geq k + 1|\theta_i)$$

$$= \frac{exp[a_j(\theta_i - b_{jk})]}{1 + exp[a_j(\theta_i - b_{jk})]} - \frac{exp[a_j(\theta_i - b_{j(k+1)})]}{1 + exp[a_j(\theta_i - b_{j(k+1)})]}. \quad (8)$$

The discrimination parameter a_j indicates the steepness of the ICCs or how the categorical response function peaks, narrowly steep or widely flat. Parameter b_{jk} is the difficulty parameter of transitioning from one lower category to the adjacent higher category. For $k = 1$ and $k = m$, b_{j1} and b_{jm} are the latent ability points when the probabilities of passing/endorsing an item are 0.5 on the lowest category and highest category,

respectively. For the intermediate category, b_{jk} is the point of the peak of the corresponding response category. If an item has $m+1$ response categories, there will be $m+1$ categorical response curves and m threshold parameters. GRM does not have requirement on the same number of response options within a test. Items can have different formats in terms of response categories.

Partial credit model (PCM) is another polytomous IRT model when the partial credit is desired to be given to those who have finished some intermediate steps or on the medium levels of an aptitude test. Following the same assumption of having $m+1$ response options, there will be m-1 possible partial credits could be given in addition to 0 for no credit and m for full credit. In partial credit model, all items are assumed to be equally discriminating which is the same as the Rasch model. The general expression of IRF for partial credit model of passing/endorsing the k_{th} category conditioning on completing $(k-1)_{th}$ category correctly, is written as,

$$P\left(\left(X_{ij}=k\middle|\theta_i, X=k-1\ \right)\right)=\frac{P\left(X_{ij}=k\middle|\theta_i\right)}{P\left(X_{ij}=k-1\middle|\theta_i\right)+P\left(X_{ij}=k\middle|\theta_i\right)}$$

$$=\frac{exp\left(\theta_i-b_{jk}\right)}{1+exp\left(\theta_i-b_{jk}\right)}. \quad (9)$$

The category response function, also known as the unconditional response function for each response option under PCM is written as,

$$P\left(X_{ij}=k\middle|\theta_i\right)=\frac{exp\left[\sum_{c=1}^{k}\left(\theta_i-b_{jc}\right)\right]}{1+\sum_{k=1}^{m}exp\left[\sum_{c=1}^{k}\left(\theta_i-b_{jc}\right)\right]}, \quad (10)$$

where $k=1,2,3,\ldots,m$, and

$$P(X_{ij} = 0|\theta_i) = \frac{1}{1 + \sum_{k=1}^{m} exp\left[\sum_{c=1}^{k}(\theta_i - b_{jc})\right]}, \qquad (11)$$

when k=0. The step difficulty b_{jk} is the point of latent variable score where two adjacent response curves intersect. If an item has five response options, there will be four step difficulties, b_{j1} is the step difficulty intersecting at $k = 0$ and $k = 1$, b_{j2} is the step difficulty intersecting at $k = 1$ and $k = 2$, and so on.

Another commonly used polytomous IRT model is rating scale model (RSM), which is a modified PCM. The step difficulty in the PCM is divided into two parts, the base item location and the relative difficulty of each step across all items. The model is obtained by replacing the b_{jk} with $b_{jk} = \gamma_j + \delta_k$ in the PCM for both conditional probability function and categorical response functions. For example, item 1 has five response categories, and the base item location is 1.66 (γ_j). The four step difficulties are -0.34 (δ_1), -0.05(δ_2), 0.67(δ_3), 0.83(δ_4). Item 2 also has five response options with $\gamma_j = 0.78$, while item 1 and item 2 will have the same δ_k. Hence, the four step difficulties for item 2 are also -0.34 (δ_1), -0.05(δ_2), 0.67(δ_3), 0.83(δ_4).

Other polytomous IRT models included the generalized PCM, generalized PCM with rating scale for step difficulty, GRM with rating scale for step difficulty. Due to the similarity in response function, these modified models will not be introduced in detail.

2.2. Overview of Equating

Equating is of concern when testing scores on parallel or matched forms of a test from two groups are compared. Equating exists as a statistical method and a procedure to adjust for difficulty difference between test forms while the forms are constructed to the

same content specifications and statistical restrictions (Kolen & Brennan, 2013). Adequate equating is possible when several requirements are met. The test forms are parallel in terms of content (unidimensional), difficulty, validity and reliability of scores (Cook & Petersen, 1987; Harris & Crouse, 1993).

Two types of equating are generally used for different data collection designs, i.e., horizontal equating and vertical equating. Horizontal equating is applied to equate scores on similar forms of a test with equivalent groups of test takers, while vertical equating is used to equate scores on alternate forms of a test with nonequivalent groups of test takers (Loyd & Hoover, 1980). Common items are usually placed in adjacent forms or among alternate forms in vertical equating. For example, a reading comprehension test may have alternate forms for grade 3 through 6. For each grade there are items are grade unique in terms of curriculum instructions, and there are common items placed between adjacent forms, like grade 3 and grade 4, and nonadjacent forms, like grade 3 and grade 5. Common items are used as linkage between forms. The design with common items is especially advantageous when using IRT modeling due to the population independent quality of item parameter estimates. The moments of item parameter estimates or the item response characteristics for the same items from different groups are used to build the linking functions between forms.

2.2.1. Equating Property

Adequate equating has several properties. Based on these properties, equating methods are developed, e.g., linear method, equipercentile method. After equating being done, these properties can also be used evaluative criteria.

The symmetry (exchangeable) property of equating requires that the equation that is used to find a score on form *T* (target form) that is equivalent to a specified score on form *S* (base form) (e.g., 85 is the form *T* equivalent score of 88 on form *S*) is the reverse function of getting a score on form *S* that is equivalent to a specified score on from *T*. The symmetry property could be expressed as

$$S^* = f_{T\to S}(X) = f_{S\to T}^{-1}(X), \quad (12)$$

AND

$$T^* = f_{S\to T}(Y) = f_{T\to S}^{-1}(Y), \quad (13)$$

Where $f_{S\to T}$ is the function of equating form *S* scores Y to the form *T* scale, and $f_{T\to S}$ is the function of equating form *T* scores X to the form *S* scale. *X* is the random variable score on form *T*, and *Y* is a random variable score on form *S*. X^* is the equated random variable score of *Y*. Y^* is the equated random variable score of *X*.

The equity property requires the mean, standard deviation, and the distributional shape of the scores that is equated from form *T* scale to the form *S* scale is the same as the scores originally on form *S*, conditioning on that the two groups of examinees have the same true score, or at least the mean of the true score is equal (Lord, 1980). Linear methods, mean and linear equating, are based on this assumption. The equity property is presented as

$$S^*[y^* = (f_{T\to S}(x)|\tau)] = S(y|\tau). \quad (14)$$

τ is the true ability level for all examinees; S is the cumulative distribution of form *S* scores; and S^* is the cumulative distribution on form *S* scale for all equated form *T* scores. x is a specific score of random variable score *X* on form *T*, and y is a specific score

of random variable score Y on form S. y^* is the equated score of x on form S scale. The Lord's equity property is restrictive. A relaxed version of equity property is proposed by Morris (1982), which requires the expectation of the equated scores is the same as the expected values of the scores on the base scale, i.e.,

$$E[Y^* = (f_{T\to S}(X)|\tau)] = E[(Y|\tau)], \quad (15)$$

where τ is the true score of all examinees. Y^* is the equated random variable score of random variable X.

Observed score equity property just requires the converted scores of form T has the same distribution as scores on from S without condition on examinees true scores. This assumption is applied when using equipercentile method. The observed score equity property is presented as

$$S^*[y^* = f_{T\to S}(x)] = S(y), \quad (16)$$

Where S and S^* are the same as defined in Lord's equity property.

Adequate equating is also group invariant in terms of the equating relationship. The equating function should match well no matter using the subpopulation of examinees or the whole examinees, or using the two subpopulations of the examinees (Cook & Petersen, 1987; Dorans & Holland, 2000; Dorans, Liu, & Hammond, 2008; Kolen, 2004; Petersen, 2007, 2008; Yi, Assessment, Harris, & Gao, 2008). In other words, the equating function should be population independent. Let G represents the group membership, group invariance property could be expressed as

$$[f_{T\to S}(X)|G] = f_{T\to S}(X), \quad (17)$$

OR

$$[f_{S\rightarrow T}(Y)|G] = f_{S\rightarrow T}(Y). \qquad (18)$$

2.2.2. Common Items Nonequivalent Groups Design

The most commonly used equating designs include random groups design, counterbalanced single group design, and common-item nonequivalent groups design. Random groups design is two randomly selected groups (G1 and G2) assigned to finish from *T* and form *S*, respectively. Form *T* and form *S* are parallel forms of a test. Counterbalanced single group design involves two random groups, G1 and G2. G1 is assigned to take form *T* first, then form *S*, while G2 is assigned to take form *S* first, then form *T*. Common items nonequivalent groups design involves using anchor items in both form *T* and form *S*, usually accounting for 20% of the total items in an alternate form. Two groups of people, not required to be randomly selected are taking form *T* and form *S*, respectively. The current study is performed under the common items nonequivalent groups design.

It is a design of anchoring with the same items in different groups. When common/anchor items are used as an external set, scores on them are not included as test scores. When common/anchor items are used as an internal set, scores are included as test scores. The common/anchor items bridge different forms. In traditional equating, moments of the common items are used to synthesize and then equating the scores of two groups. In IRT modeling, item parameter estimates from common items are used to calculate equating/transforming constants A and B. Adequate equating is largely determined by the characteristics of the common items. Therefore, common items should have the following qualities. First, common items are representative to the construct and

difficulty of the entire test. Cook and Petersen (1987) stated that the common items should be a miniature of the whole test. Second, the placement of common items in different forms should be arranged roughly the same to avoid item location effect. Sometimes, common items can be held out as separate testing session. Third, common items respond the same to the different groups of test taker, i.e., measurement invariance. If a common item has different response functions under different groups of test takers, item bias on that item might present.

2.2.3. Traditional Equating Methods

Traditional equating methods include mean, linear, and equpercentile equating. Mean equating is obtained that each score point on form *T* is adjusted for the unsigned distance in means between form *T* and form *S*, or each score point on form *S* is adjusted for the unsigned distance in means between form *T* and form *S*. The unsigned distance is added to scores from the lower mean group, and subtracted from the higher mean group.

Using linear equating, the standardized normal scores (i.e., z scores) are set to equal on parallel forms. Under linear equating function, based on the known group means and variances, the score level on either form *T* or form *S* is specified first. Then equivalent score on form *S* or form *T* can be deducted.

Equipercentile equating (e.g., finding the equivalent score of form *T* on form *S*) is achieved by finding scores on form *T* and *S* have the same percentile ranks. These methods are especially used with random groups design.

When linear and equipercentile methods are used with nonequivalent groups, more assumptions are needed to be specified and all the methods require synthetizing the

populations, two nonequivalent groups. Linear methods under nonequivalent groups design have Tucker method, Levine observed score method, and Levine true score method. Equipercentile equating upon having two populations has the pool of frequency estimation, Braun-Holland linear method, and chained equating. Those methods are not as straightforward as the methods used in random groups design. They are not the studied method of current research, more detailed explanation and description of those methods can be found in (Kolen & Brennan, 2013).

2.2.4. IRT Equating Methods

As a model based method, using item response theory for equating, a general model that fits the data well should be specified. Item parameters estimated using this fitted model will be used to develop equating functions. Equating using IRT are basically a procedure to put all item parameter and person parameter estimates on the same scale, then either true score or observed score calculated using item parameter estimates will be on the same scale. In addition, the item parameter estimates from IRT are population independent, which has given the IRT equating a lot convenience and flexibility.

Altogether, scaling or equating is not necessary with random groups design when using IRT, because all the parameter estimates have already been put on the same scale no matter the item parameter estimates are obtained in separate steps or in concurrent estimation. In separate estimation, the group abilities are default as mean of 0 and standard deviation of 1, which is legitimate and not mixing the scale up because two groups are equivalent. Using concurrent estimation, the item parameter estimates are on the same

scale simultaneously. No extra step is required for equating. Person standing on the latent scale and item parameter estimates could be compared directly without further adjustment.

When using IRT equating with common-items nonequivalent groups design, the parameter estimates can be obtained at the same time, i.e., concurrent estimation. In this situation, all parameter estimates are on the same scale already, and linking and equating is not necessary. Scale transformation or linking is only necessary for the item parameters and person parameters are estimated separately. It is very common that form *T* is fitted and estimated at the time when form *T* is given, while later on form *S* is given, and parameters will be estimated on form *S* only. Since this the most often situation in reality, the current study will focus on common-item nonequivalent groups design with separate estimations.

Given an IRT three-parameter logistic model fits the data with parameter estimates of $\Theta_{Ti}, a_{Tj}, b_{Tj}, c_{Tj}$, the linear transformation of the target scale *T* configured in estimation to the base scale is possible. A linear transformation of $\Theta_{Ti}, a_{Tj}, b_{Tj}, c_{Tj}$ on scale *T* to scale *S*, respectively, are

$$\Theta_{Si} = A\Theta_{Ti} + B, \qquad (19)$$

$$a_{Sj} = \frac{a_{Tj}}{A}, \qquad (20)$$

$$b_{Sj} = Ab_{Tj} + B, \qquad (21)$$

$$c_{Sj} = c_{Tj}. \qquad (22)$$

The transformed scale fits the model exactly the same, i.e.,

$$p_(x_{ij} = 1|\left(\Theta_{Si}, a_{Sj}, b_{Sj}, c_{Sj}\right)) = p(x_{ij} = 1|\left(A\Theta_{Ti} + B, \frac{a_{Tj}}{A}, Ab_{Tj} + B, c_{Tj}\right)). \qquad (23)$$

The proof of such equal relationship is easy after replacing all the original scale *T* estimates with transformed values on scale *S*.

$$c_{Si} + (1 - c_{Si}) \frac{\exp[a_{Sj}(\Theta_{Si} - b_{Sj})]}{1 + \exp[a_{Sj}(\Theta_{Si} - b_{Sj})]}$$

=

$$c_{Tj} + (1 - c_{Tj}) \frac{\exp\{\frac{a_{Tj}}{A}[(A\Theta_{Ti} + B) - (Ab_{Tj} + B)]\}}{1 + \exp\{\frac{a_{Tj}}{A}[(A\Theta_{Ti} + B) - (Ab_{Tj} + B)]\}}$$

=

$$c_{Tj} + (1 - c_{Tj}) \frac{\exp[a_{Tj}(\Theta_{Ti} - b_{Tj})]}{1 + \exp[a_{Tj}(\Theta_{Ti} - b_{Tj})]}$$

The same relationship after linear transformation is also applicable to graded response model. All remain the same except for the difficult parameter.

$$b_{Sjk} = Ab_{Tjk} + B, \qquad (23)$$

Where k refers to the response category in the polytomous response pattern, e.g., the Likert scale.

After identifying the equal relationship with transforming the scale *T* estimates on to scale *S* using linear transformation, the focal is to find the transformation constant or equating coefficients, A and B. Scale *S* is not arbitrary but determined by the old form of a test in the scenario of equating the new form scale to the old scale for meaningful comparison. In general, there are two types of methods that can be used to obtain the equating coefficients, i.e., the first and second moment (mean and sigma) method, and the

item and test characteristic curve method. All methods introduced here are using the item parameter estimates or ability estimates on the common items.

2.2.4.1. Mean/Mean Method

Using mean/mean method, both equating coefficients are expressed as the ratio or differences between the means of common parameter estimates as transforming from scale *T* to scale *S* (Loyd & Hoover, 1980).

$$A = \frac{\mu(\hat{a}_T)}{\mu(\hat{a}_S)}\ , \qquad (24)$$

$$B = \ \mu(\hat{b}_S) - A\mu(\hat{b}_T), \qquad (25)$$

OR

$$B = \ \mu(\hat{\theta}_S) - A\mu(\hat{\theta}_T). \qquad (26)$$

As can be noticed from the equations, means of discriminating, difficulty, and ability estimates are used to calculate linking or equating coefficients.

2.2.4.2. Mean/Sigma Method

Using mean/sigma method, the equating coefficient A is expressed using the standard deviations of the parameter estimates, while the equating coefficient B is expressed the same as the mean/mean method (Marco, 1977).

$$A = \frac{\sigma(\hat{b}_S)}{\sigma(\hat{b}_T)}\,, \qquad (27)$$

$$OR$$

$$A = \frac{\sigma(\hat{\theta}_S)}{\sigma(\hat{\theta}_T)}\,, \qquad (28)$$

$$B = \ \mu(\hat{b}_S) - A\mu(\hat{b}_T), \qquad (29)$$

OR

$$B = \mu(\hat{\theta}_S) - A\mu(\hat{\theta}_T). \qquad (30)$$

2.2.4.3. Haebara Method

Haebara (1980) proposed a characteristic curve method. Ideally, the item characteristic curve for an item j that is estimated on scale *S* and the item characteristic curve for the same item estimated on scale *T* but transformed to scale *S* with a given θ_i, should be the same as displayed in equation 23. In reality, it is not the same when estimates instead of parameters are used in the item characteristic curve. Nevertheless, the difference should be small. The smaller the difference is, the more adequate the scale transformation is. According to Haebara (1980), the difference between item characteristic curves for an item j, one estimated on scale *S* and one estimated on scale *T* but transformed to scale *S* is squared and summed as

$$Hdiff(\theta_i) = \sum_{j:N} [p_{ij}(\theta_{Si}, \hat{a}_{Sj}, \hat{b}_{Sj}, \hat{c}_{Sj}) - p_{ij}(A\theta_{Ti} + B, \frac{\hat{a}_{Tj}}{A}, A\hat{b}_{Tj} + B, \hat{c}_{Tj})]^2, \qquad (31)$$

where *j:N* is the space for common items. Then the difference is summed over examinees as

$$H_{crit} = \sum_i Hdiff(\theta_i). \qquad (32)$$

The method proceeds with solution to A and B that makes the accumulated differences the smallest.

2.2.4.4. Stocking-Lord Method

Stocking and Lord (1983) provided with another perspective using test characteristic curves instead of item characteristic curves. For a given θ_i, the Stocking-

Lord distance is the squared difference between summation of common item characteristic curves on scale *S* and the summation of common item characteristic curves on scale *S* that are transformed from scale *T*. The distance can be written as

$$SLdiff(\theta_i) = [\sum_{j:N} p_{ij}(\theta_{Si}, \hat{a}_{Sj}, \hat{b}_{Sj}, \hat{c}_{Sj}) - \sum_{j:N} p_{ij}(A\theta_{Ti} + B, \frac{\hat{a}_{Tj}}{A}, A\hat{b}_{Tj} + B, \hat{c}_{Tj})]^2. \quad (33)$$

The distance is also summed over examinees as

$$SL_{crit} = \sum_i SLdiff(\theta_i). \quad (34)$$

The equating coefficients, A and B, are optimized when the distance is the smallest.

2.2.4.5. Divgi Minimum Chi-squared Method

Divgi (1985) developed the method of getting A and B by minimize the sum of the quadratic forms across common items, which is expressed as

$$\sum_j Q_j = \left(\hat{a}_{Sj} - \frac{\hat{a}_{Tj}}{A}, \hat{b}_{Sj} - (A\hat{b}_{Tj} + B)\right)(\sum \boldsymbol{Sj} + \sum \boldsymbol{Tj}^*)^{-1}\left(\hat{a}_{Sj} - \frac{\hat{a}_{Tj}}{A}, \hat{b}_{Sj} - (A\hat{b}_{Tj} + B)\right)^T, \quad (35)$$

Where $\sum \boldsymbol{Sj}$ is 2 by 2 covariance matrix for is $(\hat{a}_{Sj}, \hat{b}_{Sj}), and\ \sum \boldsymbol{Tj}^*$ is the 2 by 2 covariance matrix for transformed $(\hat{a}_{Tj}, \hat{b}_{Tj})$, i.e., $(\frac{\hat{a}_{Tj}}{A}, A\hat{b}_{Tj} + B)$.

2.2.5. IRT True Score Equating

IRT true score equating is defined as when the latent abilities on each form (form *T* and form *S*) are the same, the number-correct true scores on two forms are viewed as equivalent. The number-correct true scores on form *T* and *S* are defined as,

$$T_T(\theta_i) = \sum_{j=1}^{M} P_{ij}\left((\theta_i|a_j, b_j, c_j)\right), \quad (35)$$

$$AND$$

$$T_S(\theta_i) = \sum_{j=1}^{N} P_{ij}\left(\left(\theta_i|a_j, b_j, c_j\right)\right), \qquad (36)$$

where M is the number of items on form *T* and N is the number of items on form *S*. T_T or T_S is the summation of the probabilities of passing/endorsing all items on the corresponding form. Upon using IRT true score equating, the item parameter estimates on two forms have already been put on the same scale. The procedure includes finding the latent ability θ_i on form *T* that can have a number-correct true score of T_T first. The same latent ability θ_i is used to find out the number-correct true score on form *S*. Since the same latent ability is involved in obtaining the number-correct true scores on form *T* and S, $T_S(\theta_i)$is taken as the form *S* equivalent to $T_T(\theta_i)$on form *T*. The challenge lies in finding the θ_i on form *T* to have the specified T_T. The Newton Raphson method is usually applied here to find the corresponding θ_i.

2.2.6. IRT Observed Score Equating

When the observed number-correct scores on scale *T* are equated to find the equivalents on scale *S*, IRT observed score equating is applied. First, distributions of observed scores on form *T* and *S* are estimated. Then, the two distributions are matched by equipercentile method to find the equivalents on one form to another. With a given ability and a test form with given number of items, probabilities for all possible response patterns are calculated as the estimated distribution of observed scores at one point on latent ability. For example, with certain examinee and a test form of three items, possible response patterns include (0,0,0) for passing none of the three items, and (1,0,0), (0,1,0),

and (0,0,1) for passing one item, and (1,1,0), (1,0,1), and (0,1,1) for passing two items, and (1,1,1) for passing three items (0=not passing, 1=passing). Then the distributions are summed over examinees' ability range to get the estimated distribution of observed scores on a form. The example is given on binary data with only three items. As the level of data and items increase, a recursion formula can be applied. If the ability is continuous, the estimated distribution is written as

$$h(x) = \int_{\theta} h(x|\theta)\pi(\theta)d\theta, \qquad (37)$$

where $\pi(\theta)$is the distribution of latent ability. If the ability of examinees is finite, the estimated distribution is written as

$$h(x) = \frac{1}{N}\sum_{i} h\big((x)\big|\theta_i\big), \qquad (38)$$

where N is the number of examinees.

The IRT observed score equating is computational intensive than IRT true score equating, but it is more practical than IRT true score equating because the true scores remain unknown all the time. The estimated distributions of observed scores on two forms need to be combined to yield a synthetic distribution for equation.

2.2.7. Anchoring, Linking, Scaling, and Equating

It is a point where the procedure of anchoring, linking, scaling, and equating can be compared and contrast for clear understanding of each.

The anchoring procedure can have two alternatives. Test developer can have the same group of people take the different tests, or different groups of people take a common

set of items that are put on different test forms (Vale, 1986). Simply, one is anchoring using the same latent trait (θ_i), the other is using the same item characteristics (a_j, b_j, c_j).

Linking is a more general procedure, which is an intermediate step of equating. It is a procedure to put the parameter estimates or observed scores on the common scale without restrictions on the test forms' difficulty and content similarity. Test forms that are built with different difficulty and content specifications can be linked to capture the growth of knowledge with test takers, but do not allow to be equated.

Equating is only used when the test forms are built to the same content and numerical specification. Equated test forms should be similar in test difficulty and score reliability. Linking is especially necessary when using IRT equating without concurrent estimation for the common-item nonequivalent groups design. Adequate equating is dependent on the adequate linking.

Raw scores are often transformed to scaled scores for score interpretation purpose. The mean and standard deviation of the raw scores are manipulated to have certain values. For example, the mean and standard deviation of the raw scores on form *S* is 28.56 and 13.24, respectively. A mean of 100 and a standard deviation of 15 are expected on scaled scores. Then the raw scores are transformed to scale scores by getting rid of the original mean and standard deviation of 28.56 and 13.24 (resulting z scores of raw scores), and applying the expected mean and standard deviation of 100 and 15 to the z scores. Therefore, scaling is the manipulation of mean and standard deviation within the same form or the same scale. Equating is to find the score equivalent of a specified score on two forms interchangeably. Form *T* raw scores are equated to have raw score equivalents on

form *S*. Then the score equivalents are scaled to have the expected mean and standard deviation.

2.2.8. Standard Error of Equating

The standard error of equating is the standard deviation of the sampling distribution of obtained equivalent scores for one score point, given an equating method is applied many times with different samples of examinees each time between two forms. Let's define form *T* as the new from and form *S* as the old form. We want to find the equivalent score on form *S* for a score of 88 on form *T*. With the first 1000 sample examinees, 500 taking form *T* and 500 taking form *S*, and we obtain the score of 86 on form *S* that is the equivalent to the score of 88 on form *T*. With the second 1000 sample examinees, we obtain 83, and with third 1000 sample examinees, we obtain 85, and so on. In this example, the score of 86, 83, and 85 on form *S* are all equivalents of 88 on form *T*. The standard deviation of 86, 83, and 85 are the standard error of equating. As can be seen here, the empirical process of documenting the standard error of equating involves resampling and equating calculation each time. The analytic method could also be applied to get the standard errors of equating, which involves the using of available variance and covariance structure of parameter estimates and other available information of the design and the given method of equating to deduct the standard error. It is intuitive that the standard error of equating or equating error comes from the different samples of the examinees, given an equating method. Each sample comes with a different score equivalent for score 88. Thus, the standard error of equating which originates from random sampling of the examinees is the random error, given an equating method. It should be

distinguished from the systematic error of equating which usually results from using different equating methods, test implementations, equating designs, and DIF problems.

2.2.9. Evaluation of Equating Results

Harris & Crouse (1993) had pointed out the criteria for equating was not fully developed and applied to the extent it should be. However, the discussions in their papers are holistic concerns starting with the question of when equating is appropriate, how equating should be performed based on the collected data, and how to evaluate the equated results. The evaluation discussed here are restricted to the last question, how to evaluate the equated results given an equating method.

Like other statistical procedures, the standard error of equating is the most important evaluative criteria that can be used to check the quality of equating, given an equating method. However, it is not realistic to approach standard error of equating in empirical studies. In order to assure adequate equating is applied, the associated properties of equating can be used as evaluative criteria also. After an equating being done, the symmetry property, the equity property, or the observed score equity property could be evaluated according to the equating methods used. Equating functions should be population invariant, which could also be used as a criterion for adequate equating. When the properties of equating do not hold in most cases, the equating procedure might be problematic. In some situation, no equating is better than equating since equating could add systematic error into the obtained equivalent scores.

2.3. Overview of Differential Item functioning

2.3.1. Differential Item Functioning

In the context of psychological measurement, measurement invariance is defined as a measurement device or instrument has the following characteristics. The assignment of scores to represent certain latent trait of a person is only determined by the target latent variable while independent of all other unrelated variables, latent or observed(Millsap, 2012). The idea is easy to understand in the context of a physical measurement situation. Suppose that a thermometer is used to measure the body temperature of participants from three age groups, 0 to 18 years old, 19 to 50 years old, and above 50 years old. Here the body temperature is the focal variable, and the age group is the irrelevant variable, that is the reading of temperature using the thermometer should not be a function of age group. If there is a relationship between body temperature reading and age groups, measurement bias appears and measure invariance does not hold in such case.

Conditional probability is used to express the definition of measurement invariance, i.e.,

$$P(X|W,I) = P(X|W), \qquad (39)$$

where X is the measured variable, W is the target latent variable, and I is the irrelevant variable, which usually are demographic variables, like gender, ethnicity, citizenship, and culture background. The irrelevant variables could also be research specific grouping variables. If the equation does not hold, the measurement invariance is violated and measurement bias exists.

The most commonly used method for testing measurement invariance is multi-group confirmatory factor analysis (MGCFA). The testing of measurement invariance is a step by step procedure in finding out at which level the measurement invariance quality holds, i.e., the four steps of factorial invariance testing. The four steps are configural invariance, pattern invariance, strong invariance (metric invariance), and strict invariance (scalar invariance). Group means comparison is meaningful when measurement invariance holds (Millsap, 2012).

Using IRT model, the measurement bias can be studied both at the item level and the test level. Item bias is referred to as differential item functioning (DIF). Differential item functioning is defined as the same item has the different item parameters and consequently yields different probability of correct response between the focal and the reference groups given the target latent trait or ability in two groups are matching, i.e., the item response function is not the same between groups for the same item with same latent ability (Ellis, 1989; Mellenbergh, 1989; Zumbo, 1999). Graphically, the same item will have two different item characteristic curves on focal and reference groups (Mellenbergh, 1989). Like in any other situation, biased items will give rise to a lot of challenges and problems in item response theory modeling for parameter estimation, linking, and equating. After all, the measurement invariance is the prerequisite for a lot statistical procedure and other important issues in educational and testing context. Due to the indeterminacy of item parameters and person parameter estimates, the ability or latent trait estimates with biased items in the test are not reliable. The present of biased items also affect the estimation of parameters on other unbiased items. The existence of biased items

in the common items set will affect the configuration of linking constants and affect the adequate equating procedure. Wingersky, Cook, and Eignor (1987) recommended to study the efficiency of linking items against the whole test rather than focusing on the estimation of standard errors of linking items themselves.

2.3.2. Forms of Bias

Uniform bias is defined as no interaction effect between item parameter estimates and the group membership, i.e., an item is estimated as more difficult or easier in the focal group than in the reference group across the range of matching variable (latent ability scale) (Ellis, 1989; Mellenbergh, 1989; Millsap, 2012; Millsap & Everson, 1993; Zumbo, 1999). Non-uniform bias is defined as having an interaction effect between item parameter estimates and the group membership. For example, an item is displayed to be more difficult in the focal group on the lower end of the matching variable, while the same item is displayed to be easier in the focal group on the higher end of the matching variable (Ellis, 1989; Mellenbergh, 1989; Millsap, 2012; Millsap & Everson, 1993; Zumbo, 1999).

2.3.3. Item Bias Detection

On the one hand the differences between the same item parameter estimates could be due to sampling error, on the other the differences between the same item parameter estimates could be the produce of item bias between focal and reference groups. How to quantify the difference of the same item parameter estimates in different groups and measure its magnitude is the discussion of the following section. Methods of detecting item bias include Likelihood-Ratio (LR) tests, Wald Statistic, Mantel-Haenszel Statistics (Dorans & Holland, 1992; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006;

Holland, Thayer, Wainer, & Braun, 1988; Steinberg & Thissen, 2006; Thissen, Steinberg, & Wainer, 1993; Wald, 1943), Raju area statistics, and differential functions of items and tests (DFIT) (Millsap & Everson, 1993; Raju, 1988, 1990; Raju, Van der Linden, & Fleer, 1995).

2.3.3.1. Likelihood-ratio Test

Likelihood-ratio test is a model based method (Edelen et al., 2006; Thissen et al., 1993). In order to use LR method, a baseline model M_1 without any invariant constraints is specified except for constraints needed for model identification. Then a second model M_0 is specified adding constraints to item parameters (item discrimination, item location, step difficulties, and threshold). M_0 and M_1 are nested models. With model configuration, the likelihood values will be calculated for M_0 and M_1 as L_0 and L_1, respectively. The natural logarithm of difference between L_0 and L_1 multiplied by -2 will follow a chi-square distribution with a degree of freedom of the number of constraints added to M_1 to get M_0, i.e., $df_{LR} = df_{M_0} - df_{M_1}$. The LR test statistics denoted by Q_L is written as,

$$Q_L = -2\log\left(\frac{L_0}{L_1}\right). \qquad (40)$$

LR test procedure can be applied to both dichotomous and polytomous IRT models. Along with LR test is an omnibus test of null hypothesis that all concerned (constrained) item parameters are group invariant. The alternative is at least one of them is not group invariant. A series of post hoc test is involved upon the rejection of null hypothesis.

2.3.3.2. Wald Test

Wald statistics is an item level test statistics, which allows the direct comparison of concerned item parameters across groups(Thissen et al., 1993; Wald, 1943). Supposing the item location parameter is under concern and it is defined that b_{jR} (reference group) and b_{jF} (focal group) are estimates of item j difficulties in two groups. The Wald statistic of testing the null hypothesis of H_0: $b_{jR} = b_{jF}$ for item j is given by

$$Z_j = \frac{\left(\hat{b}_{jR} - \hat{b}_{jF}\right)}{\sqrt{Var\left(\hat{b}_{jR}\right) + Var\left(\hat{b}_{jF}\right)}}. \qquad (41)$$

The Wald statistics is compared to the standard normal distribution for the significant difference reference. The method could be easily applied when more than items or more than two groups are under concern.

2.3.3.3. Mantel-Haenszel Procedure

Mantel-Haenszel procedure is applied in DIF by testing the null hypothesis of the odds ratio of answering an item in reference group equals that in focal group (Dorans & Holland, 1992). The count of correct and incorrect answers to item j for the level i ($i = 1,2,3, \ldots m$)on matching variable is listed in the following table. The partial table of response to item j across the levels of latent matching variable i is given by Table 1. The null hypothesis for DIF analysis using MH procedure is,

$$H_0: \frac{A_i D_i}{C_i B_i} = 1. \qquad (42)$$

The chi-square test statistics for MH procedure is,

$$Q_{MH} = \frac{\left[\sum_i A_i - \sum_i E\left(\widehat{A_i}\right)\right]^2}{\sum_i Var(A_i)}, \qquad (43)$$

where $i = 1,2,3, \dots, m$, and

$$E\left(\widehat{A_i}\right) = E\left(\widehat{A_i}|H_0\right) = \frac{N_{Fi}N_{Ti}}{N_i},$$

$$Var(A_i) = \widehat{Var}(A_i|H_0) = \frac{N_{Fi}N_{Ri}N_{Ti}N_{Wi}}{N_i^2(N_i - 1)}.$$

The $MH - \chi^2$ statistic is based on the hypergeometric distribution. For the continuity on i, matching variable score levels, the statistics should be corrected by the ways of the table, which is 2 here, the grouping variable and answer type. The correction is given,

$$Q_{MH} = \frac{\left[\sum_i A_i - \sum_i E\left(\widehat{A_i}\right)\right]^2}{\sum_i Var(A_i)} \frac{(n-1)}{n}, \qquad (44)$$

where n is the number of ways. Q_{MH} approximates a chi-square distribution with $df = 1$.

2.3.3.4. Raju Area Statistics

Raju area statistics are used to quantify the area (difference or distance) between two item response curves (ICC) (Raju, 1988). Let's define F_{jF} and F_{jR} are two ICCs for item j on focal group and reference group. The area between two curves is given by,

$$Signed\ Area: SA = \int_{-\infty}^{\infty} \left(F_{jF} - F_{jR}\right) d\theta, \qquad (45)$$

$$AND$$

$$Unsigned\ Area: UA = \int_{-\infty}^{\infty} \left|F_{jF} - F_{jR}\right| d\theta. \qquad (46)$$

Under Rasch model,

$$SA = \left(b_{jR} - b_{jF}\right), \qquad (47)$$

$$AND$$

$$UA = \left|b_{jR} - b_{jF}\right|. \qquad (48)$$

Under 2PL model,

$$SA = \left(b_{jR} - b_{jF}\right), \qquad (49)$$

$$AND$$

$$UA = \left|b_{jR} - b_{jF}\right|, if\ a_{jF} = a_{jR}, \quad (50)$$

$$OR$$

$$UA = \left|\frac{2\left(a_{jR} - a_{jF}\right)}{D a_{jF} a_{jR}} ln\left(1 + exp\left(\frac{D a_{jF} a_{jR}\left(b_{jR} - b_{jF}\right)}{a_{jR} - a_{jF}}\right)\right) - \left(b_{jR} - b_{jF}\right)\right|, if\ a_{jF} \neq a_{jR.} \quad (51)$$

Under 3PL model,

$$SA = (1 - c)\left(b_{jR} - b_{jF}\right), \qquad (52)$$

$$AND$$

$$UA = (1 - c)\left|b_{jR} - b_{jF}\right|, if\ a_{jF} = a_{jR}, \quad (53)$$

$$OR$$

$$UA = (1 - c)\left|\frac{2\left(a_{jR} - a_{jF}\right)}{D a_{jF} a_{jR}} ln\left(1 + exp\left(\frac{D a_{jF} a_{jR}\left(b_{jR} - b_{jF}\right)}{a_{jR} - a_{jF}}\right)\right) - \left(b_{jR} - b_{jR}\right)\right|, if\ a_{jF} \neq a_{jR.} \quad (54)$$

Raju (1990) had presented the mean and variance for the sampling distribution of SA and UA statistics under different models. The statistical significance tests of SA and UA under different situations are possible with z test.

2.3.3.5. Differential Functions of Items and Tests

Not like Raju area statistics, DFIT (Raju et al., 1995) is not different in forms from model to model. DFIT is applied with the identification of a well fitted model. Therefore, either different or the similar item response functions are estimated. The item response functions are not expected be exactly the same due to sampling error. Therefore, for a person i, the difference in the probabilities of passing/endorsing items on a test level between two groups is captured by difference in true score, which is defined as

$$D_i^2 = (T_F - T_R)^2. \qquad (55)$$

DTF is defined as the expectation of accumulated differences across examinees using either examinees from focal group only or reference group only,

$$DTF = E_F(D_i^2) = \int_\theta D_i^2 f_F(\theta) d\theta = \sigma_D^2 + \mu_D^2, \quad (56)$$

where E_F is the expected values across focal group and $f_F(\theta)$ is the density function of θ in the focal group. The non-compensatory DIF (NCDIF) for an item j across examinees in focal group is given as,

$$NCDIF_j = E_F\left[P_{jF} - P_{jR}\right]^2 = E_F\left(d_j^2\right) = \sigma_{d_j}^2 + \mu_{d_j}^2. \quad (57)$$

The compensatory DIF (CDIF) for an item j across examinees in focal group is given as,

$$CDIF_j = E_F\left(Dd_j\right) = Cov\left(D, d_j\right) + \mu_D \mu_{d_j}. \qquad (58)$$

The DTF can be viewed as the summation of CDIF over test items, which is

$$DTF = \sum_{j=1}^{n} CDIF_j. \qquad (59)$$

The chi-square statistic proposed by Raju et al (1995) for testing the null hypothesis of,

$$H_{0:}\, NCDIF_j = 0, \qquad (60)$$

is given by,

$$\chi^2_{NCDIF_j} = \frac{\widehat{NCDIF_j}}{\hat{\sigma}^2_{d_j}/N_f}, \qquad (61)$$

where N_f is the number of test takers in focal group and the degree of freedom for the chi-square test statistic. A lot of simulation studies have been conducted in profiling the performance of each method. Collins, Raju, and Edwards (2000) have investigated the differential functioning detection on a satisfaction scale with Lord's chi-square method, Raju SA method, and differential functioning of items and tests (DFIT). DFIT has a more consistent performance compared to other procedures. Due to the limited test statistics available for the significance of DFIT indices, Oshima, Raju, and Nanda (2006) has proposed a bootstrapping alike method to obtain the sampling distribution of paired item parameter estimate difference score between focal group and reference group. The difference score on the 99th ($critical\ \alpha = 0.01$) percentile of the sampling distribution is selected cutoff value for decision making for statistical significant difference. The method is often referred as item parameter replication method.

2.3.4. Test Purification

Upon having biased items in a test, the first and the most need is to eliminate those biased items. However, the identification of biased items is not as straightforward as only

applying one or two of the DIF detection methods aforementioned. It is a computational iterative procedure. Item parameter estimates from focal and reference groups must be placed on a common metric so that the paired item parameter estimates comparison is meaningful.

An iterative procedure under three parameters logistic model is proposed in the literature. The procedure (Lord, 1980; Marco, 1977) is given by following steps:

1. Use all data from two groups and run the model with item location parameters following a standard normal distribution $N(0,1)$ and save the guessing parameter estimates.
2. Fit the same 3PL models in separate groups with location in each group following a standard normal distribution $N(0,1)$. Fix the guessing parameter estimates saved from step 1.
3. Remove all biased items using any of the bias detection methods discussed previously.
4. Fit the same 3PL model with the remaining items using data from two groups to estimate the latent ability. Save the latent ability estimates.
5. Fit the same 3PL models in separate groups using all items (no dropping) and fixing the latent ability using estimates from step 4.
6. Repeat step 3.

Park and Lautenschlager (1990) modified Lord and Marco's procedure by repeating step 3 through 5 until in each iteration the same items are flagged as biased items. The concurrent estimation of latent ability on two groups rendered linking

unnecessary. However, the latent ability is estimated repetitively in both procedures, which is time and resource expensive.

Segall (1983) proposed another separate parameter estimation and bias detection procedure. This iterative procedure is given by,

1. Use the same fitted model and estimate item parameters in focal and reference group separately.
2. Use one of the available linking methods to obtain a linking function using the parameter estimates from step 1.
3. Put two groups of parameter estimates on a common scale using the linking function identified in step 2.
4. Examine the item parameters estimates for bias using one of the bias detection methods available, and remove all biased items.
5. Generate an updated linking function using the remaining items.
6. Put two groups of parameter estimates of all items (no dropping) on a common scale using the updated linking function from step 5.
7. Evaluate item parameter estimates for bias detection and remove the biased items.
8. Repeat step 5 through 7 until each time with the same biased items identified.

Application researches and simulation studies have testified the usability of Segall's method (Candell & Drasgow, 1988; Drasgow, 1987; Park & Lautenschlager, 1990). Iterative linking plan has better results than single linking procedure (Kim & Cohen, 1992).

3. METHODOLOGY

3.1. Study Design

The simulation study is performed under common items nonequivalent groups design, i.e., two groups with different levels of ability taking alternate forms of a test with shared items. Group 1 is simulated with a relative low ability profile and Group 2 is with a relative high ability profile to distinguish as nonequivalent groups. The differences in abilities are assumed to have an impact on the response patterns to common items between two groups. Test length is 50 items through the study. Each group will take a form of a test of 50 items. The two forms given to group 1 (form 1) and group 2 (form 2) are parallel forms, which means they are constructed according to the same content and statistical specifications. Specifically in this study, the forms are parallel in terms of three aspects: 1) the total number of items are the same on two forms, i.e., 50 items in total on each form; 2) both unique and common items on two forms are of similar level of item difficulty and discrimination, i.e., they are random number generated from the same distribution in terms of each parameter and the mean and standard deviation of corresponding parameter in separate groups are comparable; 3) common items are placed at the same position on two forms, i.e., the item index are the same. For example, if item j is placed at the position with an index of 36 in group 1, the same item will be placed in group 2 with the same index of 36. Different levels and types of DIF will be assigned to some common items on two forms. In the current study, DIF items will be assigned to only one group at a time,

either group 1 or group 2, no mixed type assignment of DIF items, e.g., all designated DIF items will be easier and less discriminative among group 1 or group 2.

3.2. Simulation Factors

Six factors and their associated effects on equating coefficients and item parameter recovery under common items nonequivalent groups design are explored in the current study. The factors are common item ratio, DIF item ratio, direction of DIF, form of DIF, magnitude of DIF, and sample size. Table 2 displays the simulation factors in the current study. Following, each factor is explained in detail.

3.2.1. Common item ratio

Among the fifty items, two percentages of common items are assigned, i.e., 20% and 30%, which results in 10 common items and 15 common items out of 50, respectively. The two percentages are selected due to the fact that twenty percent of common items has been widely tested, suggested, and required in literature. Less than twenty percent of shared items would result in inadequate anchoring. Thirty percent of common items is also tested in this study as a suffice condition for common items in nonequivalent groups design. Against this factor, research question 1 will be what is the effect of common item ratio to the number of test items in each form on the equating coefficients and item parameters recovery with nonequivalent groups.

3.2.2. DIF item ratio

Within the common items, three percentages of DIF items are tested, i.e., 20%, 40%, and 60%, which results in 2, 4, and 6 DIF items out of 10 common items, and 3, 6, and 9 DIF items out of 15 common items. The three levels of DIF item ratio are selected

to represent a small, medium, and large percentage of different item behavior between groups. Against this factor, research question 2 will be what is the effect of DIF item ratio to the number of common items on the equating coefficients and item parameters recovery with nonequivalent groups.

3.2.3. Form of DIF

Within each DIF item, two types of DIF are tested, i.e., uniform and non-uniform. The uniform DIF happens when there is a difference in item location parameters between groups even though the item is the same, i.e., $b_{2j} \neq b_{1j}$, while b_{1j} is the item j location parameter from group 1 and b_{2j} is the same item j location parameter from group 2. The non-uniform DIF happens when there is a difference in item discrimination and/or item location for the same item between different testing groups.

With this factor, research question 3 will be what is the effect of form of DIF in common items on the equating coefficients and item parameters recovery with nonequivalent groups. Is there any difference between uniform and nonuniform DIF in terms of their effect on equating coefficients and item parameter recovery?

3.2.4. Magnitude of DIF

Within uniform DIF, three different levels of DIF magnitudes will be tested. The small uniform DIF is represented by $b_{2j} - b_{1j} = 0.3 \; or \; b_{1j} - b_{2j} = 0.3$, medium by $b_{2j} - b_{1j} = 0.6 \; or \; b_{1j} - b_{2j} = 0.6$, and large by $b_{2j} - b_{1j} = 0.9 \; or \; b_{1j} - b_{2j} = 0.9$. In this study only one level of item discrimination is tested, which is represented by $a_{2j} - a_{1j} = 0.3 \; or \; a_{1j} - a_{2j} = 0.3$, representing a small level of non-uniform bias. The medium and large difference in discrimination parameters between groups are avoided

because the manipulation of discrimination parameter could easily make the changed values extremes or exceed the extremes, situations that will not be the focus of current study. For this factor, research question 4 will be what is the effect of magnitude of DIF in common items on the equating coefficients and item parameters recovery with nonequivalent groups.

3.2.5. Direction of DIF

Both uniform and non-uniform DIF are simulated as non-directional in terms of group membership. Although it is intuitive that the same item would favor toward group 2, the group that has a relative high ability profile, it is possible that DIF items would be against group 2. For example, under the circumstance of encountering an easy item, the high ability person might get confused or simply having a hard time recalling the simple fact, which is very likely to give a wrong answer. DIF items are simulated to be more difficult and discriminative to group 2 first and then to be more difficult and discriminative in group 1. The non-directional DIF between groups is simulated to eliminate the uncertainty. Against this factor, research question 5 will be whether there is any difference in terms of equation coefficients and item parameter recovery when DIF direction changes from against to in favor of group 2 test takers.

3.2.6. Sample size

Sample size of the simulated participants is another factor examined in this study. For each condition specified above, a small sample size of 500, a medium of 1000, and a large of 3000 for each group will be tested for linking and equating. Against this factor, research question 6 will be what is the effect of sample size on equating coefficients and

item parameter recovery when DIF item presents in common items nonequivalent groups design equating.

3.2.7. Reference condition

All conditions aforementioned are situations when DIF items presents. In order to have a reference when DIF is absent, null conditions with two percentages of common items and three levels of sample size are simulated in the current study to see the behaviors of interested parameters. And research question 7 will be what equating coefficients and item parameters recovery look like when DIF is absent. Is there any difference under different common item ratios? Is there any difference under different sample sizes?

To summarize, there are 2 (common item ratio) * 3 (DIF item ratio) * 2 (form of DIF) * 3 (magnitude of DIF) * 2 (direction of DIF) * 3 (level of sample size) = 216 conditions when DIF items present. There are also 2 (common item ratio) * 3 (level of sample size) = 6 null condition when DIF item is absent. All conditions are run in R, a free open source package for statistical computation (https://cran.r-project.org). Specifically, the package of "irtoys": *A Collection of Functions Related to Item Response Theory* is employed in the current study (Partchev, 2016). For each condition, 500 replications are conducted.

3.3. Data Generation

Generally, there are two population distributions are associated with this study, the simulated participant and item parameter distributions. According to literature, the ability of simulated test takers is usually assumed either to be a normal, a uniform, or a β distribution, and the item parameters, i.e., the item difficulty, item discrimination are

chosen either from a normal, a uniform, a β, or a lognormal distribution. The decisions are made according to the research interests, e.g., whether a normal or a skewed distribution of population is of concern, and whether the realistic item parameters or extreme cases are focal (Han, 2007).

The ability of simulated test takers is generated from a normal distribution. Group 1, which has a relative low ability profile, is generated from a standard normal distribution $N\ (0,1)$. Group 2, which has a relative high ability profile, is generated from a normal distribution with both mean and standard deviation equal 1, i.e., $N\ (1,1)$. For example, the sample size tested is 1000. Then 1000 of ability scores will be randomly generated from $N\ (0,1)$ for group 1, and 1000 of ability scores will be randomly generated from $N\ (1,1)$ for group 2. For each replication, the ability scores in each group will be regenerated.

The item location/difficulty parameter is generated from a standard normal distribution $N\ (0,1)$. The item discrimination parameter is generated from a uniform distribution $U(0.8, 1.7)$. The item parameter distributions are specified reflecting the general acceptable ranges of item location/difficulty and item discrimination.

In the current study, the maximum number of unique items on each test form is 40, and the maximum number of common items is 15. In order to add flexibility to the manipulation of study conditions, a unique item pool of 80 items is constructed, and a common item pool of 15 items is constructed. The descriptive statistics of unique item pool by form and common item pool are shown in Table 3. Another 12 common item pools of 15 items with different direction of DIF, different levels of uniform DIF and one level of non-uniform DIF are also constructed. There are six out of twelve with DIF items

generally favoring toward group 2: 1) DIF item pool with small uniform DIF only; 2) DIF item pool with medium uniform DIF only; 3) DIF item pool with large uniform DIF only; 4) DIF item pool with small uniform DIF and non-uniform DIF; 5) DIF item pool with medium uniform DIF and non-uniform DIF; 6) DIF item pool with large uniform DIF and non-uniform DIF. There are another six out of twelve with DIF items generally favoring toward group 1. It is noted that there is only one level of non-uniform DIF examined in the current study. For example, the condition of 15 common items having 3 DIF items with small uniform and non-uniform DIF is tested. Then item 1 through item 35 in the unique item pool of 80 items will be taken out as the unique items on form 1. Item 1 through item 15, i.e., all the common items in the common item pool will be taken out as the rest 15 items on form 1. Together there are 50 items on form 1. Let's index them as item 1 through item 50. Item 41 through item 75 in the unique item pool of 80 items will be taken out as the unique items on form 2. Item 1 through item 15, i.e., all the common items in the common item pool will be taken out as the rest 15 item on from 2. Let's also index them as item 1 through item 50. However, there are 3 DIF items with small uniform and non-uniform DIF. We need to replace 3 out of the 15 common items using DIF items from the item pool that has small uniform and non-uniform DIF. If the DIF items existed in group 1, the DIF item 13, 14, and 15 in the corresponding DIF item pool replaces item 48, 49, and 50 on form 1, respectively. If the DIF items existed in group 2, the DIF item 13, 14, and 15 in the corresponding DIF item pool replaces item 48, 49, and 50 on form 2, respectively.

Given the conditions to be tested in the study, item parameter data sets generated include 2 sets when DIF is absent, and 2 (common item ratio) * 3 (DIF item ratio) * 2 (form of DIF) *2 (direction of DIF) * 3 (magnitude of DIF) = 72 sets when DIF presents. The item parameter generated will be used to obtain the probability of answering an item correctly using the 2PL model. The calculated probability will be compared with a random number drawing from a uniform distribution $U(0,1)$. If the calculated probability is larger than the random number, the observed response of 1 (correct) for that item given the person ability will be assigned. If the calculated probability is smaller than the random number, the observed response of 0 (incorrect) for that item given the person ability will be assigned. For example, if the sample size 500 is tested. Then the observed response data set within each group upon each replication will be a 100×50 matrix of 0s and 1s. The row is participant ID and the column is item index. All the data will be generated using 2PL model and also fitted into 2PL model after obtaining the observed response. Given the 74 data sets of item parameters, 3 different sample sizes, and 100 replications, there are $74 \times 3 \times 100$ times of model fittings and estimations by group of test takers in the current study.

3.4. Analysis Procedure

3.4.1. IRT linking with Stocking-Lord method

Common item parameter estimate linking is conducted in the IRT framework with separate model estimation. When separate model estimation is applied to observed response on each group, the ability for each group during the estimation is default as $N(0,1)$, i.e., designated as equivalent groups. However, two groups are nonequivalent in

ability and linking is required to put common item parameter estimates on the same scale that honors relative standing among groups.

Among the various IRT linking methods, Stocking-Lord, also called test characteristic curve method is employed due to the literature endorsing Stocking-Lord as the reliable and robust method especially when item parameter estimates are problematic, i.e., large differences in results between two groups of estimation. However, it is also reasonable to assume that test characteristic method will more likely to be affected by DIF items since this method uses all the raw differences from items between groups. Given the popularity of and widely acknowledgement to the test characteristic method and the possible problem associated with test characteristic method, it is necessary to quantify the performance of such method under the presence of DIF items.

3.4.2. Expected values of A and B

When linking group 2 onto the scale of group 1, the expected value of A, which is the slope for scale transformation, is 1. The expected value of B, which is the intercept for scale transformation, is also 1. The reason traces back to the true population parameter of two groups. Group 1 follows $N\ (0,1)$ and group 2 follows $N\ (1,1)$. However, during the estimation stage both groups are fixed as $N\ (0,1)$. The location of group 2 shifts downward (left) by 1, while the dispersion of group 2 remains the same. In order to honor the original scale, the location should shift 1 upward (right) and the dispersion remains unchanged. A as the slope is to hold dispersion the same and the expected value should be 1. B as the intercept is to make the location move upward (right) by 1 and the expected value should be 1.

3.4.3. Linking and calibrating plan

Form 1 is set as the old form and form 2 is set as the new form. Item parameter estimates on new form is transformed/linked back to the scale of old form. In order to compare the item parameter estimates to the generating parameters, i.e., the true value. All item parameter estimates on from 1 and transformed item parameter estimates on form 2 need to be put on the scale of generating parameters. As described, two steps of linking take place. Details involve and need clarification.

Step 1: Put item parameter estimates on from 2 and from 1 on the common metric of form 1

After fitting 2PL models for each group of response data, there are two groups of item parameter estimates for common items. These estimates are used with Stocking-Lord method to obtain equating constants A and B. All item parameter estimates for both unique and common items on form 2 taken by group 2 are transformed using A and B so that the transformed item parameter estimates will be on the same scale of form 1 taken by group 1. Finally, the common item parameter estimates on two forms, i.e., the originals on form 1 and the transformed on form 2, will be averaged and taken as the common item parameter estimates (Hambleton & Swaminathan, 1985). The linking constants A and B in this step will be retained for each condition under each replication. The retained value under each replication will be compared to the expected value of A and B for evaluation.

The model estimation is completed using the function $est(\ )$ in the packages of "irtoys". The linking is performed using the function $sca()$ also in the package of

"irtoys". In order to apply Stocking-Lord method, quadrature points and quadrature weights are supplied in the function.

Step 2: Put item parameter estimates resulted from step 1 and generating parameters on the common metric of generating parameters.

The same linking method and procedure is applied to transform a) unique item parameter estimates on form 1, b) unique but transformed item parameter estimates on form2, and c) average common item parameter estimates on to the generating scale. For example, if the condition being dealt with is 10 common items and 40 unique items on each form. Then according to the aforementioned procedure, 40 item parameter estimates on form 1, 40 transformed item parameter estimates on form 2, and 10 common items with average common item parameter estimates (90 in total) will be transformed back to the generating scale. Under this situation, linking constants A and B are obtained using all 90 items as common items. After getting the constants A and B, the 90 item parameter estimates will be transformed to the generating scale. The transformed item parameter estimates of the 90 items will be compared with the 90 generating item parameters for item parameter recovery evaluation. Linking/equating constants A and B obtained in this step will not be retained.

3.5. Evaluation Criteria

The current study will examine the performance or recovery of four parameters, i.e., the linking/equating constants A and B, item difficulty, and item discrimination. Generally, two indexes will be used as the evaluative criteria, i.e., bias and root mean

square error (RMSE) of four parameters (Harris & Crouse, 1993). Bias of *A* is calculated as

$$Bias_A = \frac{\sum_{l=1}^{NREP}(\hat{A}_l - 1)}{NREP}, and$$

bias of *B* is calculated similarly as

$$Bias_B = \frac{\sum_{l=1}^{NREP}(\hat{B}_l - 1)}{NREP},$$

where $\hat{A}_l$ and $\hat{B}_l$ are linking constants obtained from each replication given the simulated condition, 1 is the expected value of linking constants A and B, and $NREP$ is the number of replications, which is 100 in the current study. With the fact that both *A* and *B* had the expected value of 1, the relative bias would be the same to the bias itself. A rule of thumb for acceptable relative bias is that the value is not greater than 0.05 (Hoogland & Boomsma, 1998).

RMSE of *A* is calculated as

$$RMSE_A = \sqrt{\frac{\sum_{l=1}^{NREP}(\hat{A}_l - 1)^2}{NREP}}, and$$

RMSE of *B* is calculated as

$$RMSE_B = \sqrt{\frac{\sum_{l=1}^{NREP}(\hat{B}_l - 1)^2}{NREP}}.$$

Bias of item discrimination *a* is calculated as

$$Bias_a = \frac{\sum_{j=1}^{N}(\hat{a}_j - a_j)}{N}, and$$

bias of item location *b* is calculated similarly as

$$Bias_b = \frac{\sum_{j=1}^{N}(\hat{b}_j - b_j)}{N},$$

where $\hat{a}_j$ is the estimated item discrimination and a_j is the true generating value, and $\hat{b}_j$ is the estimated item location and $b_j$is the true generating value, given an item j. N is the number of items on each form, which is 50 in the current study. Mean bias of *a* and *b* are obtained as the mean of bias of *a* and *b* over the number of replication, which is 100 in the current study. The relative mean bias of *a* and *b* is not as straightforward as the relative bias of *A* and *B* because bias of *a* and *b* is not on the individual item level but on the level of test form across 50 items. However, with the mean statistics of *a* (around 1.2) and *b* (around 0.2) shown in Table 3 in mind, the mean bias of *a* and *b* is acceptable when mean bias of *a* is no greater than 0.06, and mean bias of *b* is no greater than 0.01.

RMSE of a is calculated as

$$RMSE_a = \sqrt{\frac{\sum_{j=1}^{N}(\hat{a}_j - a_j)^2}{N}}, and$$

RMSE of b is calculated as

$$RMSE_b = \sqrt{\frac{\sum_{j=1}^{N}(\hat{b}_j - b_j)^2}{N}}.$$

Mean RMSE of item discrimination *a* and item location *b* are then calculated as average of RMSE of *a* and average of RMSE of *b* over 100 replications.

4. RESULTS

The results were reported under five sections, i.e., 1) null condition without DIF items, 2) uniform DIF against group 2 condition, 3) uniform DIF favoring group 2 condition, 4) uniform and nonuniform DIF against group 2 condition, and 5) uniform and nonuniform DIF favoring group 2 condition. Under each section, results were presented on patterns of sample size, number of common items, number of DIF items, magnitude of DIF items, forms of DIF, and direction of DIF if applicable in terms of biases and RMSEs of linking constants *A* and *B*, and mean biases and mean RMSEs of item parameters *a* and *b*.

4.1. Null Condition Without DIF Item

The biases and RMSEs of linking constants *A* and *B* over 100 replications are shown for three sample sizes (500, 1000, and 3000) and two number of common items (10 or 15 out of 50) in Figure 1. The biases of *A* and *B* were smaller than 0.05 in absolute value across conditions, which meant the relative biases were also smaller than 0.05. The RMSEs of *A* and *B* were smaller than 0.1 in most cases. The biases and RMSEs of linking intercept B were larger than those of linking slope A in general. The overall mean biases and mean RMSEs of item discrimination parameter *a* and item location parameter *b* over 100 replications are also presented in Figure 1. The mean biases of *a* and *b* are smaller than 0.06 and 0.01, respectively. The mean biases are positive for *a*, whereas negative for *b* in most cases. The mean RMSEs of *a* were similar under the same sample size across different number of common items. It was the same with the

mean RMSEs of b. Generally, as the sample size increased, the biases and RMSEs decreased, a trend more obvious for RMSEs of *A* and *B*, and mean RMSEs of *a* and *b*. The sample size effect leveled off at 1000 for the mean biases of *a* and *b*. In general, the biases and RMSEs decreased as the number of common items increased.

Mean biases and mean RMSEs of *a* and *b* by group — group 1 unique item, group 2 unique item, and common item — over 100 replications were shown for three sample sizes and two number of common items in Figure 2. There were no significant differences on the magnitude of mean biases of *a* (<0.06 in absolute value) by group and mean bias of *b* (<0.01in absolute value) by group, and the differences between groups only started from the third decimal place. There were differences on the magnitudes of mean RMSEs of *a* and mean RMSE of *b* by group. Group 2 unique items had the highest level of mean RMSEs, followed by group 1 unique items and then common items. The differences started from the second decimal place between groups in general. The mean RMSEs were larger than 0.1 when the sample size was small at 500, and smaller than 0.1 when the sample size was large at 3000. The trends of mean biases of *a* and *b* by group had some differences. In group 1 unique items, *a* and *b* were all positively biased. In group 2 unique items, *a* was still positively biased, whereas *b* was negatively biased. In common items, *a* and *b* were positively biased in most cases except for the case of 15 common items and 500 students, in which *b* had a negative mean bias. The trends of mean RMSEs of *a* and *b* were similar across three groups. However, in group 2 unique items, the mean RMSEs of *a* and *b* were clustered together and rank highest in magnitude among groups

To summarize, the biases were negligible under null condition where all common items were invariant items. Under null condition, linking was adequate under 2 parameter logistic model using Stocking-Lord method and common-item nonequivalent groups design with as few as 500 participants in each group and 20% of items as common items.

4.2. Uniform DIF Against Group 2 Condition

Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* over 100 replications are shown for conditions with 10 common items containing 2, 4, and 6 number of small uniform DIF items each time against group 2 across three sample sizes in Figure 3. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 15 common items containing 3, 6, and 9 numbers of small uniform DIF items each time against group 2 across three sample sizes in Figure 4. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 10 (Figure 7) and 15 (Figure 8) common items with medium uniform DIF items in Figure 7 and 8, respectively. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 10 and 15 common items with large uniform DIF items in Figure 11 and 12, respectively.

The biases of *A* were smaller than 0.05 in absolute value across conditions. The biases of *B* were larger than 0.05 in absolute value across all conditions. The RMSEs of *A* were smaller than 0.1 across cases. The RMSESs of *B* were larger than 0.1 in most cases. The mean biases of *a* and *b* were smaller than 0.06 and 0.01 in absolute values under small and

medium uniform DIF against group 2, respectively. The mean biases of *b* were larger than 0.01 in absolute value when sample size was 500 under large uniform DIF against group 2. Item discrimination *a* was positively biased in most cases, whereas item location *b* was negatively biased in most cases. The mean RMSEs of *a* were close to each other under the same sample size across different number of DIF items. The mean RMSEs of *b* were dispersed under the same sample size across different number of DIF items. Generally, as the sample size increased, the biases and RMSEs decreased, a trend more obvious with mean RMSEs of *a* and *b*. The sample size effect leveled off at 1000 across conditions for recovery on *A* and *B*, and *a* and *b*. As the number of common items increased, biases and RMSES decreased slightly. As the number of DIF items increased, biases and RMSEs increased quickly. As the magnitude of DIF increased, biases and RMSEs also increased quickly.

Mean biases and mean RMSEs of *a* and *b* by group — group 1 unique item, group 2 unique item, and common item — are shown for conditions with 10 and 15 common items with small uniform DIF items each time against group 2 across three sample sizes in Figure 5 and 6, respectively. Mean biases and mean RMSEs of *a* and *b* by group are shown for conditions with 10 and 15 common items with medium uniform DIF in Figure 9 and 10, respectively. Mean biases and mean RMSEs of *a* and *b* by group are shown for conditions with 10 and 15 common items with large uniform DIF in Figure 13 and 14, respectively.

There were no significant differences on the magnitude of mean biases of *a* by group and mean biases of *b* by group. The mean biases of *a* were all smaller than 0.06 in

absolute value. The mean biases of *b* were smaller than 0.01 in absolute value in most cases. Differences among groups appeared on the third decimal place. Common items had the most dispersed mean biases of *a*, and mean biases of *b*, compared to the other two groups. There were differences on the magnitude of mean RMSEs of *a* by group, and mean RMSEs of *b* by group. The differences appeared on the second place between groups in general. Within common item group, the mean biases and mean RMSEs of *a* and *b* were sensitive to condition changes, i.e., increased as the DIF magnitude increased and decreased as the number of common items increased. The mean RMSEs were larger than 0.1 when the sample size 500, smaller than 0.1 when the sample size was 3000. The trends of mean biases of *a* and *b* by group had some differences. In group 1 unique items, *a* and *b* were positively biased in most cases. In group 2 unique items, *a* was still positively biased, while *b* was negatively biased. In common items, *a* became negatively biased in some cases, and *b* was mainly negatively biased. The trends of mean RMSEs of *a* and *b* were similar across three groups. However, in group 2 unique items, the mean RMSEs of *a* and *b* were more clustered together. Compared with null conditions, the biggest change happened within common items. Both the mean biases and mean RMSEs of *a* and *b* for common items increased in magnitude and became more dispersed under uniform DIF against group 2.

To summarize, under uniform DIF against group 2 students, the biases of *A* remained unaffected in most cases. The biases of *B* were larger than acceptable value across conditions. The mean biases of *b* were large when sample size was small and the DIF magnitude was large. Compared with group 1 and group 2 unique items, mean biases

and mean RMSEs of *a* and *b* in common items were affected most when uniform DIF presented.

4.3. Uniform DIF Favoring Group 2 Condition

Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* over 100 replications are shown for conditions with 10 common items containing 2, 4, and 6 number of small uniform DIF items each time favoring group 2 across three sample sizes in Figure 15. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 15 common items containing 3, 6, and 9 numbers of small uniform DIF items each time favoring group 2 across three sample sizes in Figure 16. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 10 and 15 common items with medium uniform DIF items in Figure 19 and 20, respectively. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 10 and 15 common items with large uniform DIF items in Figure 23 and 24, respectively.

The biases of *A* were smaller than 0.05 in absolute value and clustered. The biases of *B* were larger than 0.05 in absolute value across all conditions. The RMSEs of *A* were smaller than 0.1 and clustered across cases. The RMSEs of *B* were larger than 0.1 and scattered in most cases. The mean biases of *a* and *b* were smaller than 0.06 and 0.01 in absolute values across uniform DIF conditions, respectively. Item discrimination *a* was positively biased in most cases, while the item location *b* was negatively biased in most cases. The mean RMSEs of *a* were close to each other under the same sample size across

different number of DIF items. The mean RMSEs of *b* were dispersed under the same sample size across different number of DIF items. Generally, as the sample size increased, the biases and RMSEs decreased, a pattern more obvious with mean RMSEs of *a* and *b*. The sample size effect leveled off at 1000 across conditions for recovery on *A* and *B*, and *a* and *b*. As the number of common items increased, biases and RMSES decreased slightly. As the number of DIF items increased, biases and RMSES increased quickly. Similarly, as the magnitude of DIF increased, biases and RMSES increased quickly. Compared with results shown in the section of uniform DIF against group 2 conditions where *B* was negatively biased and *A* was positively biased, the bias of *A* and *B* among favoring group 2 conditions had switched in bias direction, i.e., *B* was positively biased and *A* was negatively biased.

Mean biases and mean RMSEs of *a* and *b* by group — group 1 unique item, group 2 unique item, and common item — are shown for conditions with 10 and 15 common items with small uniform DIF items each time favoring group 2 across three sample sizes in Figure 17 and 18, respectively. Mean biases and mean RMSEs of *a* and *b* by group are shown for conditions with 10 and 15 common items with medium uniform DIF items in Figure 21 and 22, respectively. Mean biases and mean RMSEs of *a* and *b* by group are shown for conditions with 10 and 15 common items with large uniform DIF items in Figure 25 and 26, respectively.

There were no significant differences on the magnitude of mean biases of *a* by group, and mean biases of *b* by group. The mean biases of *a* were all smaller than 0.06 in absolute value. The mean biases of *b* were smaller than 0.01 in most cases. Differences

between groups appeared on the third decimal place. Common items had the most dispersed mean biases of *a* and *b*, compared to the other two groups. There were some differences on the magnitude of mean RMSEs of *a* by group, and mean RMSEs of *b* by group. The differences appeared on the second decimal place between groups in general. Within common item group, the mean biases and mean RMSEs of *a* and *b* were sensitive to condition changes, i.e., increased as the DIF magnitude increased and decreased as the number of common items increased. Generally, the mean RMSEs of *a* and *b* were larger than 0.1 when the sample size was 500, smaller than 0.1 when the sample size was 3000, except for the mean RMSEs of *b* in common items, which were larger than 0.1 in most cases. The trends of mean biases of *a* by group, and mean biases of *b* by group had some differences. In group 1 unique items, both *a* and *b* were positively biased in most cases. In group 2 unique items, *a* was still positively biased, while *b* was negatively biased. In common item group, both *a* and *b* were positively biased in most cases. The trends of mean RMSEs of *a* and *b* were similar across three groups. However, compared with group 1 and common item group, the mean RMSEs of *a* and *b* were clustered together in group 2. Compared with null conditions, the biggest change happened within common items. Both the mean biases and mean RMSEs of *a* and *b* for common items increased in magnitude and became more dispersed under uniform DIF favoring group 2.

The magnitudes and trends described under uniform DIF favoring group 2 were like those under uniform DIF against group 2. The biases of *A* remained unaffected in most cases. The biases of *B* were larger than acceptable value across conditions. The mean biases of *b* were large when sample size was small and the DIF magnitude was large.

Compared with group 1 and group 2 unique items, mean biases and mean RMSEs of *a* and *b* in common items were affected most when uniform DIF presented.

4.4. Uniform and Nonuniform DIF Against Group 2 Condition

Biases and RMSEs of linking constants *A* and *B,* and overall mean biases and mean RMSEs of *a* and *b* over 100 replications are shown for conditions with 10 common items containing 2, 4, and 6 number of small uniform DIF and small nonuniform DIF items each time against group 2 across three sample sizes in Figure 27. Biases and RMSEs of linking constants *A* and *B,* and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 15 common items containing 3, 6, and 9 numbers of small uniform DIF and small nonuniform DIF items each time against group 2 across three sample sizes in Figure 28. Biases and RMSEs of linking constants *A* and *B,* and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 10 and 15 common items with medium uniform DIF and small nonuniform DIF items in Figure 31 and 32, respectively. Biases and RMSEs of linking constants *A* and *B,* and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 10 and 15 common items with large uniform DIF and small nonuniform DIF items in Figure 35 and 36, respectively.

The biases of A were larger than 0.05 in absolute value in most cases. Except for 20% (2 in 10 common items and 3 in 15 common items) DIF item conditions with small uniform and small nonuniform DIF, biases of *B* were much larger than 0.05 in absolute value. The RMSEs of *A* were around 0.1 across cases. The RMSEs of *B* were much larger than 0.1 in most cases, with some cases under small uniform and small nonuniform DIF around 0.1. The mean biases of *b* were smaller than 0.01 under small uniform and small

nonuniform DIF conditions, but larger than 0.01 in absolute value under medium and large uniform DIF and small nonuniform DIF conditions. The mean biases of *a* were smaller than 0.06 across all conditions. Item discrimination *a* was positively biased in most cases, while item location *b* was negatively biased in most cases. The mean RMSEs of *a* were relatively small and clustered under the same sample size across different number of DIF items. The mean RMSEs of *b* were relatively large and dispersed under the same sample size across different number of DIF items. Generally, as the sample size increased, the biases and RMSEs decreased, a pattern more obvious on mean RMSEs of *a* and *b*. The sample size effect leveled off at 1000 across conditions for recovery on *A* and *B*, and *a* and *b*. As the number of common items increased, biases and RMSES decreased slightly. As the number of DIF items increased, biases and RMSES increased quickly. As the magnitude of DIF increased, biases and RMSEs also increased quickly. Bringing nonuniform DIF into view, biases of *A* had significant increase compared with conditions that had uniform DIF only.

Mean biases and mean RMSEs of *a* and *b* by group — group 1 unique item, group 2 unique item, and common item — are shown for conditions with 10 and 15 common items with small uniform DIF items each time against group 2 across three sample sizes in Figure 29 and 30, respectively. Mean biases and mean RMSEs of *a* and *b* by group are shown for conditions with 10 and 15 common items with medium uniform DIF items in Figure 33 and 34, respectively. Mean biases and mean RMSEs of *a* and *b* by group are shown for conditions with 10 and 15 common items with large uniform DIF items in Figure 37 and 38, respectively.

There were no significant differences on the magnitude of mean biases of *a* by group, and mean biases of *b* by group. The mean biases of *a* were all smaller than 0.06 in absolute values. The mean biases of b were smaller than 0.01 in absolute values in most cases. Differences between groups appeared on the third decimal place. Common items had the most dispersed mean biases of *a* and *b*, compared with the other two groups. There were some differences on the magnitude of mean RMSEs of *a* by group, and mean RMSEs of *b* by group. The differences appeared on the first and second decimal places. Within common item group, the mean biases and mean RMSEs of *a* and *b* were sensitive to condition changes, i.e., increased as the DIF magnitude increased and decreased as the number of common items increased. The mean RMSEs were larger than 0.1 when the sample size was 500, smaller than 0.1 when the sample size was 3000, except for the mean RMSEs of *b* in common items, which were larger than 0.1 in most cases. The trends of mean biases of *a* and *b* by group had some differences. In group 1 unique items, *a* and *b* were positively biased in most cases. In group 2 unique items, *a* was still positively biased, while *b* was negatively biased. In common item group, *a* was positively biased in most cases, and b was negatively biased in most cases. The trends of mean RMSEs of *a* and *b* were similar across three groups. However, compared with group 1 and common item group, the mean RMSEs of *a* and *b* were clustered together in group 2. Compared with null conditions, the biggest change also happened within common items. Both the mean biases and mean RMSEs of *a* and *b* for common items increased in magnitude and became more dispersed under uniform and nonuniform DIF against group 2.

With both uniform and nonuniform DIF items presented, biases of *A* became larger than 0.05 in absolute value for most cases. Compared with group 1 and group 2 unique items, mean biases and mean RMSEs of *a* and *b* in common items were affected most when both uniform and nonuniform DIF presented. Other patterns were similar to those observed in previous sections.

4.5. Uniform and Nonuniform DIF Favoring Group 2 Condition

Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* over 100 replications are shown for conditions with 10 common items containing 2, 4, and 6 number of small uniform DIF and small nonuniform DIF items each time favoring group 2 across three sample sizes in Figure 39. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 15 common items containing 3, 6, and 9 numbers of small uniform DIF and small nonuniform DIF items each time favoring group 2 across three sample sizes in Figure 40. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 10 (Figure 43) and 15 (Figure 44) common items with medium uniform DIF and small nonuniform DIF items favoring group 2 in Figure 43 and 44, respectively. Biases and RMSEs of linking constants *A* and *B*, and overall mean biases and mean RMSEs of *a* and *b* are shown for conditions with 10 and 15 common items with large uniform DIF and small nonuniform DIF items favoring group 2 in Figure 47 and 48, respectively.

Biases of A were larger than 0.05 in absolute value across conditions. Biases of *B* were much larger than 0.05 in absolute value, except for 20% (2 in 10 common items and

3 in 15 common items) DIF item conditions with small and medium uniform DIF and small nonuniform DIF favoring group 2. The RMSEs of *A* were higher than 0.1 in most cases. The RMSEs of *B* were larger than 0.1, and much larger than 0.1 under conditions of medium and large uniform DIF and small nonuniform DIF situations. Both RMSEs of *A* and *B* became more scattered compared to results in previous sections, but still RMSEs of *B* more dispersed than RMSEs of *A*. The mean biases of *a* and *b* were no larger than 0.01 and 0.06, respectively, in most cases. Item discrimination *a* was positively biased in most cases, while the item location *b* was negatively biased in most cases. The mean RMSEs of both *a* and *b* were more dispersed compared with results in uniform DIF only sections. Generally, as the sample size increased, the biases and RMSEs decreased, a pattern more obvious on mean RMSEs of *a* and *b*. The sample size effect leveled off at 1000 across conditions for recovery on *A* and *B*, and *a* and *b*. As the number of common items increased, biases and RMSES decreased slightly. As the number of DIF items increased, biases and RMSES increased quickly. As the magnitude of DIF increased, biases and RMSES also increased quickly. Bringing nonuniform DIF into view, the biases of *A* had a significant increase and became larger than acceptable value, compared with conditions that had uniform DIF only. Compared with results shown in the section of uniform and nonuniform DIF against group 2, where *B* was negatively biased and *A* was positively biased, the biases of *A* and *B* among favoring group 2 conditions had switched in bias direction, i.e., *A* was negatively biased and *B* was positively biased.

Mean biases and mean RMSEs of *a* and *b* by group — group 1 unique item, group 2 unique item, and common item — are shown for conditions with 10 and 15 common

items with small uniform DIF items each time favoring group 2 across three sample sizes in Figure 41 and 42, respectively. Mean biases and mean RMSEs of *a* and *b* by group are shown for conditions with 10 and 15 common items with medium uniform DIF in Figure 45 and 46, respectively. Mean biases and mean RMSEs of *a* and *b* by group are shown for conditions with 10 and 15 common items with large uniform DIF in Figure 49 and 50, respectively.

There were no significant differences on the magnitude of mean biases of *a* by group, and mean biases of *b* by group. The mean biases of *a* were all smaller than 0.06 in absolute values. The mean biases of b were smaller than 0.01 in absolute values in most cases. Differences between groups appeared on the second and the third decimal places. Common items had the most dispersed mean biases of *a* and *b,* compared with the other two groups. There were differences on the magnitude of mean RMSEs of *a* and *b* between groups. The differences appeared on the first and the second decimal places. Within common item group, the mean biases and mean RMSEs of *a* and *b* were sensitive to condition changes, i.e., increased as the DIF magnitude increased and decreased as the number of common items increased. The mean RMSEs of *a* and *b* were larger than 0.1 when the sample size was 500, smaller than 0.1 when the sample size was 3000, except for the mean RMSEs of *a* and *b* in common items. For example, the mean RMSEs of *b* in common items were larger than 0.1 in most cases, except for small uniform DIF and large sample size conditions. The trends of mean biases of *a* and *b* by group had some differences. In group 1 unique items, *a* and *b* were positively biased in most cases. In group 2 unique items, *a* was still positively biased, while *b* was negatively biased. In

common item group, *a* and *b* were both positively and negatively biased. The trends of mean RMSEs of *a* and *b* were similar across three groups. However, compared with group 1 and common item group, the mean RMSEs of *a* and *b* were clustered together in group 2. The mean RMSEs of *a* and *b* were most dispersed in common item group. Compared with null conditions, the biggest change also happened within common items. Both the mean biases and mean RMSEs of *a* and *b* for common items increased in magnitude and became more dispersed under uniform and nonuniform DIF favoring group 2.

To summarize, the trends and patterns observed under both uniform and nonuniform DIF favoring group 2 were close to those under uniform and nonuniform DIF against group 2. The biases of *A* became larger than 0.05 in most cases. Compared with group 1 and group 2 unique items, mean biases and mean RMSEs of *a* and *b* in common items were affected most.

5. DISCUSSION

The purpose of this dissertation is to evaluate DIF effect on common-item nonequivalent groups design linking under the two parameters logistic model. Simulation studies were conducted taking sample size, ratio of common item, ratio of DIF item, magnitude of DIF item, form of DIF, and direction of DIF into consideration. Recovery quality of linking constants and item parameters was evaluated using biases and RMSEs. Relative biases were used to scale the biases term to give a concrete idea of bias magnitude.

5.1. Good Recovery Under Null Condition

Under null conditions with all common items invariant, biases of both *A* and *B* were small across all conditions. Small RMSEs of *A* and *B* indicated good recovery on *A* and *B*. Small mean RMSEs of *a* and *b* indicated good recovery on item parameter *a* and *b* also. The recovery of linking constant *A* and *B*, item parameter *a* and *b* were in consistency to existed research on related topics(Cohen & Kim, 1998; Kim & Cohen, 1992, 1998, 2002). Null condition results informed us that under common-item nonequivalent groups design, linking was adequate using test characteristic method with a ratio of common item at 20% and a sample size as small as 500. Good recovery under null conditions with nonequivalent groups also corroborated the previous studies on the effect of examinee ability on equating constants and item parameter recovery. Test equating was generally independent of examinee ability under both random groups and nonequivalent groups

designs across traditional and IRT equating methods (Cook & Petersen, 1987; Harris & Kolen, 1986; Skaggs & Lissitz, 1988).

5.2. Large Biases of B but Small Biases of A with Uniform DIF

With uniform DIF against or in favor of group 2 participants, biases of *B* were larger than 0.05 in absolute value across three levels of DIF magnitude, while biases of *A* were smaller than 0.05 in absolute values in most cases. Biases of *A* only became larger than 0.05 when the DIF magnitude was large and the number of DIF items was high. Biases of *B* increased quickly as the magnitude of DIF increased, changing from 0.175 (small DIF highest number of DIF items) in absolute value to 0.375 (medium DIF and highest number of DIF items) in absolute value, and to 0.55 (large DIF and highest number of DIF items) in absolute value. Biases of *A* remained unaffected by the number of DIF items and level of DIF magnitude most of time, and changes were nuance even there were. The direction of DIF, i.e., against or in favor of group 2 did not affect the magnitude of biases but the direction of biases. For example, *B* was negatively biased when having DIF items against group 2 participants and was positively biased when having DIF items in favor of group 2 participants.

With the functionality of *A* and *B* in mind, i.e., *A* as slope and *B* as intercept for equating, we might be able to interpret the biases in terms of their impact on equating results. Little biases on *A* meant the standard deviation of group 2 scores after transformation would not be biased. However, the large negative biases of *B* when against group 2 meant the mean of group 2 scores after transformation was seriously underestimated. The large positive biases of *B* when favoring group 2 meant the mean of

group 2 after transformation was seriously overestimated. Both scenarios would impair the adequacy of equating and make meaningful comparison unachieved. One possible reason of *B* other than *A* being seriously biased might be the mean as the first order statistic was more likely to be affected by outliers, while standard deviation as the second order statistic was less likely to be affected by outliers. Another possible reason could be read from Figure 51-53. With uniform DIF against group 2, it can be spotted easily in figures that students with the same ability θ would be estimated with a lower possibility of passing an item. Using true score equating as an illustration here, sum of the probabilities across test items containing DIF items would be lower than sum of probabilities across all invariant test items. The achievement of group 2 students was underestimated upon having uniform DIF items against them. Linking constant *B* would mirror that occurrence as negatively biased. With uniform DIF items favoring group 2 participants, the probabilities would be overestimated. The whole situation was to the opposite of mechanism discussed for uniform DIF against group 2 participants. Therefore, *B* was positively biased when DIF was in favor of group 2 participants.

5.3. Large Biases of A and B with Uniform and Nonuniform DIF

When both uniform and nonuniform DIF presented on the common items, biases of both *A* and *B* became large than 0.05 in most cases. Under small uniform and small nonuniform DIF conditions, biases of *B* regressed toward 0 a little bit, compared with biases of *B* when having small uniform DIF only. Under medium uniform DIF and small nonuniform DIF, and large uniform DIF and small nonuniform DIF conditions, biases of *B* were at similar level with medium uniform DIF only and large uniform DIF only

conditions, respectively. The biases of A became larger than 0.05 but remained similar across different levels of uniform DIF magnitude. Therefore, direct reason for increased biases of *A* was the presence of small nonuniform DIF. Small nonuniform DIF also had certain impact on the estimation of *B* but the effect became negligible when the uniform DIF magnitude was high enough.

The reason for large biases of *B* was the same as the one given under uniform DIF only conditions in section 5.2. One possible reason for the increase in biases of *A* might also be traced out from Figure 51-53. Given a range of ability θs, the corresponding range of probabilities under DIF items on y axis would be more dispersed than the corresponding range of probabilities under invariant items. Therefore, when having nonuniform DIF against group 2, the dispersion of summed probabilities across test items containing DIF items would be inflated. Linking constant A would mirror that occurrence as positively biased. The same mechanism applied to biases of *A* under uniform and nonuniform DIF items in favor of group 2. But the direction of biases of *A* switched.

5.4. Small Mean Biases and Mean RMSEs of a and b

The collapsed mean biases and mean RMSEs of item parameters *a* and *b* were generally low across all conditions, no matter under uniform DIF only conditions or under both uniform and nonuniform DIF conditions. The mean biases and mean RMSEs of *a* and *b* were also small and did not differ very much by group 1 unique item, group 2 unique item, and common item. However, the small mean biases and mean RMSEs would accumulate across items. Considering 50 items on each form, mean biases and mean RMSEs would be 50 times larger on the test level. Still using the true score equating as

illustration, probabilities would be slightly biased on each item, but the sum of probabilities across test items would be largely biased.

5.5. Sensitive Mean Biases and Mean RMSEs of a and b Within Common Item

Mean biases and mean RMSEs of *a* and *b* with common items were most sensitive to condition changes, like the number of DIF items, and the level of DIF magnitude. Since DIF items only presented within common items, no wonder that item parameter recovery on common items were most affected by changes in DIF conditions. Also, the use of the average of the separate estimates on common items in two groups as the final estimates of common item parameters (Hambleton, Swaminathan, & Rogers, 1991) might explain why the trends of mean biases of *a* and *b* within common item group looked different (more dispersed) from those of group 1 and group 2 unique items. Another possible reason for the pattern observed might be the linking and calibrating plan used here (Battauz, 2015). In this study, everything was put back to the generating scale to make evaluations. Under this plan, group 1 had the least transformation, group 2 had the medium, and common item group had the most transformation.

5.6. Limitations and Future Research

Though explored quite a lot conditions upon having uniform and nonuniform DIF against or in favor of a group, many more conditions or factors were left untreated. One possible extension is to study the effect under different IRT models. After understanding the situations and performances with dichotomous data, the next step would be polytomous data where graded response model is of great interest. The same common item DIF effect on linking could be tested under graded response model to see whether the

results and conclusions in 2PL model are still applicable. DIF could be expanded to have more years or groups of participants with varied levels of differences in ability profile. There are many factors affecting equating coefficients and the length of chain was one of them (Battauz, 2015). In this study, only one length of chain was tested. Future study could investigate more complex chain and observe the change in biases and RMSEs. Also, the current investigation only showed common item DIF effect on nonequivalent groups design linking. The study did not take a further step of removing the DIF items and quantifying the improvement in linking constants and item parameter recovery. Future study could be performed either removing DIF items directly or in a way that identifying and then removing DIF items using various DIF detection methods.

6. CONCLUSION

To conclude, the effects of DIF common items were substantial on nonequivalent groups design linking. When only having uniform DIF against or in favor of a group, the mean of transformed scores would be either seriously underestimated or overestimated due to negatively biased or positively biased linking constant *B*, respectively. When having both uniform and nonuniform DIF against or in favor of a group, the mean and standard deviation of transformed scores would be either seriously underestimated or overestimated due to nonnegligible biases of both *A* and *B*. Under both scenarios, adequate equating was not achievable even at a large sample size of 3000 in each group. The bias increased rapidly as the number of DIF items and the level of DIF magnitude increased. The mean bias and mean RMSE of *a* and *b* were small across conditions. However, the small item level mean bias and mean RMSE would augment on the test level. Therefore, DIF common items effects were not only on linking constant, common items, but on unique items in each group. The juxtaposition of good recovery under null conditions and seriously biased results under DIF conditions underscores the importance of measurement invariance of common items to nonequivalent groups. To link adequately, it is always recommended to check DIF between concerned groups.

REFERENCES

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581-594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Battauz, M. (2015). Factors affecting the variability of IRT equating coefficients. *Statistica Neerlandica, 69*(2), 85-101.

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Bowles, R. P. (2016). Review of Test Equating, Scaling, and Linking: Methods and Practices (3rd ed.), by Michael J. Kolen and Robert L. Brennan. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(1), 155–156

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*(3), 253-260.

Chu, K.-l., & Kamata, A. (2004). Test equating in the presence of DIF items. *Journal of Applied Measurement, 6*(3), 342-354.

Cohen, A. S., & Kim, S. H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*(2), 116-130.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of applied psychology, 85*(3), 451-461.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*(3), 225-244.

Divgi, D. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9*(4), 413-415.

Dorans, N. J., & Holland, P. W. (1992). DIF Detection and Description: Mantel-Haenszel and Standardization. *ETS Research Report Series, 1992*(1).

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *ETS Research Report Series, 2000*(2).

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*(1), 81-97.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of applied psychology, 72*(1), 19-29.

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the Mini-Mental State Examination. *Medical care, 44*(11), S134-S142.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of applied psychology, 74*(6), 912-921.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*: Mahwah, NJ: Lawrence Erlbaum.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2): Sage.

Hanson, B. A., & Beguin, A. A. (1999). *Separate versus concurrent estimation of IRT item parameters in the common item equating design.* (ACT research report; No. 99-08). Iowa City, IA, USA: American College Testing Program.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*(3), 195-240.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*(1), 35-43.

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

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement, 74*(4), 627-658.

Kabasakala, K. A., & Kelecioğlub, H. (2015). Effect of Differential Item Functioning on Test Equating. *Educational Sciences: Theory & Practice, 5*, 1229-1246.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51-66.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143.

Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*(1), 25-41.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*(1), 3-14.

Kolen, M. J., & Brennan, R. L. (2013). *Test equating: Methods and practices*: Springer Science & Business Media.

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

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores:* IAP.

Loyd, B. H., & Hoover, H. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research, 13*(2), 127-143.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*: Routledge.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1).

Muraki, E. (1993). Information functions of the generalized partial credit model. *ETS Research Report Series, 1993*(1), i-12.

Muraki, E. (1997). A generalized partial credit model *Handbook of modern item response theory* (pp. 153-164): Springer.

Oshima, T., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*(1), 1-17.

Park, D. G. (1988). *Investigations of item response theory item bias detection*. (Doctoral dissertation, University of Georgia).

Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14*(2), 163-173.

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. *Linking and aligning scores and scales*, 59-72.

Petersen, N. S. (2008). A discussion of population invariance of equating. *Applied Psychological Measurement, 32*(1), 98-101.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

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

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

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin, 114*(3), 552-566.

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

Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 35*(1), 139-139.

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

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*(1), 69-82.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological methods, 11*(4), 402-415.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*(4), 333-344.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society, 54*(3), 426-482.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). Specifying the characteristics of linking items used for item response theory item calibration. *ETS Research Report Series, 1987*(1).

Yi, Q., Assessment, H., Harris, D. J., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement, 32*(1), 62-80.

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

APPENDIX A

TABLES

Table 1 Item *j* response pattern

	Answer on item *j*		
	Right(1)	Wrong (0)	Total
Focal group	A_i	B_i	N_{Fi}
Reference group	C_i	D_i	N_{Ri}
Total	N_{Ti}	N_{Wi}	N_i

Table 2 Study design factors

Factor	No. Of Levels	Details
Sample size	**3**	**500, 1000, 3000 in each group**
Model	**1**	**2PL**
Common item ratio	**2**	**20% or 30% of the total items on each form**
DIF item ratio	3	20%, 40%, or 60% of the common items
Direction of DIF	2	Harder and more discriminative on Group 1; Harder and more discriminative on Group 2
Form of DIF	2	Uniform DIF or non-uniform DIF
Magnitude of uniform DIF	3	Distance of 0.3, 0.6, or 0.9 on location parameter between two groups for the same item
Magnitude of non-uniform DIF	1	Distance of 0.3 on discrimination parameter between two groups for the same item

Note. Null conditions are bold.

Table 3 Descriptive statistics of generated items

Unique Items (Group 1)			Common Items			Unique Items (Group 2)		
Item No.	a	b	Item No.	a	b	Item No.	a	b
1	1.002	-0.385	1	1.009	1.169	1	0.943	0.989
2	1.238	0.613	2	0.813	1.031	2	1.177	0.545
3	0.836	-0.016	3	1.411	0.27	3	1.576	1.173
4	0.917	0.405	4	1.643	-1.741	4	0.926	1.55
5	0.81	1.46	5	1.068	0.472	5	1.691	-0.907
6	0.915	-0.903	6	1.193	0.842	6	1.383	0.33
7	1.25	-0.865	7	1.33	1.53	7	1.319	-0.223
8	0.829	-0.393	8	1.31	-1.172	8	1.364	1.23
9	1.039	1.312	9	1.451	-0.284	9	1.442	1.961
10	0.962	0.127	10	0.926	0.068	10	1.086	2.743
…	…	…	11	0.997	0.379	…	…	…
37	0.929	-0.685	12	1.558	0.802	37	0.918	0.146
38	1.357	-0.143	13	0.914	0.807	38	1.131	1.321
39	1.356	-1.129	14	1.122	-0.993	39	0.87	0.655
40	1.383	0.241	15	1.602	-0.75	40	1.078	-0.581
Mean	**1.232**	**0.161**		**1.223**	**0.162**		**1.246**	**0.177**
SD	**0.288**	**0.88**		**0.27**	**0.958**		**0.267**	**1.073**

APPENDIX B

FIGURES

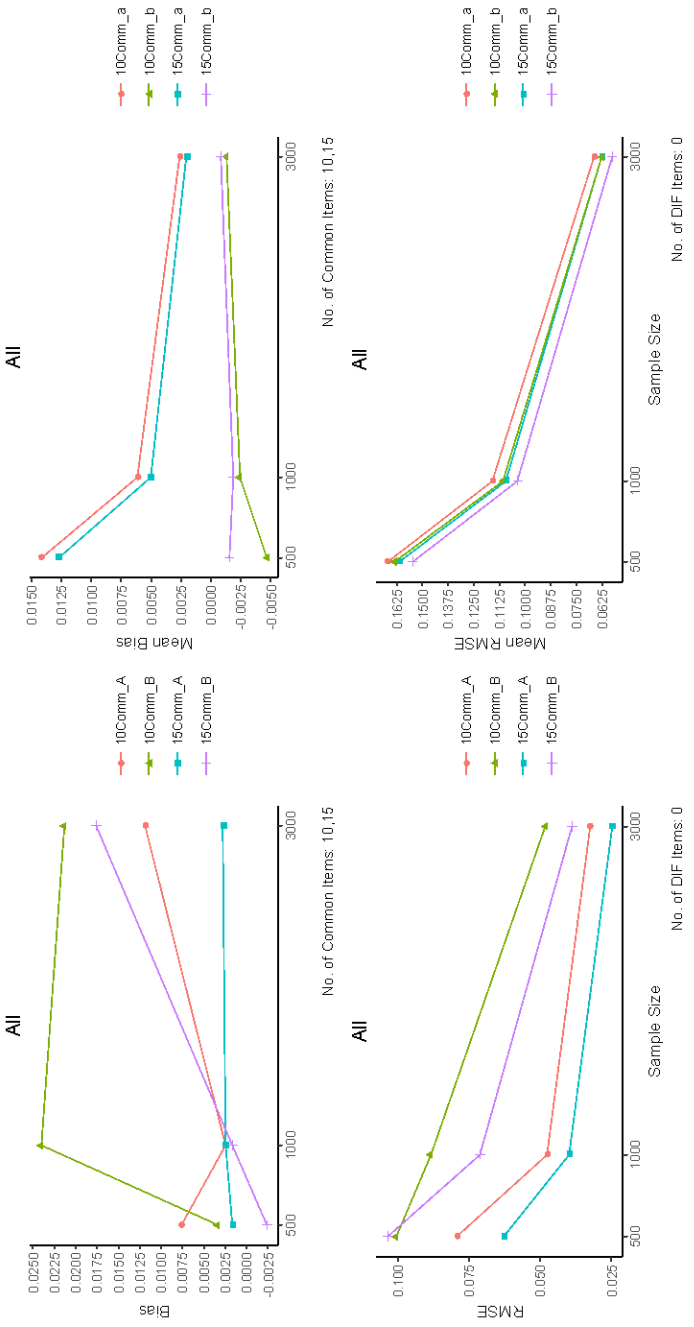

Figure 1 Null condition linking constants and item parameter recovery

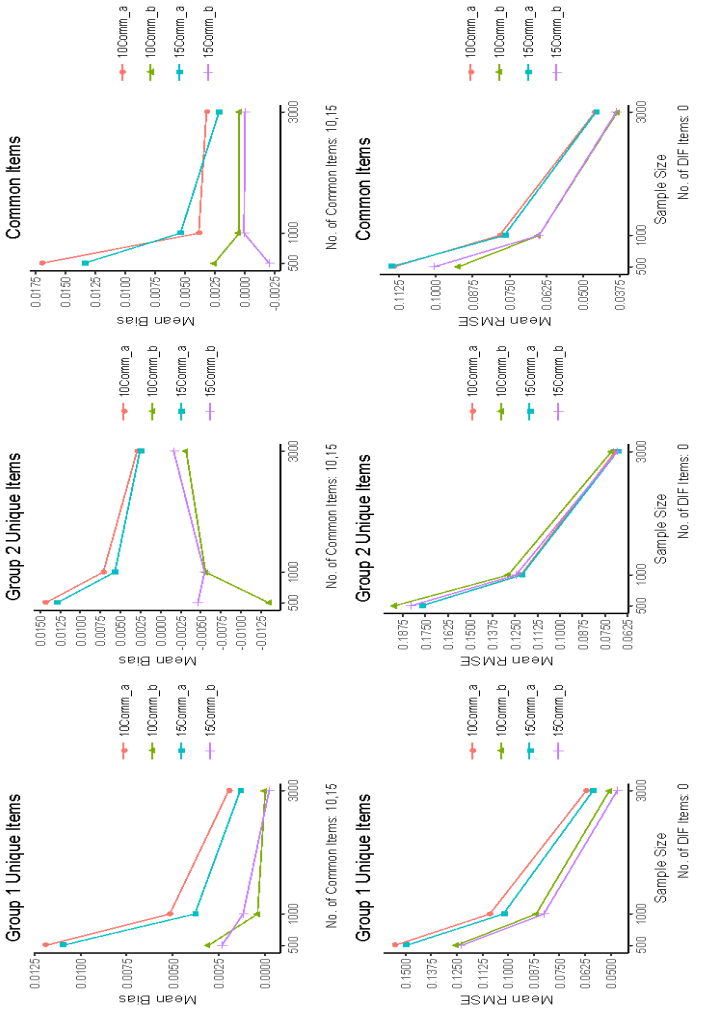

Figure 2 Null condition linking constants and item parameter recovery by groups

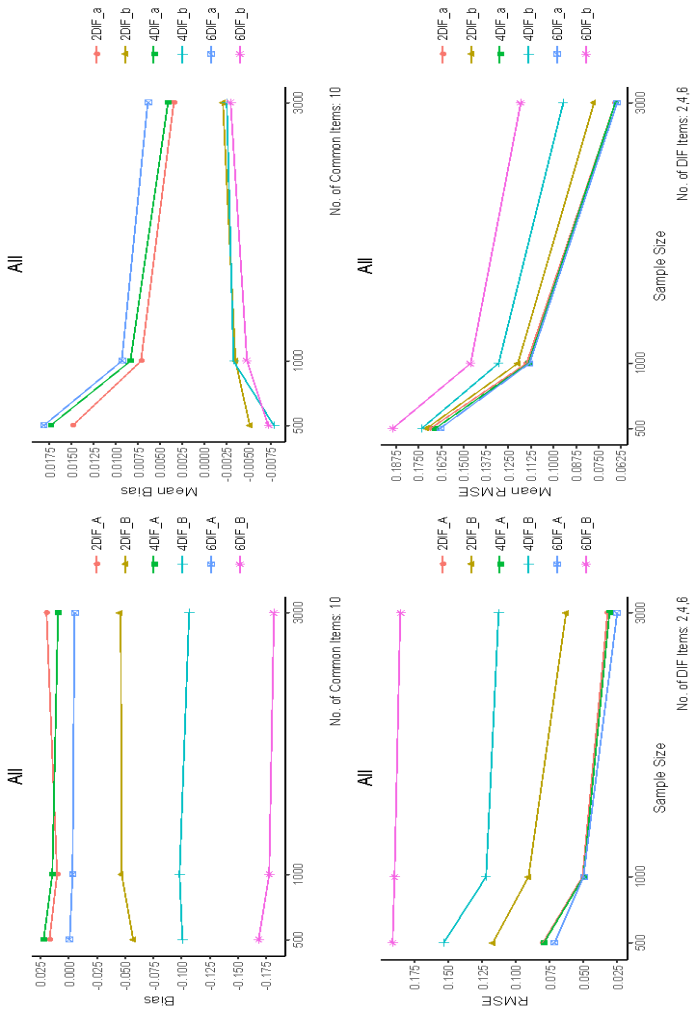

Figure 3 Small Uniform DIF Against Group 2 10 Common items Linking Constants and Item Parameter Recovery

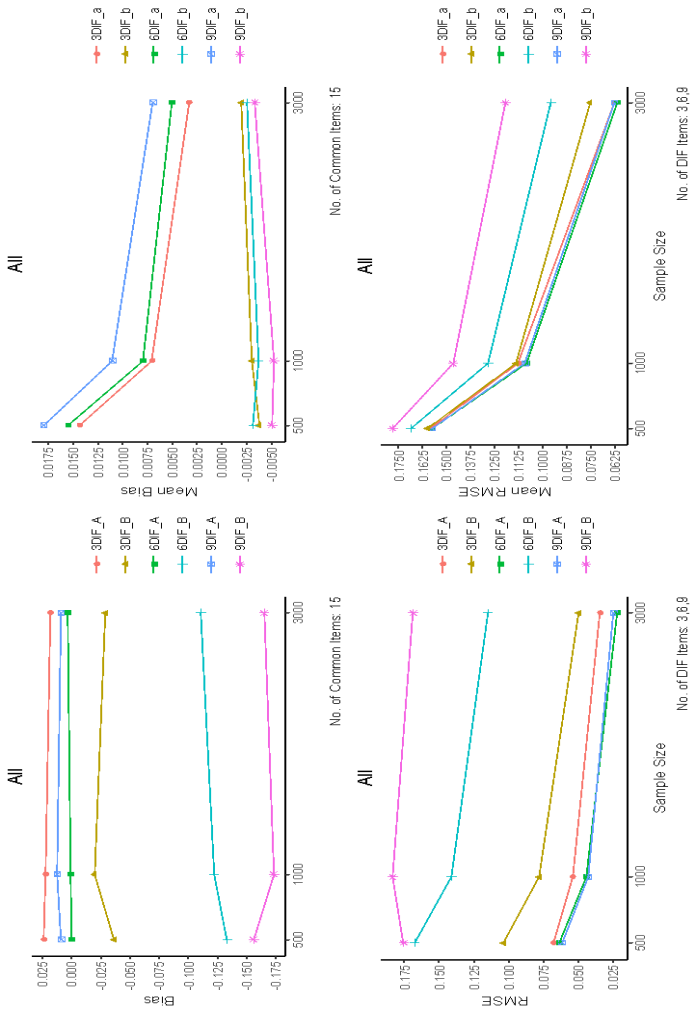

Figure 4 Small Uniform DIF Against Group 2 15 Common Items Linking Constants and Item Parameter Recovery

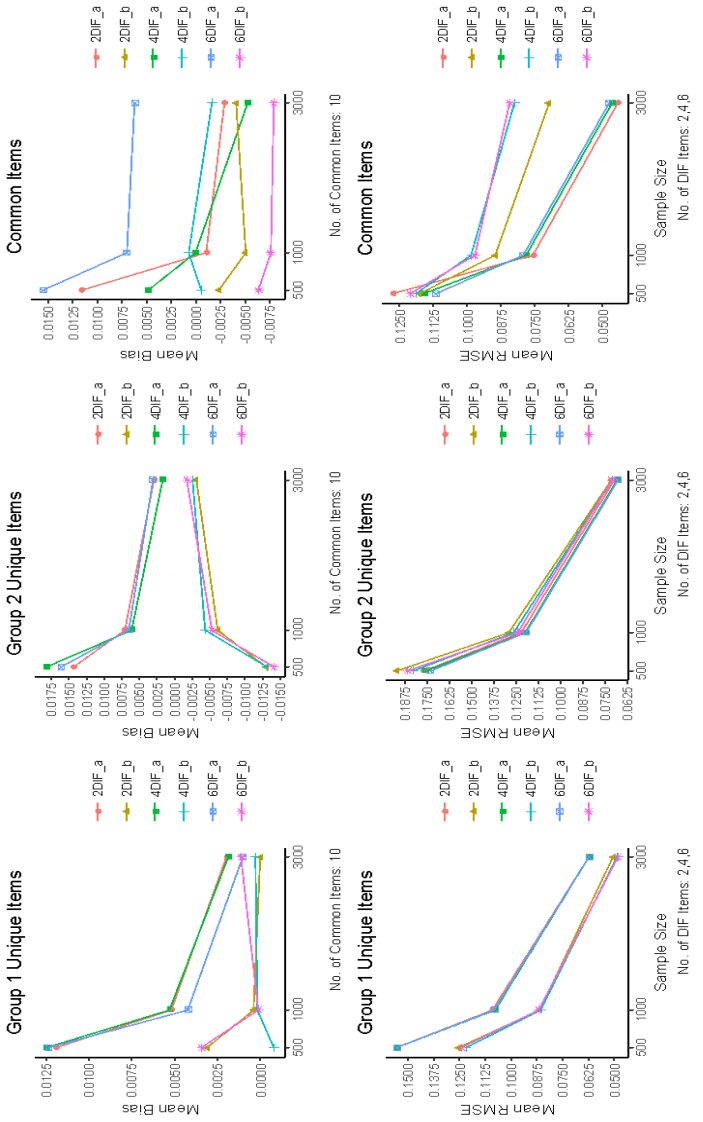

Figure 5 Small Uniform DIF Against Group 2 10 Common Items Item Parameter Recovery by Group

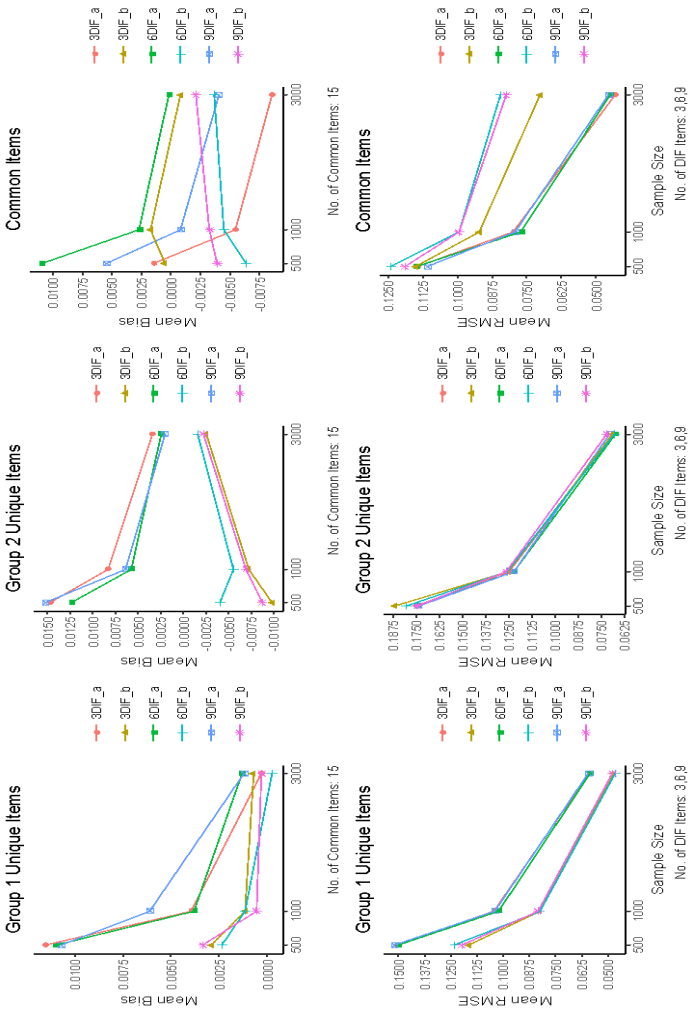

Figure 6 Small Uniform DIF Against Group 2 15 Common Items Item Parameter Recovery by Group

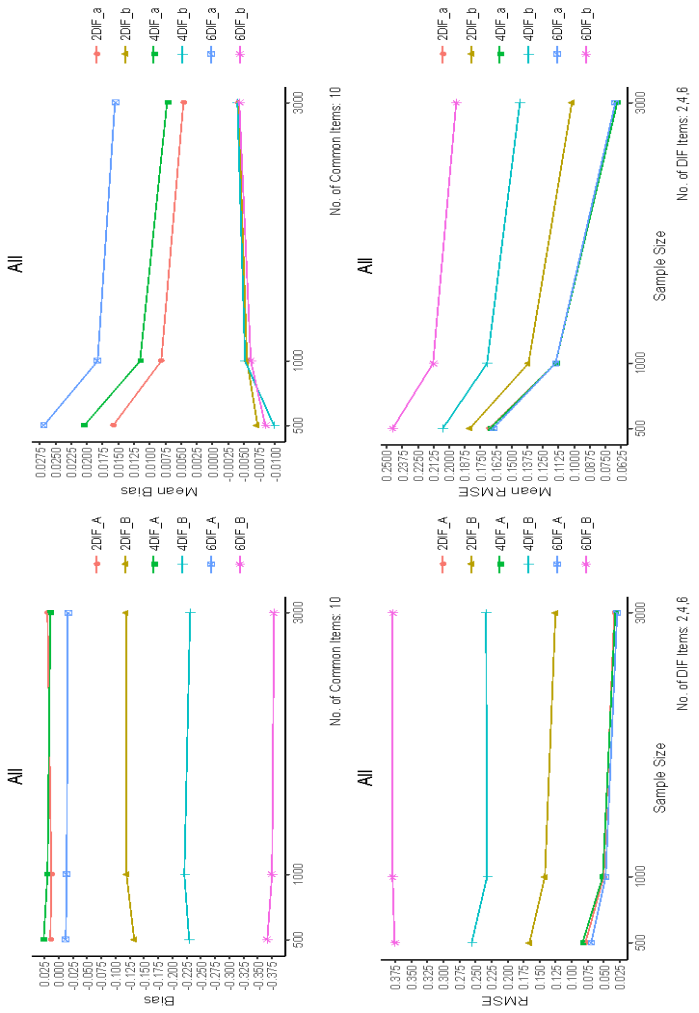

Figure 7 Medium Uniform DIF Against Group 2 10 Common Items Linking Constants and Item Parameter Recovery

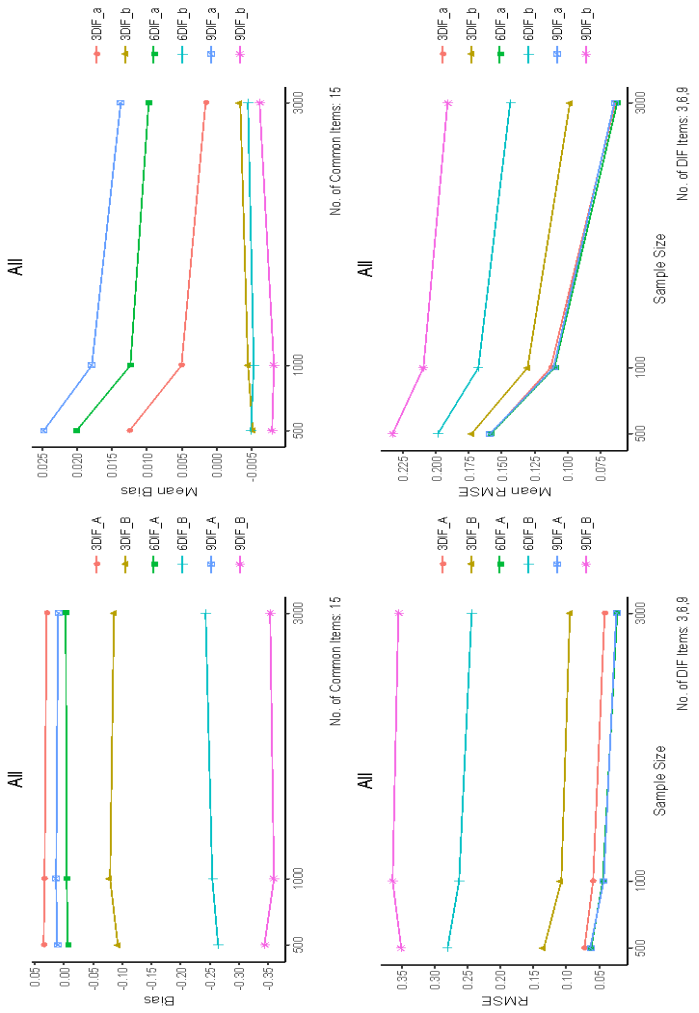

Figure 8 Medium Uniform DIF Against Group 2 15 Common Items Linking Constants and Item Parameter Recovery

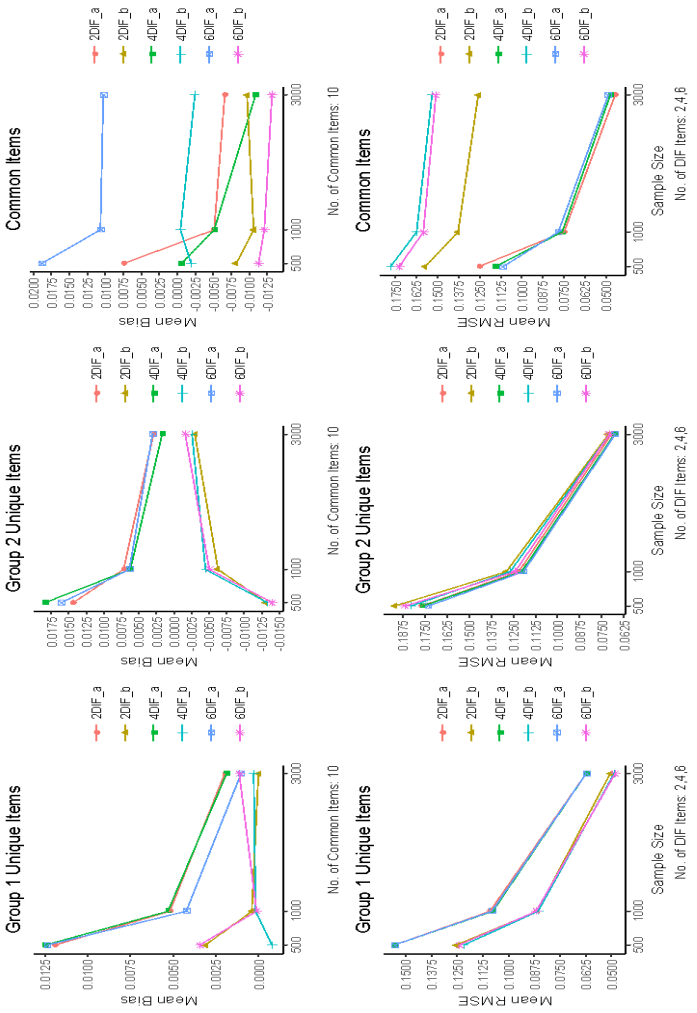

Figure 9 Medium Uniform DIF Against Group 2 10 Common Items Item Parameter Recovery by Group

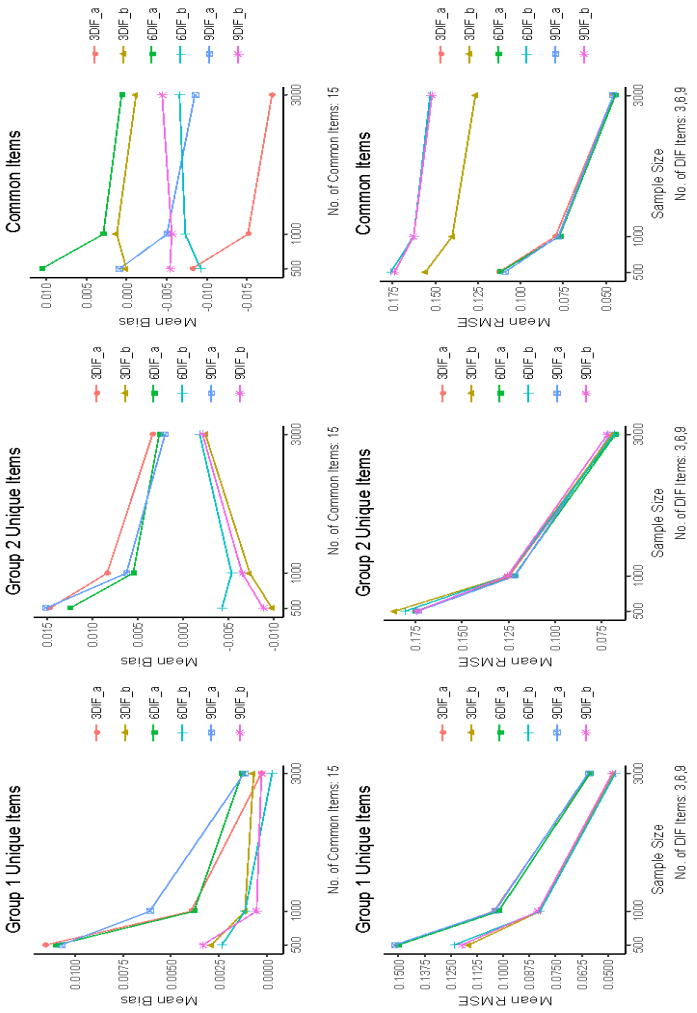

Figure 10 Medium Uniform DIF Against Group 2 15 Common Items Item Parameter Recovery by Group

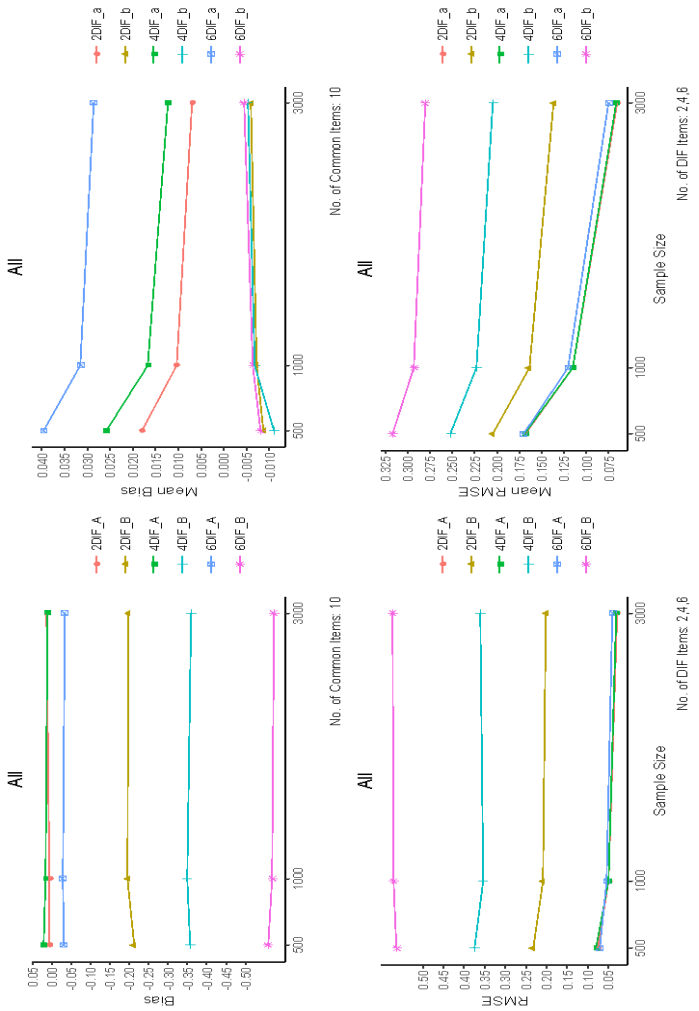

Figure 11 Large Uniform DIF Against Group 2 10 Common Items Linking Constant and Item Parameter Recovery

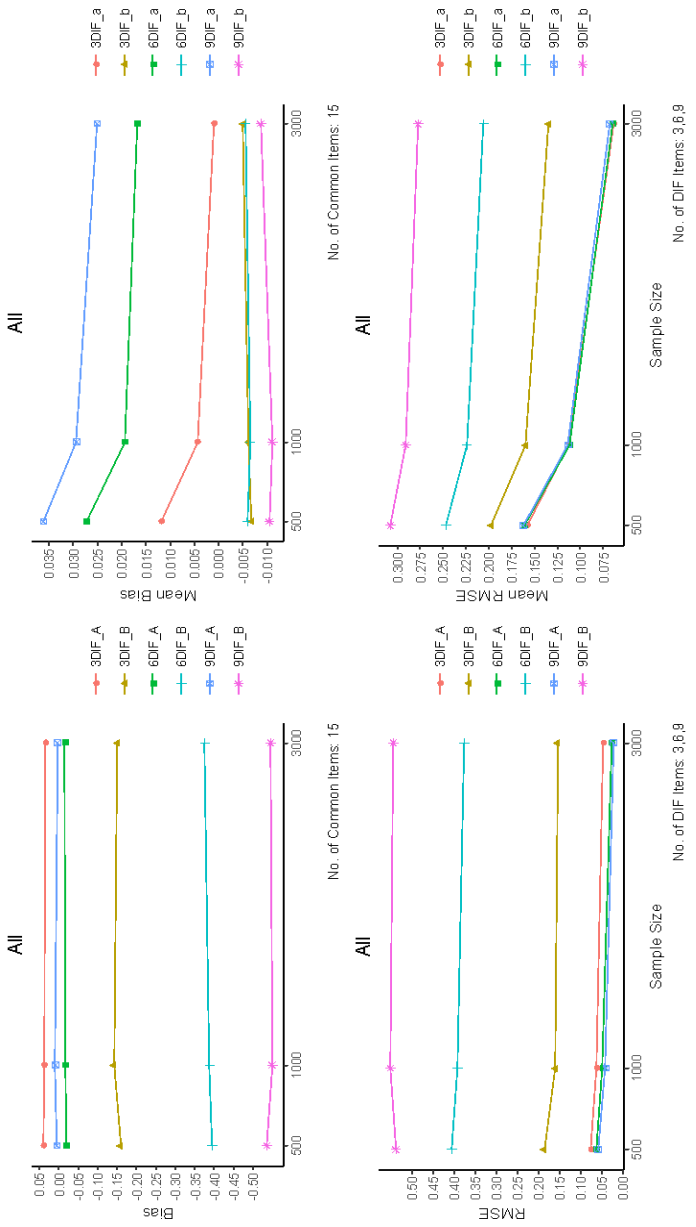

Figure 12 Large Uniform DIF Against Group 2 15 Common Items Linking Constants and Item Parameter Recovery

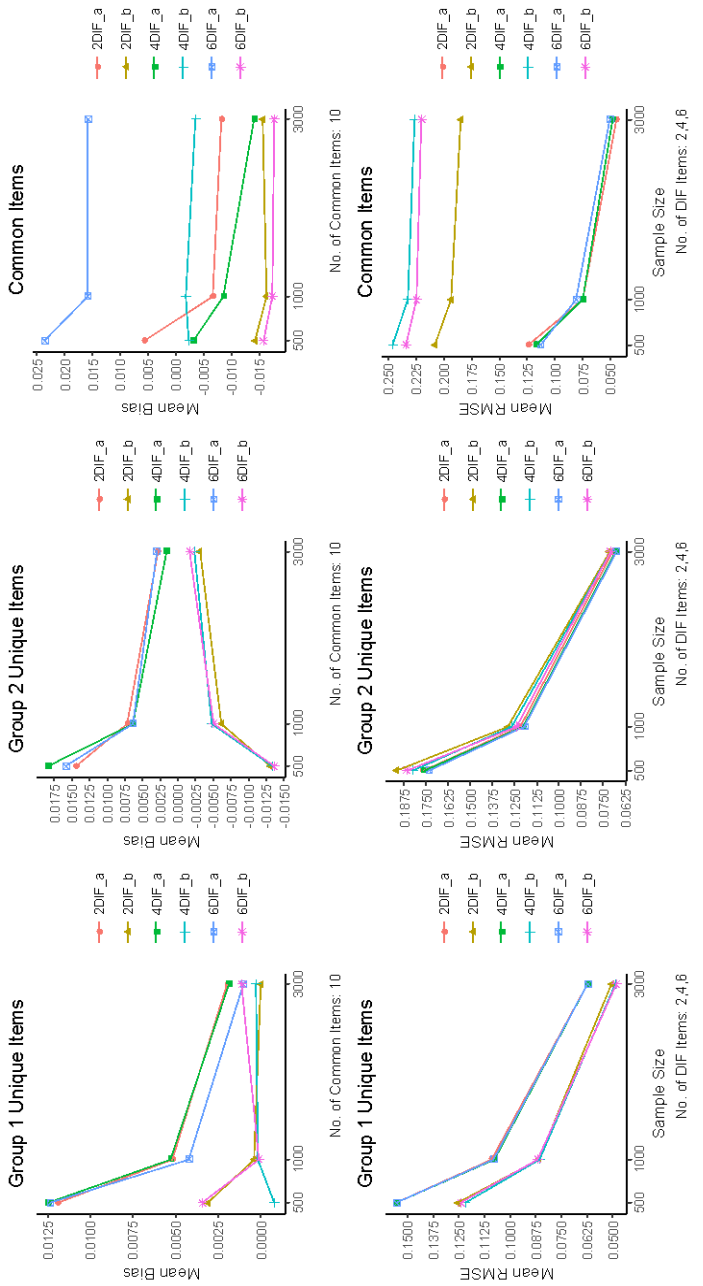

Figure 13 Large Uniform DIF Against Group 2 10 Common Items Item Parameter Recovery by Group

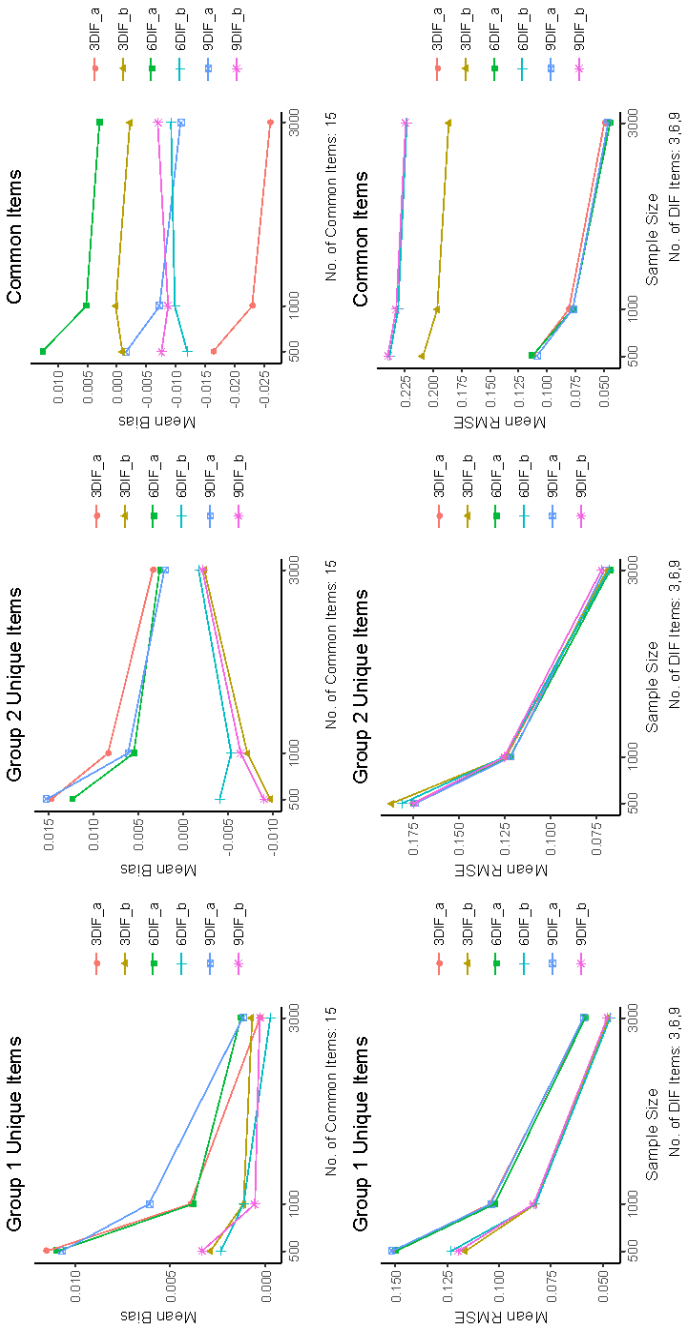

Figure 14 Large Uniform DIF Against Group 2 15 Common Items Item Parameter Recovery by Group

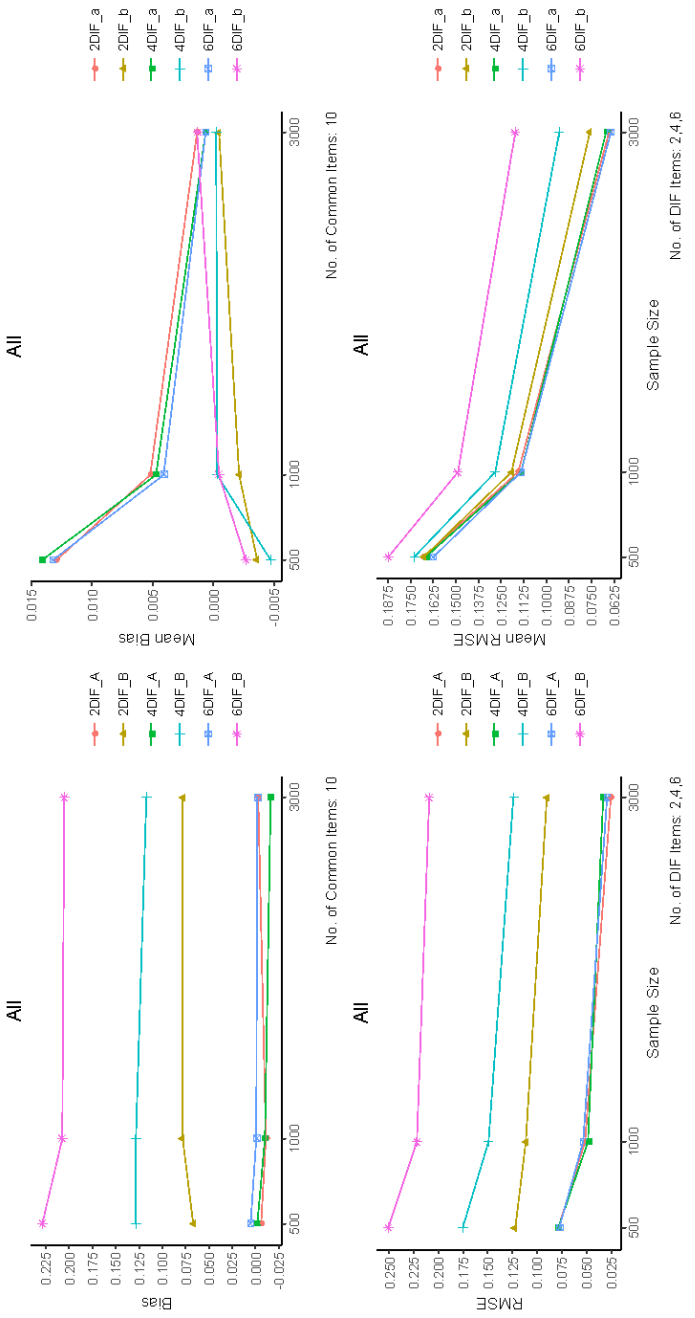

Figure 15 Small Uniform DIF Favoring Group 2 10 Common Items Linking Constants and Item Parameter Recovery

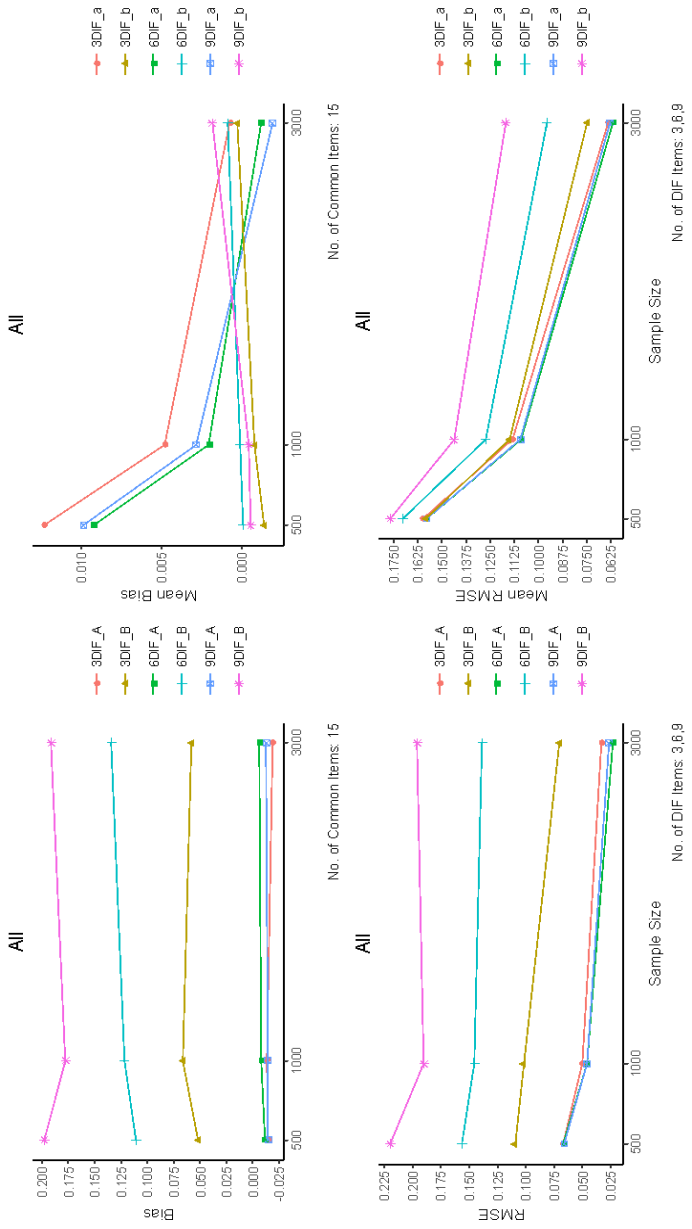

Figure 16 Small Uniform DIF Favoring Group 2 15 Common Items Linking Constants and Item Parameter Recovery

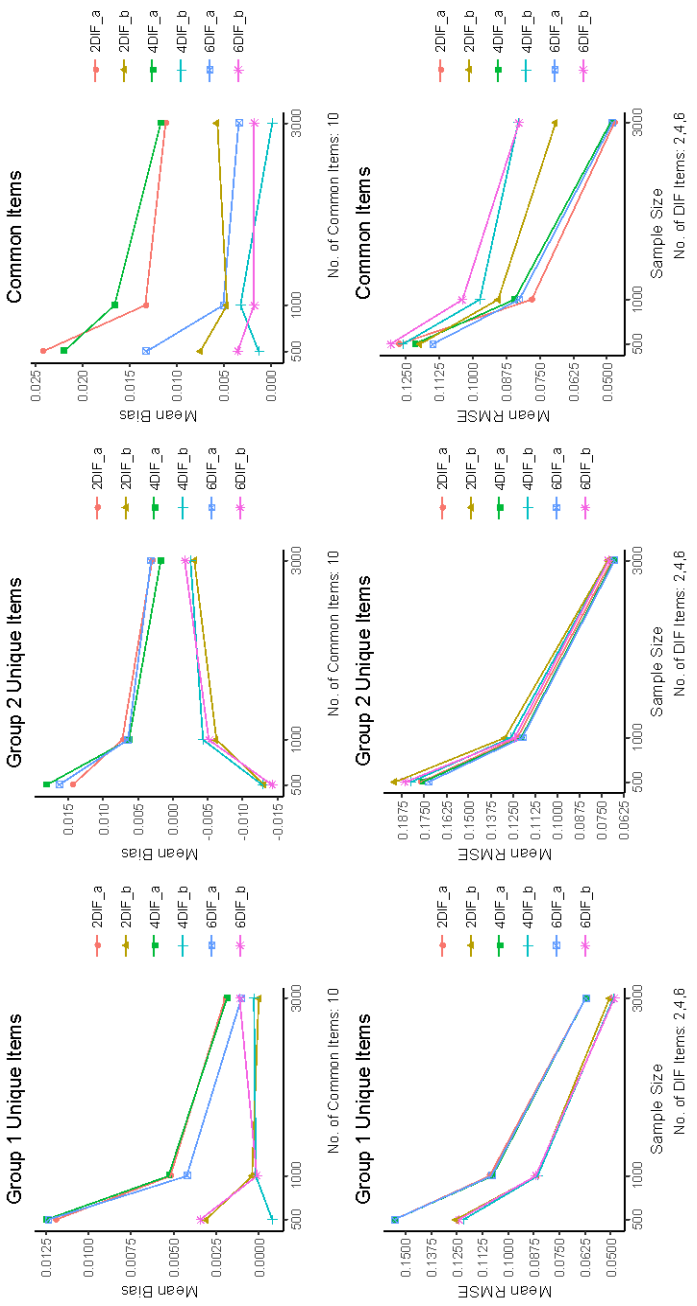

Figure 17 Small Uniform DIF Favoring Group 2 10 Common Items Item Parameter Recovery by Group

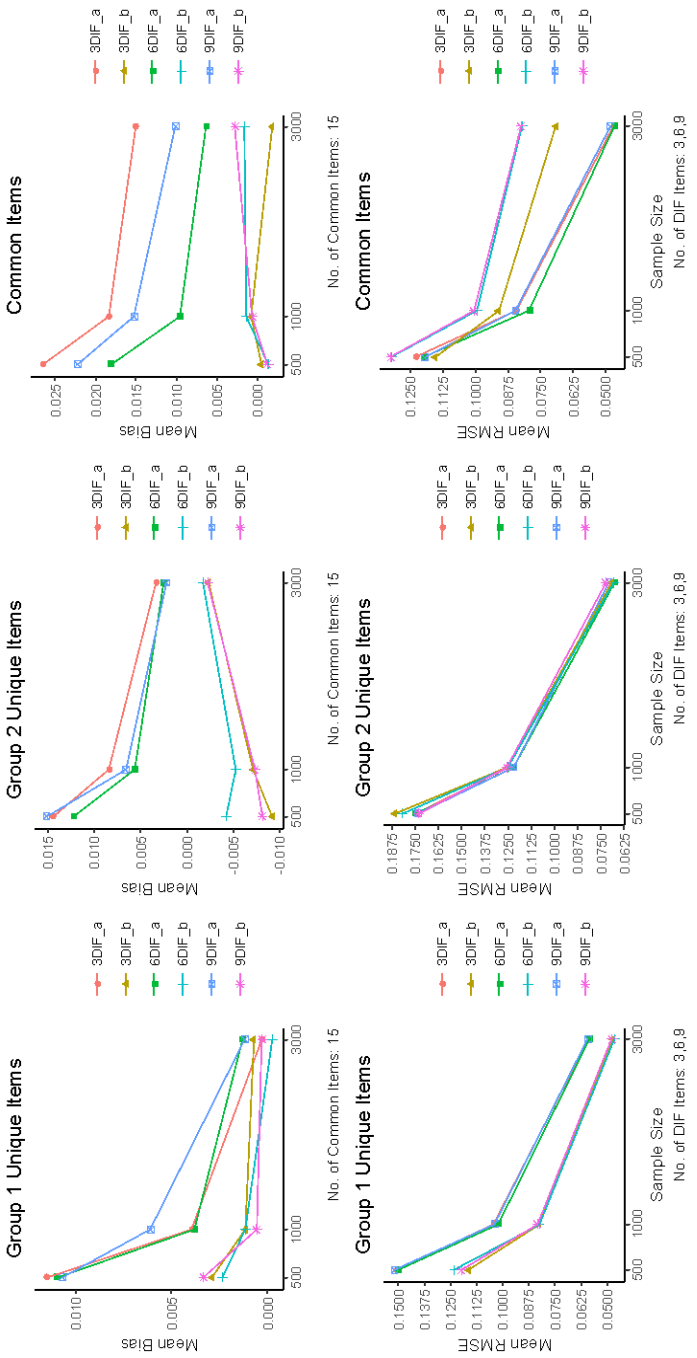

Figure 18 Small Uniform DIF Favoring Group 2 15 Common Items Item Parameter Recovery by Group

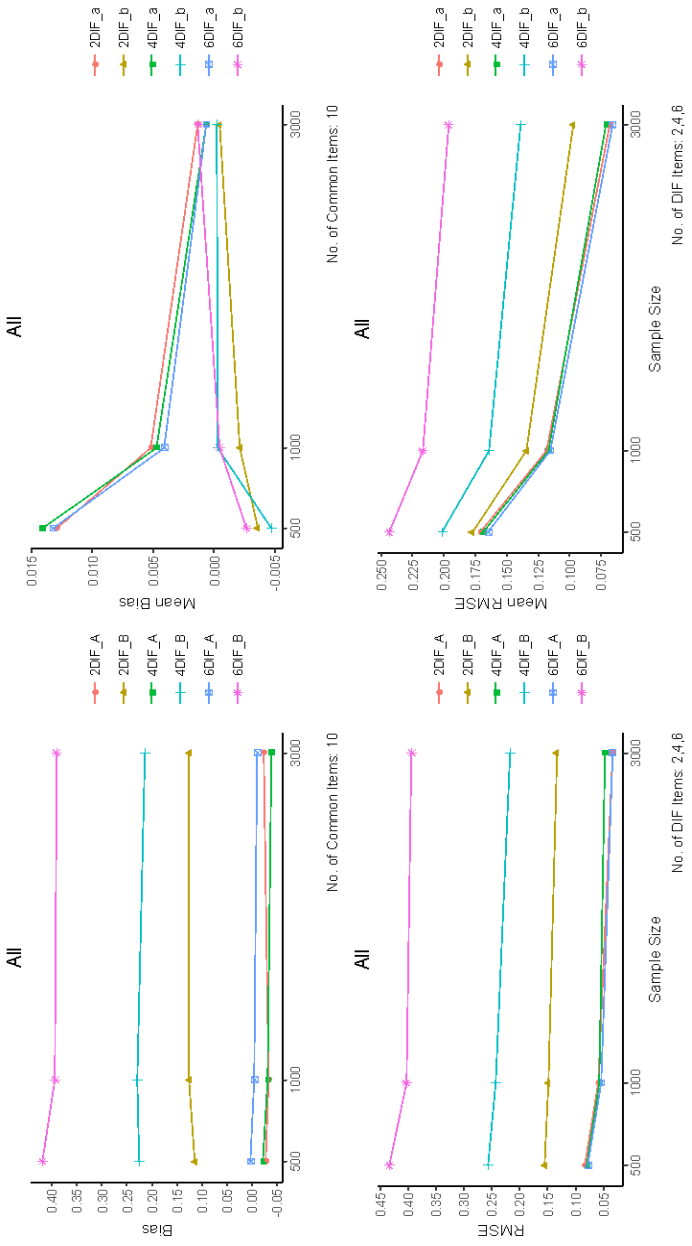

Figure 19 Medium Uniform DIF Favoring Group 2 10 Common Items Linking Constants and Item Parameter Recovery

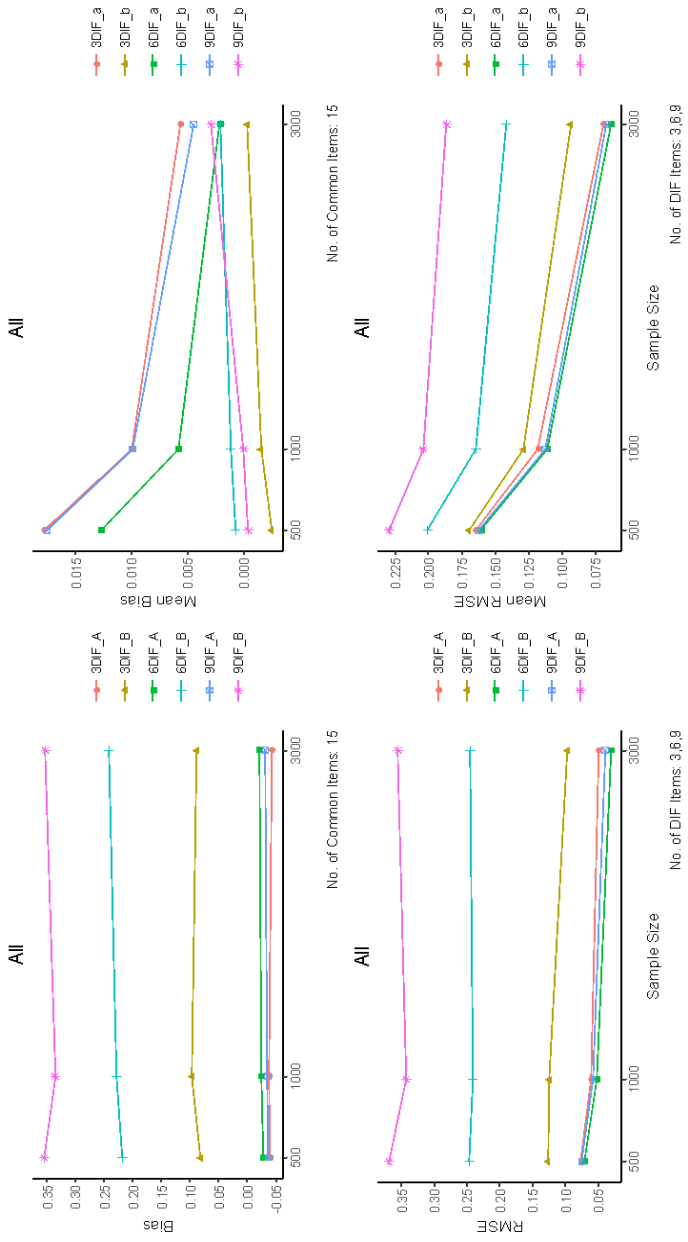

Figure 20 Medium Uniform DIF Favoring Group 2 15 Common Items Linking Constants and Item Parameter Recovery

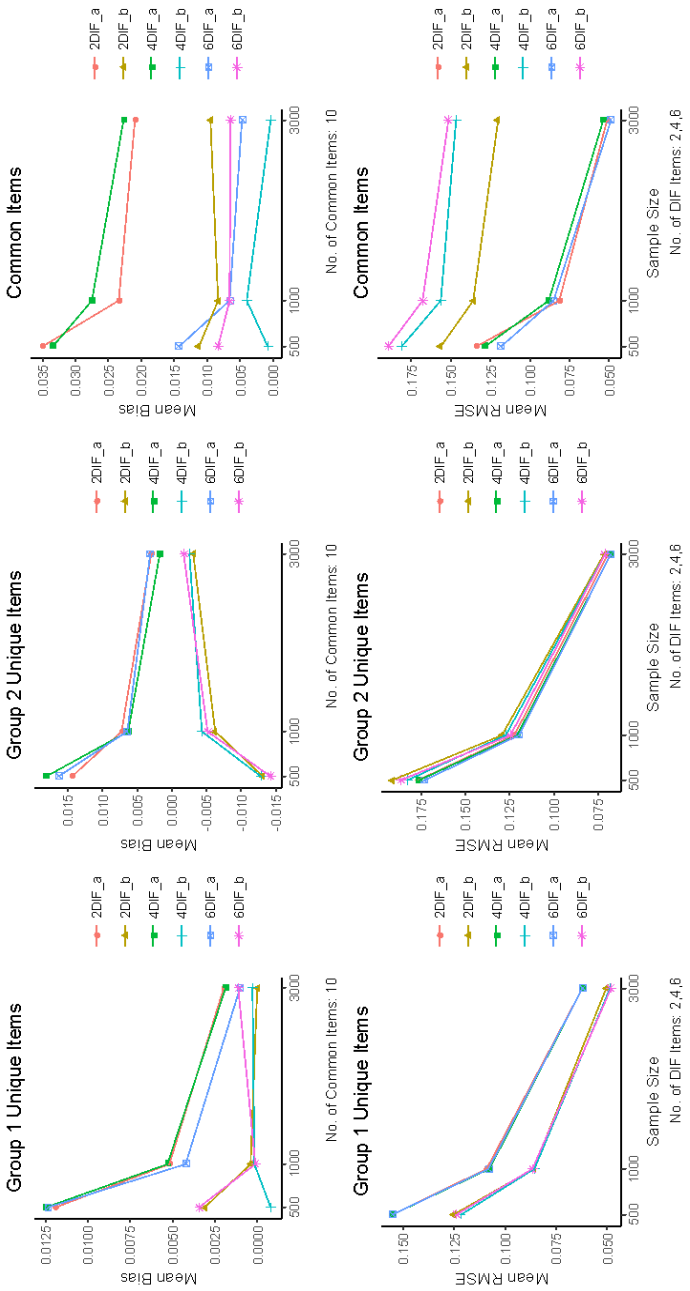

Figure 21 Medium Uniform DIF Favoring Group 2 10 Common Items Item Parameter Recovery by Group

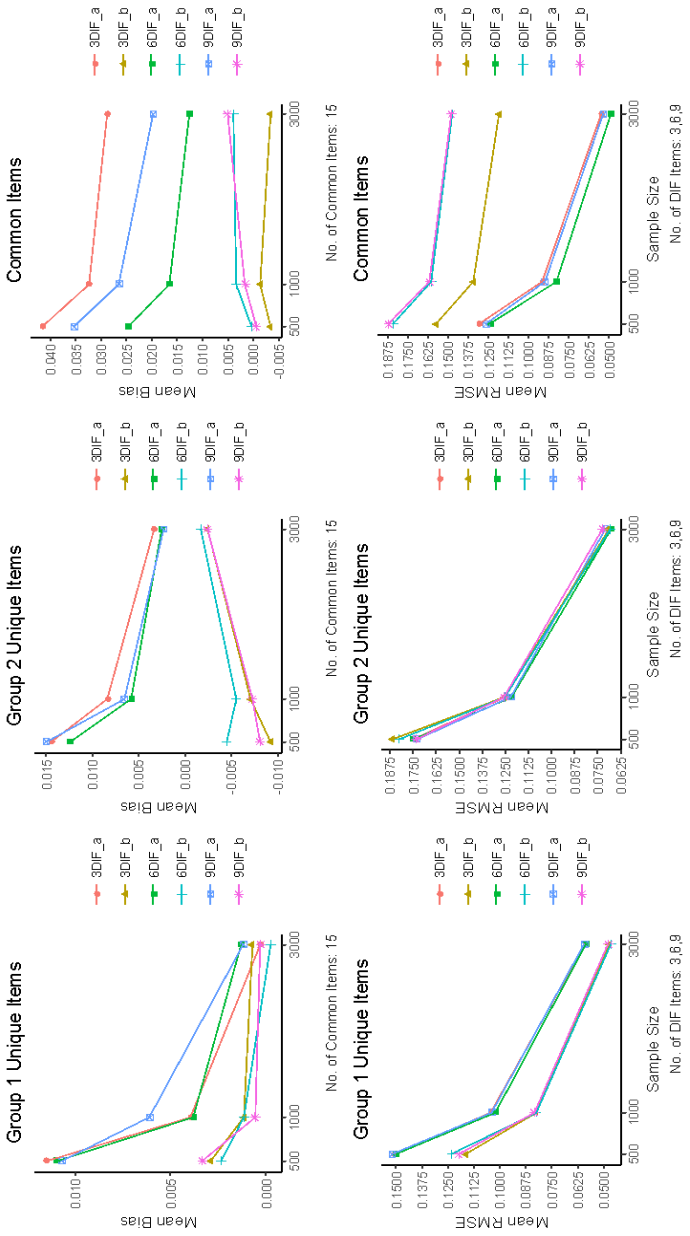

Figure 22 Medium Uniform DIF Favoring Group 2 15 Common Items Item Parameter Recovery by Group

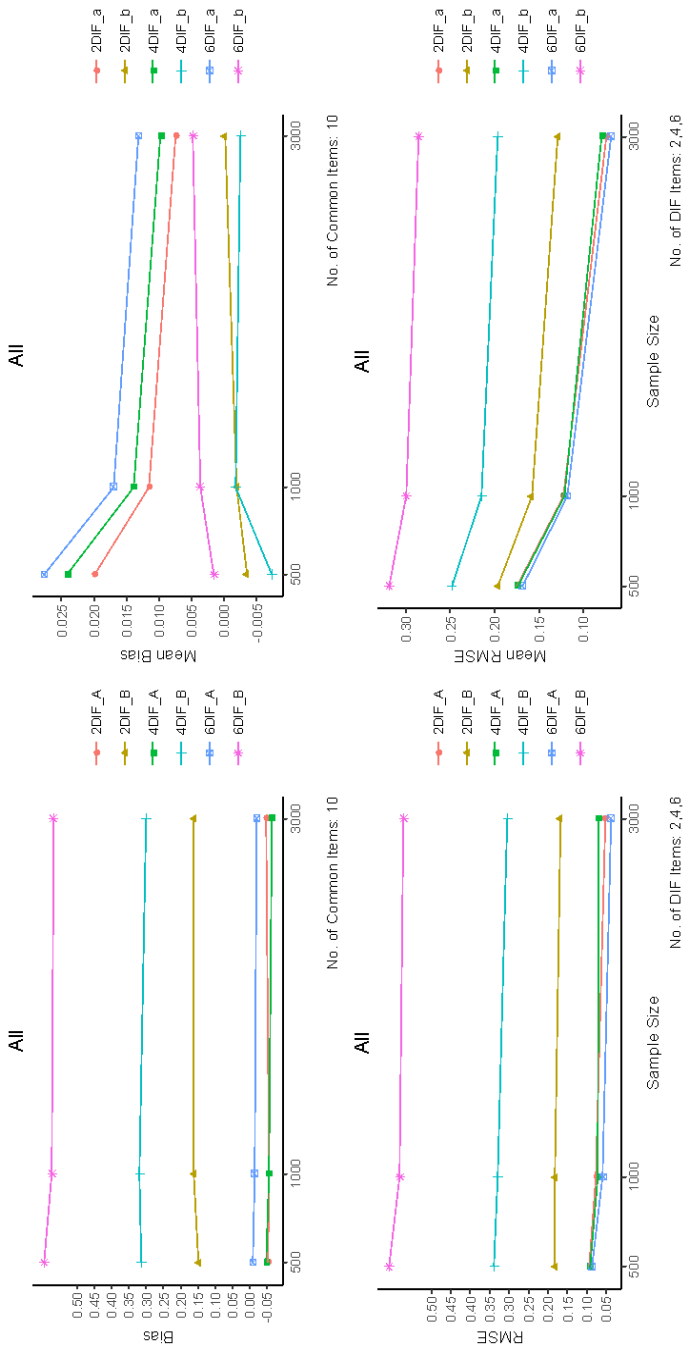

Figure 23 Large Uniform DIF Favoring Group 2 10 Common Items Linking Constants and Item Parameter Recovery

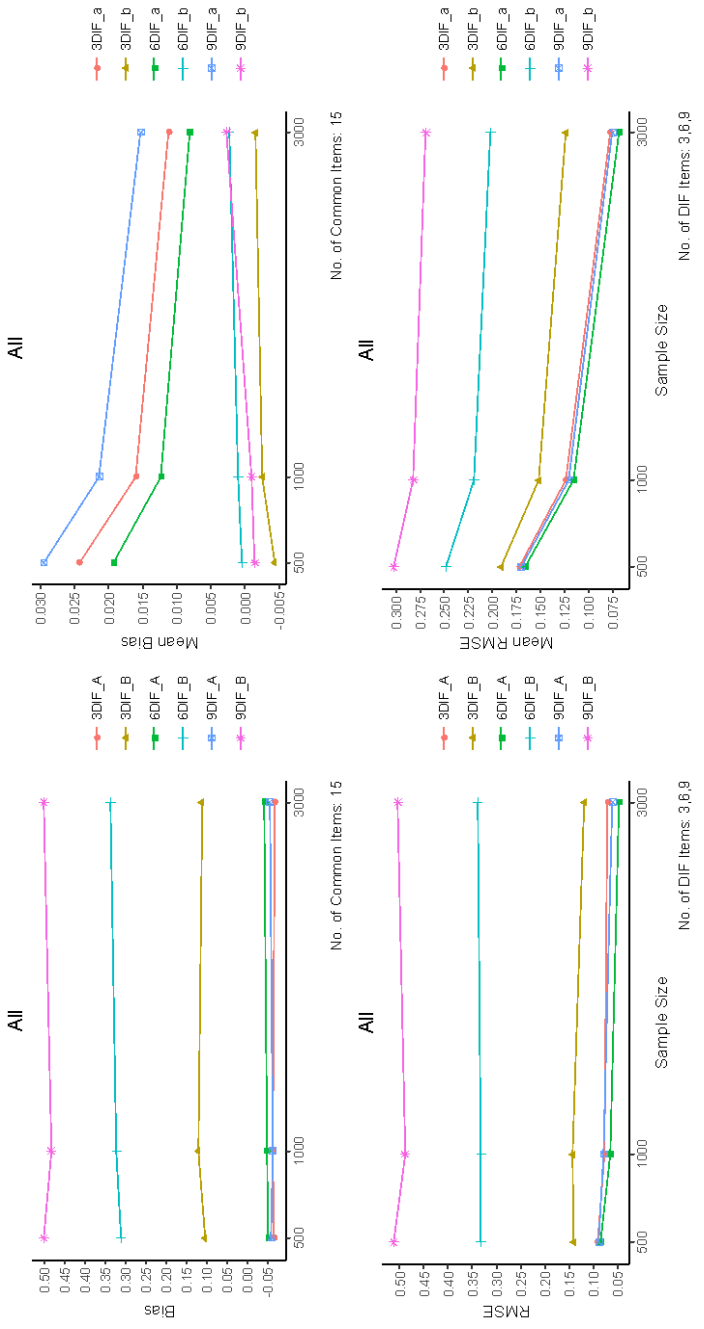

Figure 24 Large Uniform DIF Favoring Group 2 15 Common Items Linking Constants and Item Parameter Recovery

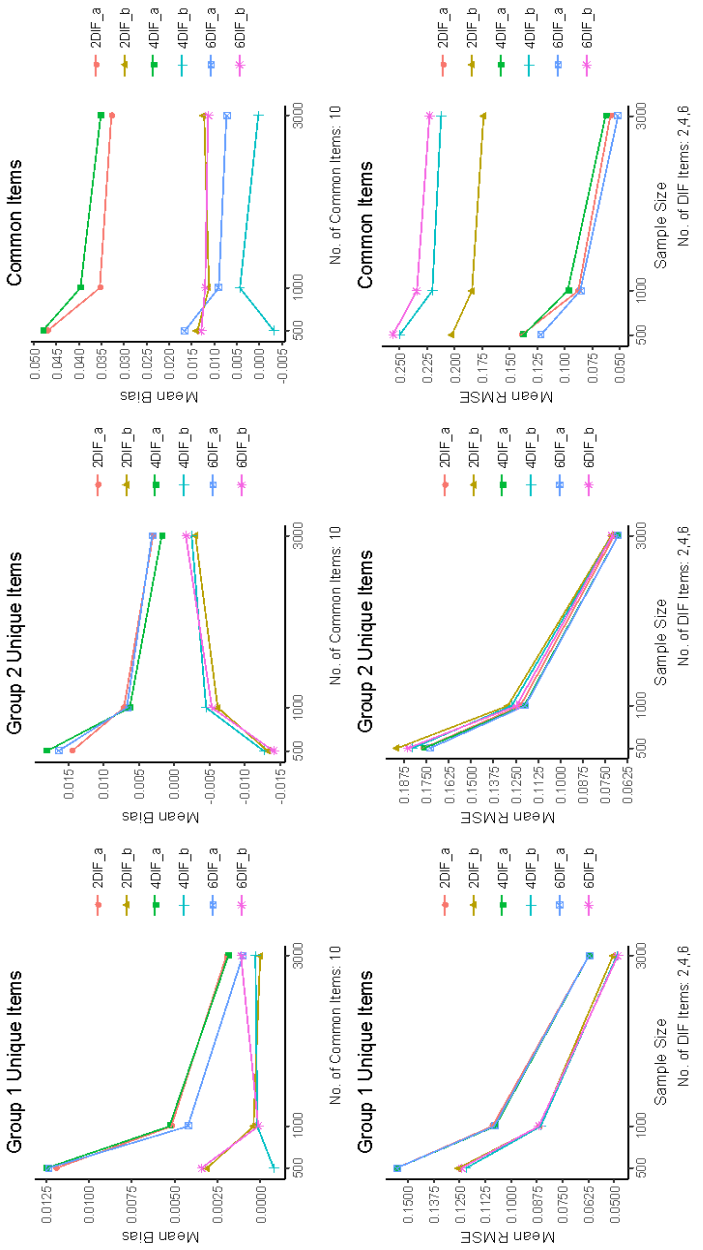

Figure 25 Large Uniform DIF Favoring Group 2 10 Common Items Item Parameter Recovery by Group

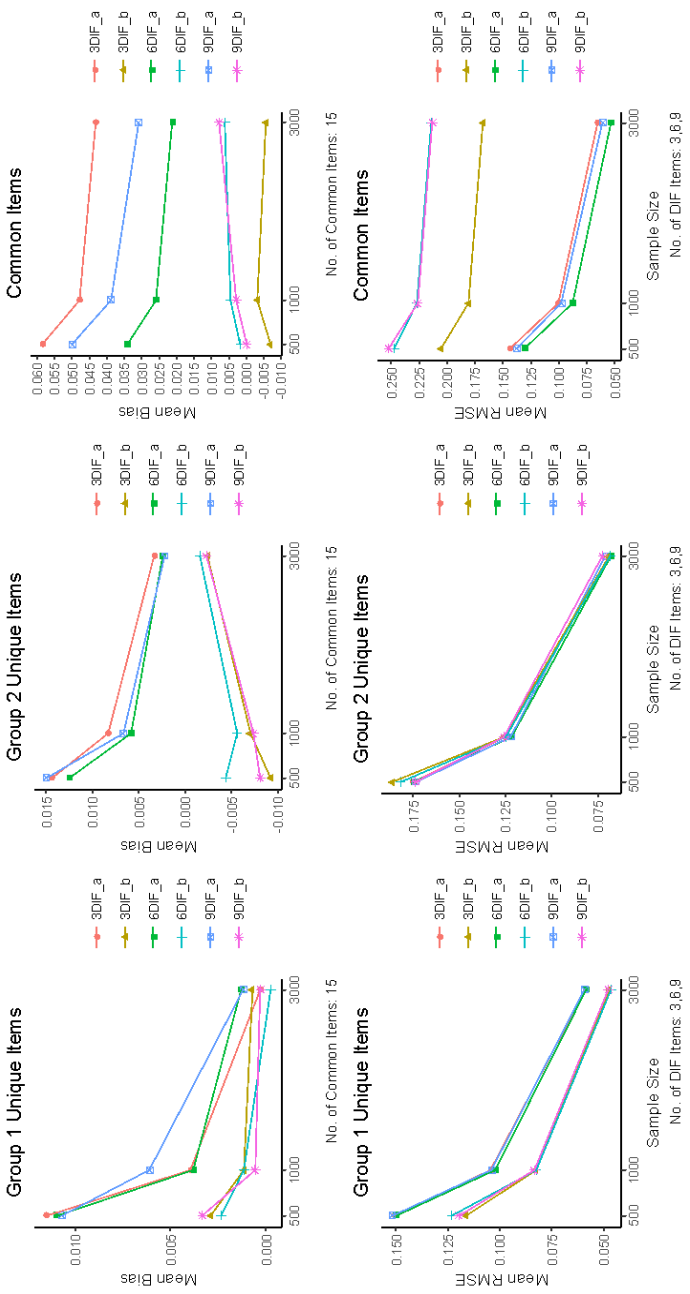

Figure 26 Large Uniform DIF Favoring Group 2 15 Common Items Item Parameter Recovery by Group

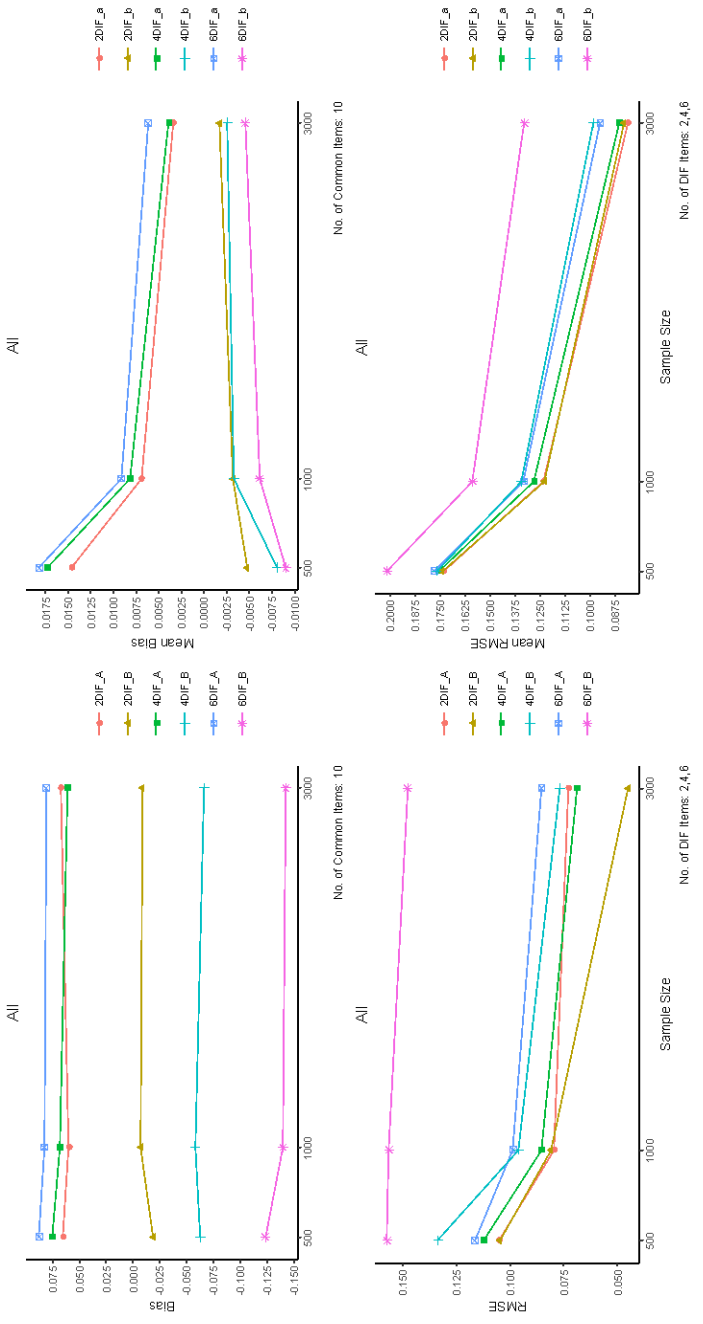

Figure 27 Small Uniform and Nonuniform DIF Against Group 2 10 Common Items Linking Constants and Item Parameter Recovery

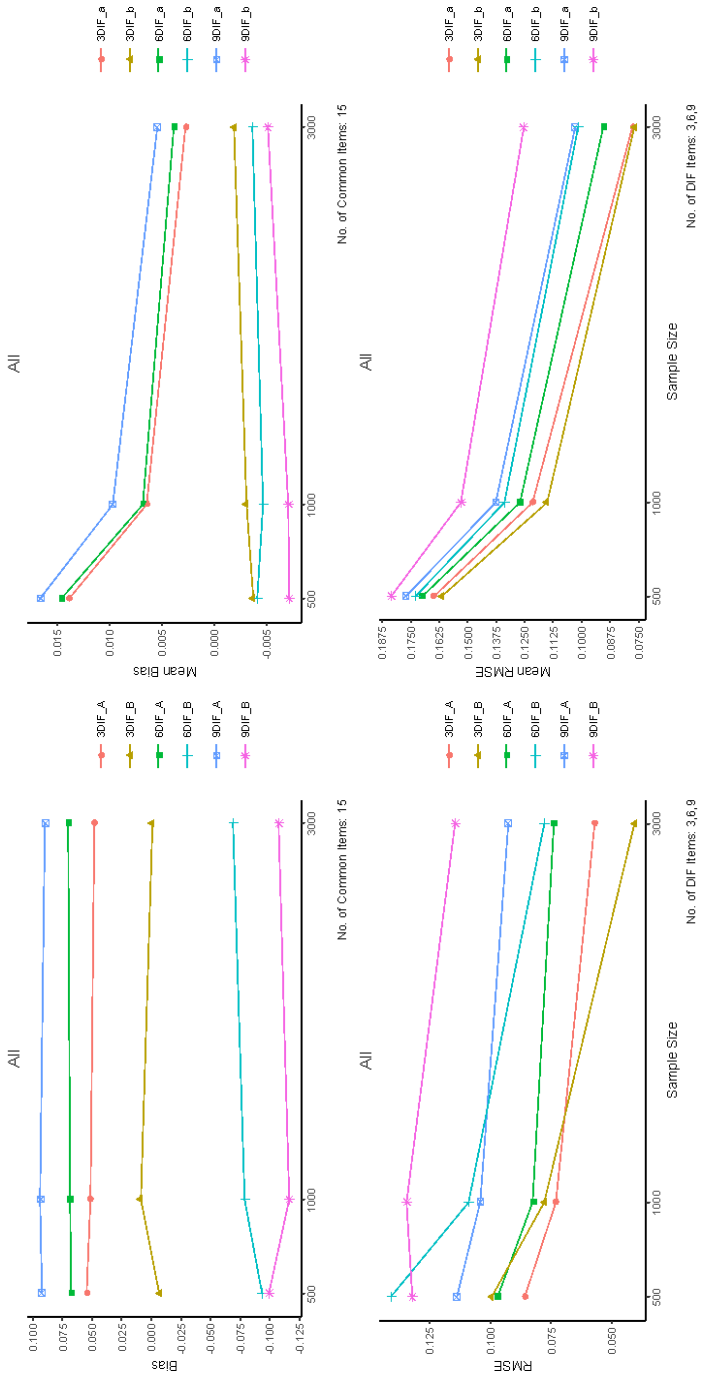

Figure 28 Small Uniform and Nonuniform DIF Against Group 2 15 Common Items Linking Constants and Item Parameter Recovery

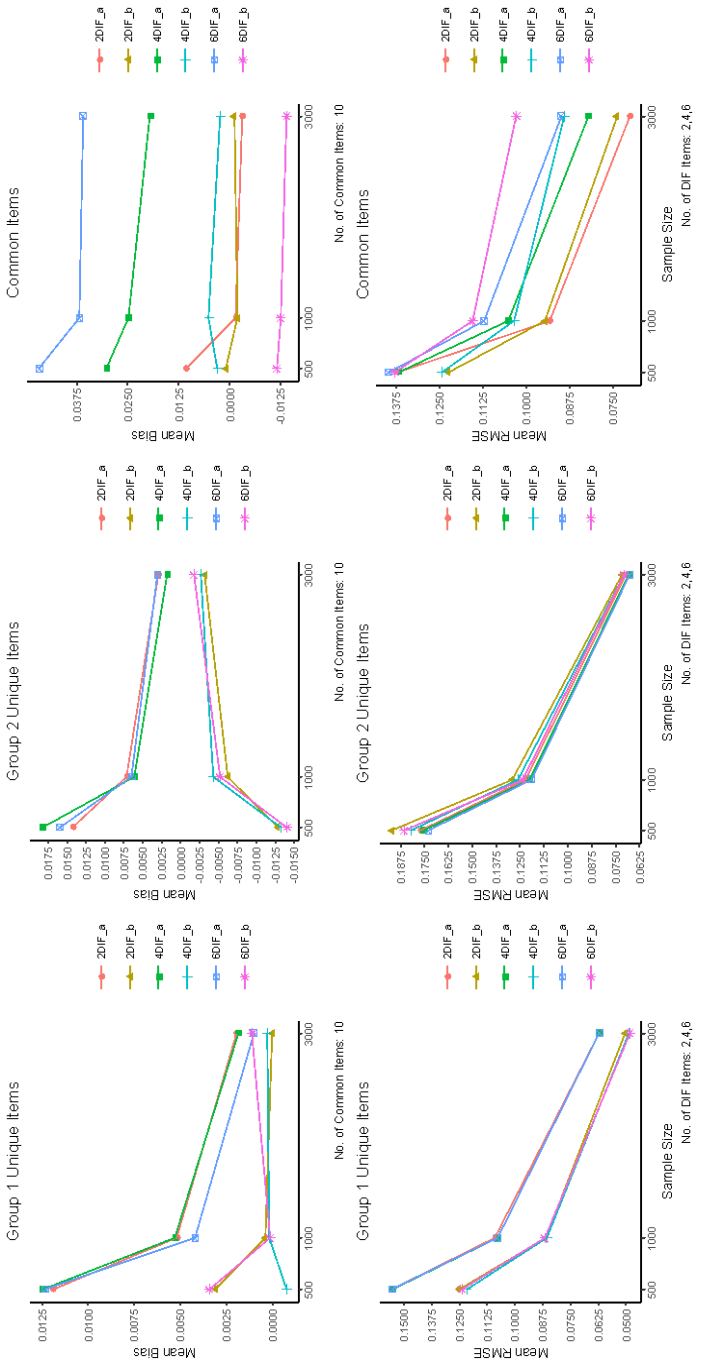

Figure 29 Small Uniform and Nonuniform DIF Against Group 2 10 Common Items Item Parameter Recovery by Group

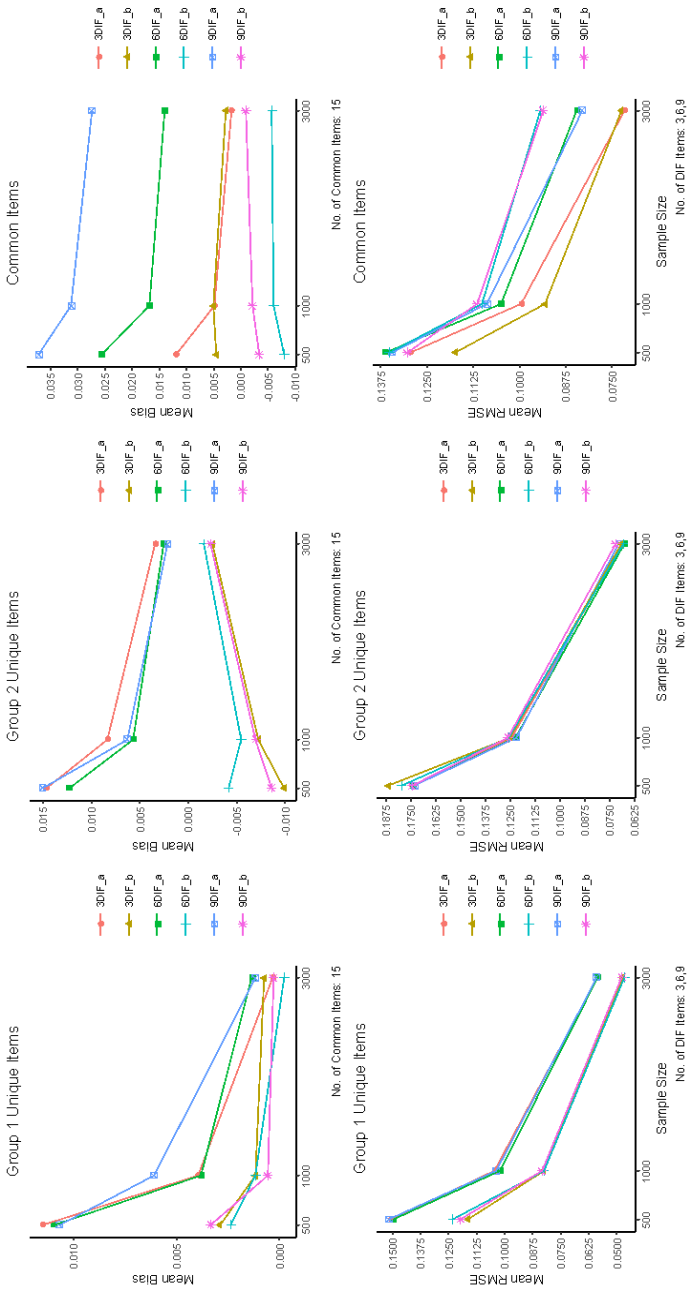

Figure 30 Small Uniform and Nonuniform DIF Against Group 2 15 Common Items Item Parameter Recovery by Group

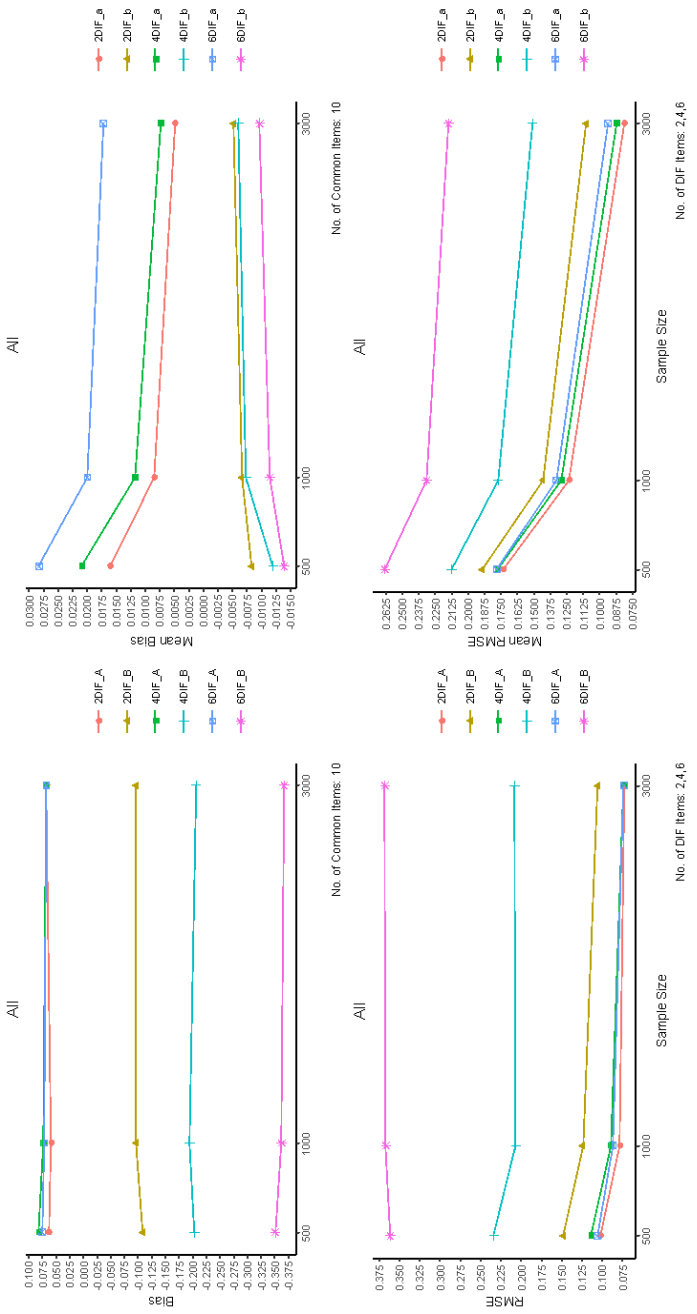

Figure 31 Medium Uniform and Small Nonuniform DIF Against Group 2 10 Common Items Linking Constants and Item Parameter Recovery

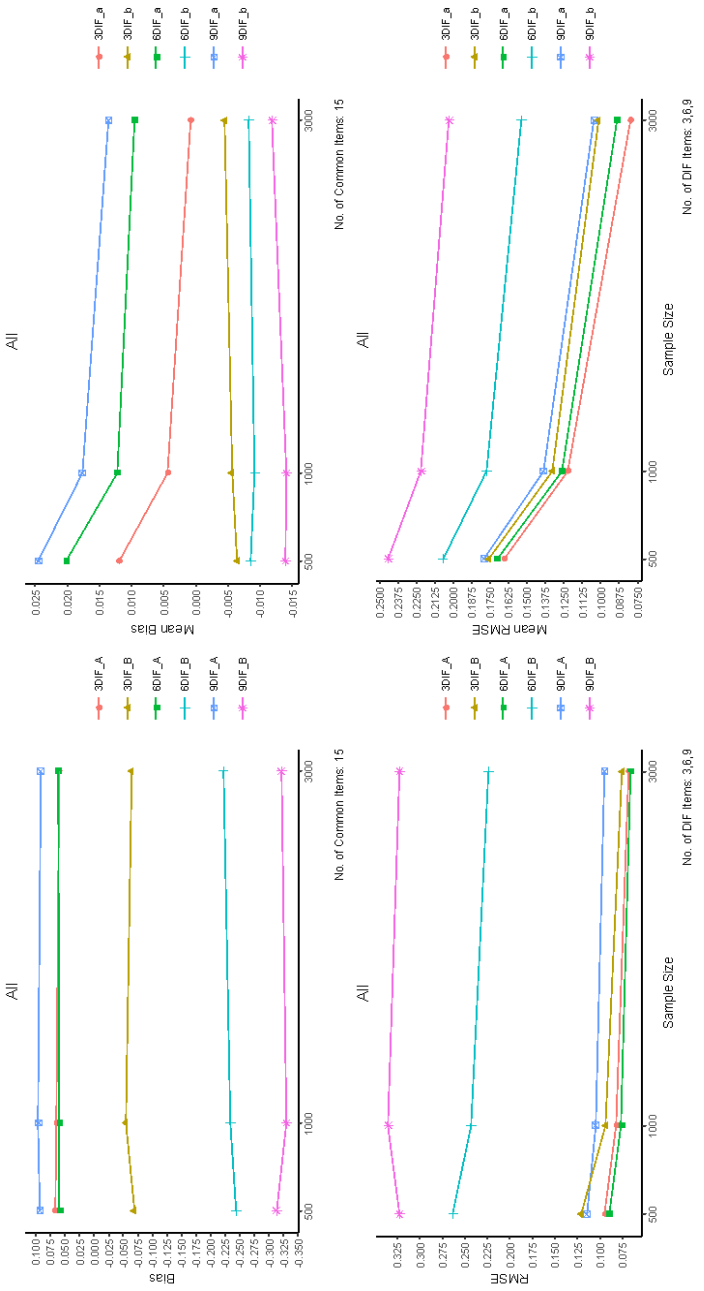

Figure 32 Medium Uniform and Small Nonuniform DIF Against Group 2 15 Common Items Linking Constants and Item Parameter Recovery

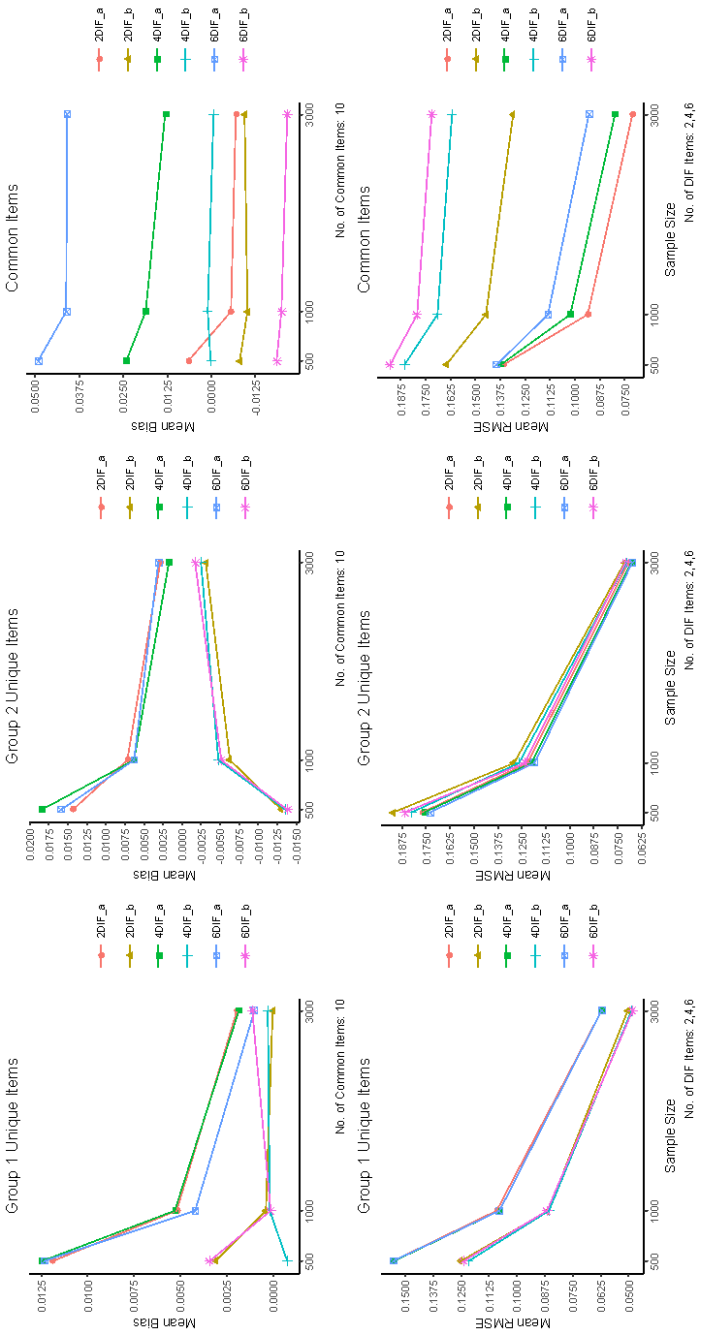

Figure 33 Medium Uniform and Small Nonuniform DIF Against Group 2 10 Common Items Item Parameter Recovery by Group

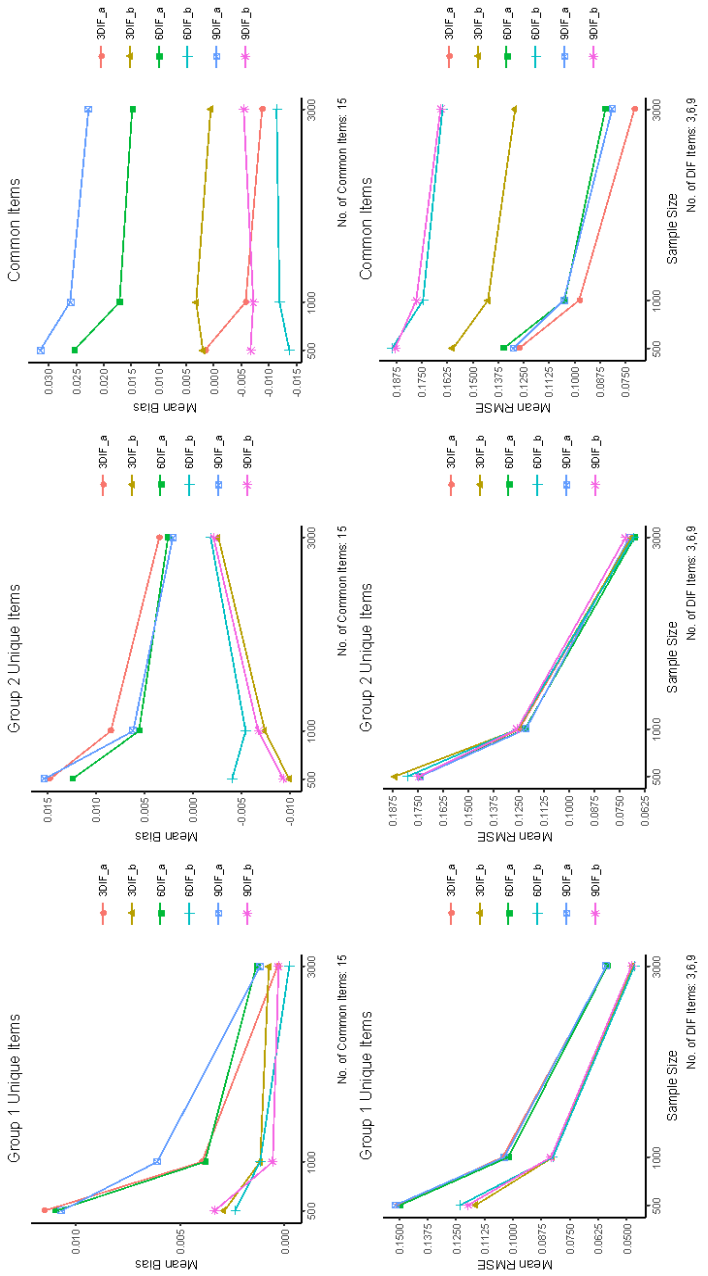

Figure 34 Medium Uniform and Small Nonuniform DIF Against Group 2 15 Common Items Item Parameter Recovery by Group

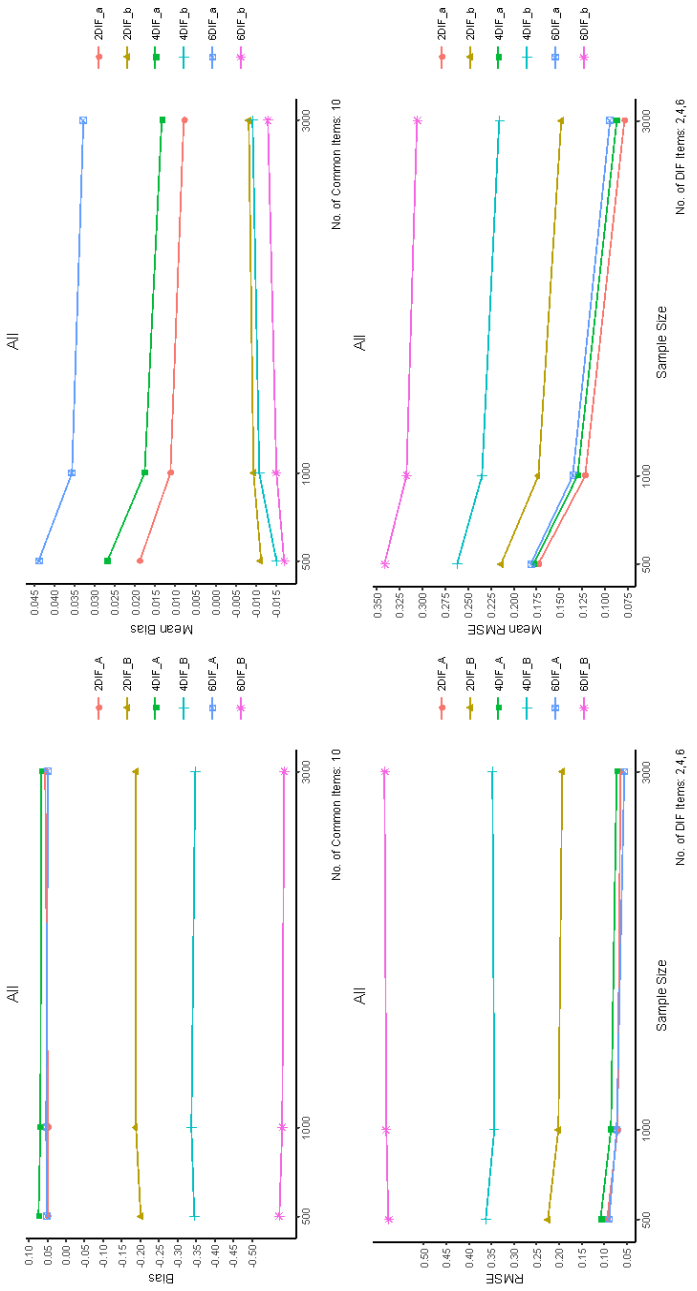

Figure 35 Large Uniform and Small Nonuniform DIF Against Group 2 10 Common Items Linking Constants and Item Parameter Recovery

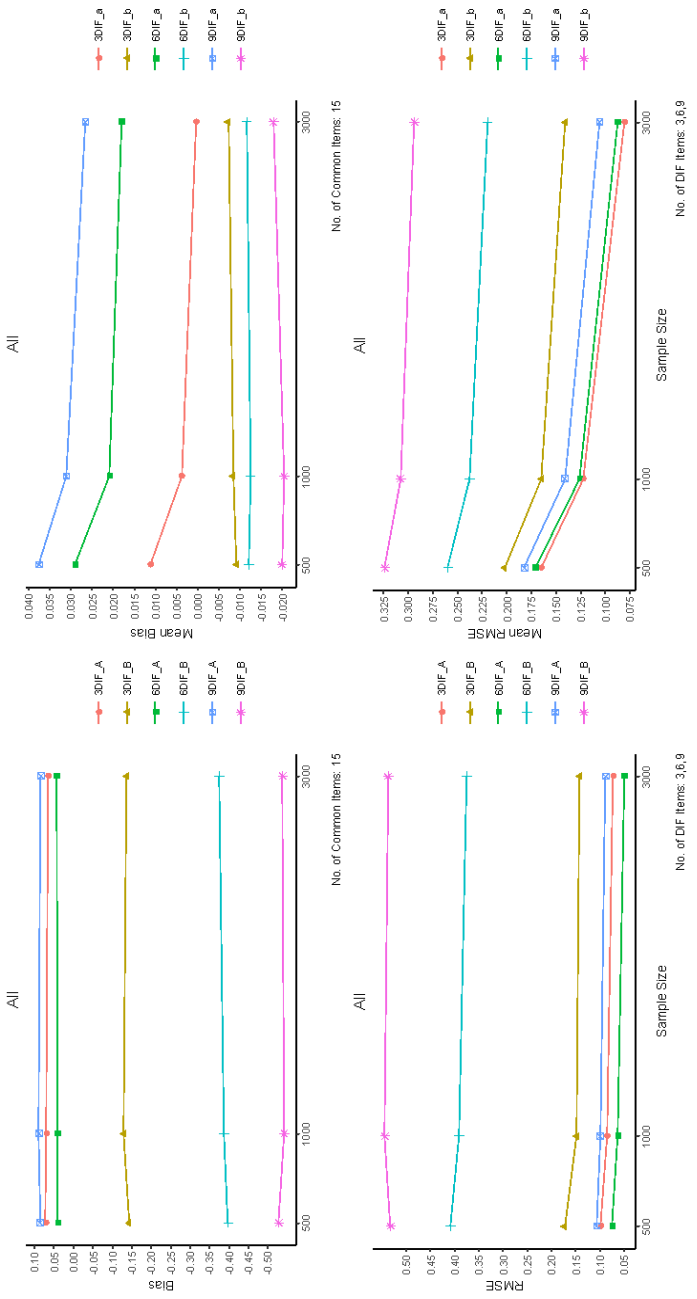

Figure 36 Large Uniform and Small Nonuniform DIF Against Group 2 15 Common Items Linking Constants and Item Parameter Recovery

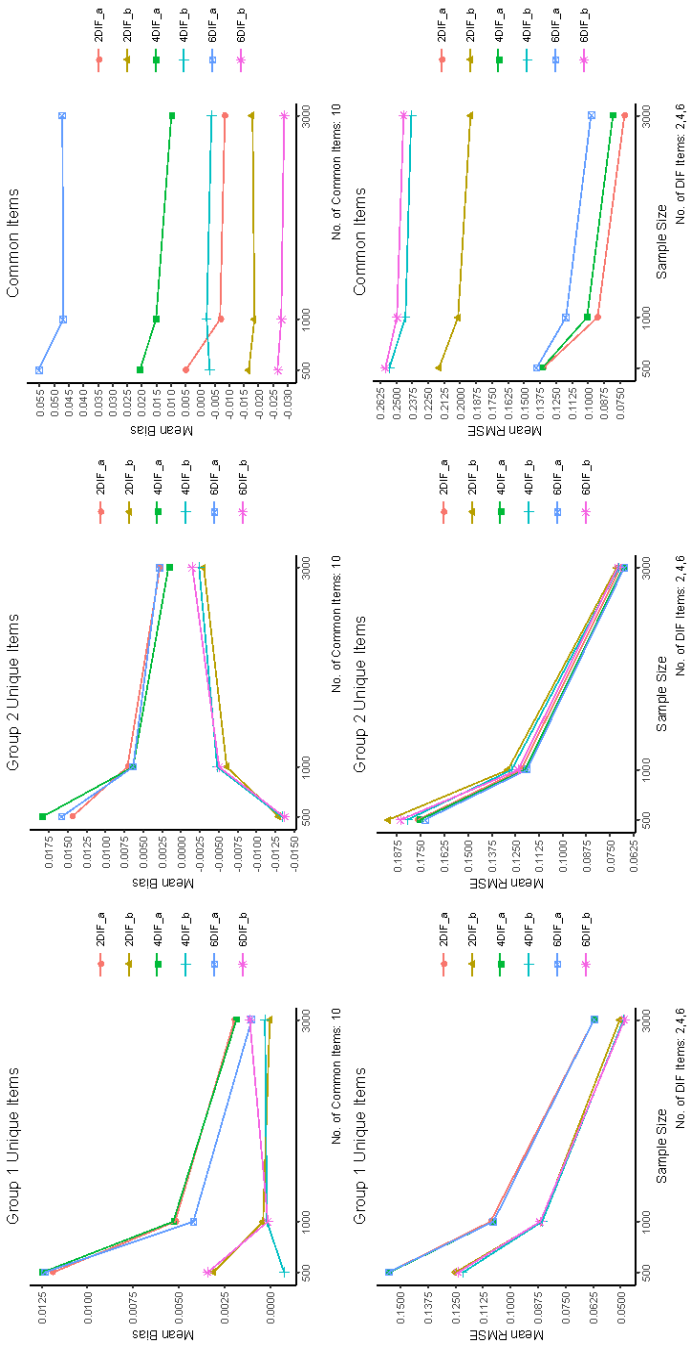

Figure 37 Large Uniform and Small Nonuniform DIF Against Group 2 10 Common Items Item Parameter Recovery by Group

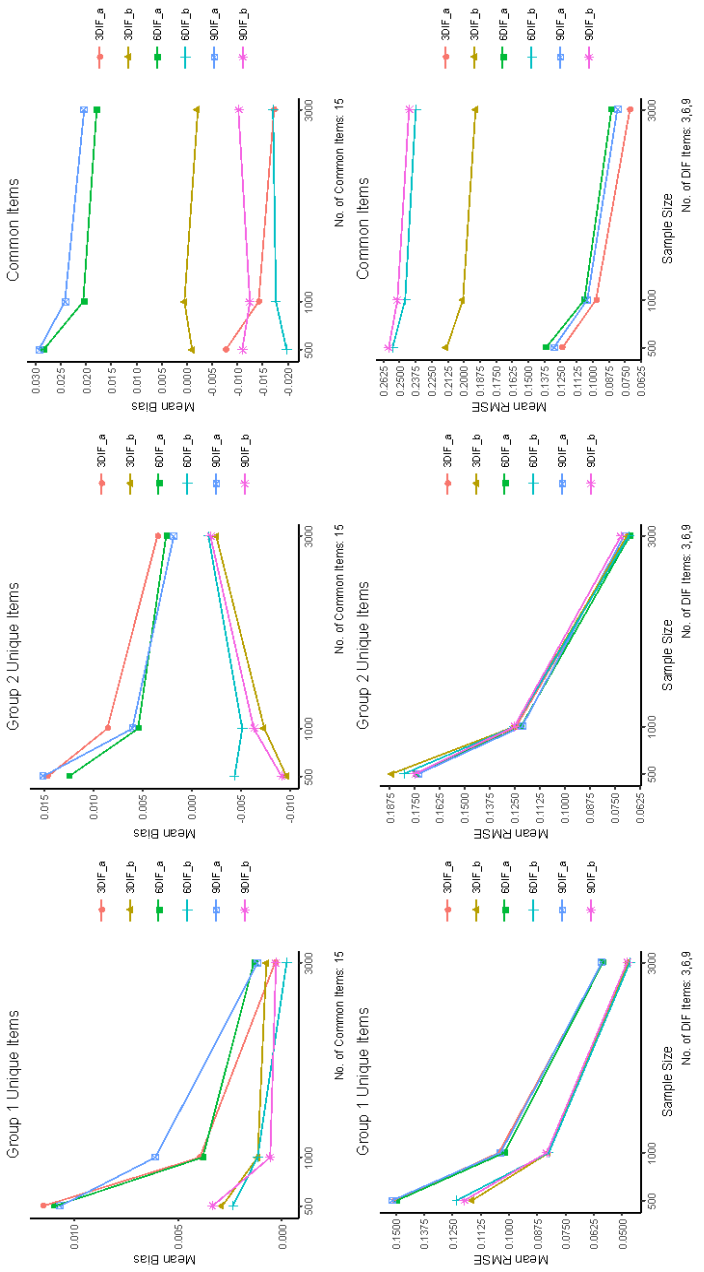

Figure 38 Large Uniform and Small Nonuniform DIF Against Group 2 15 Common Items Item Parameter Recovery by Group

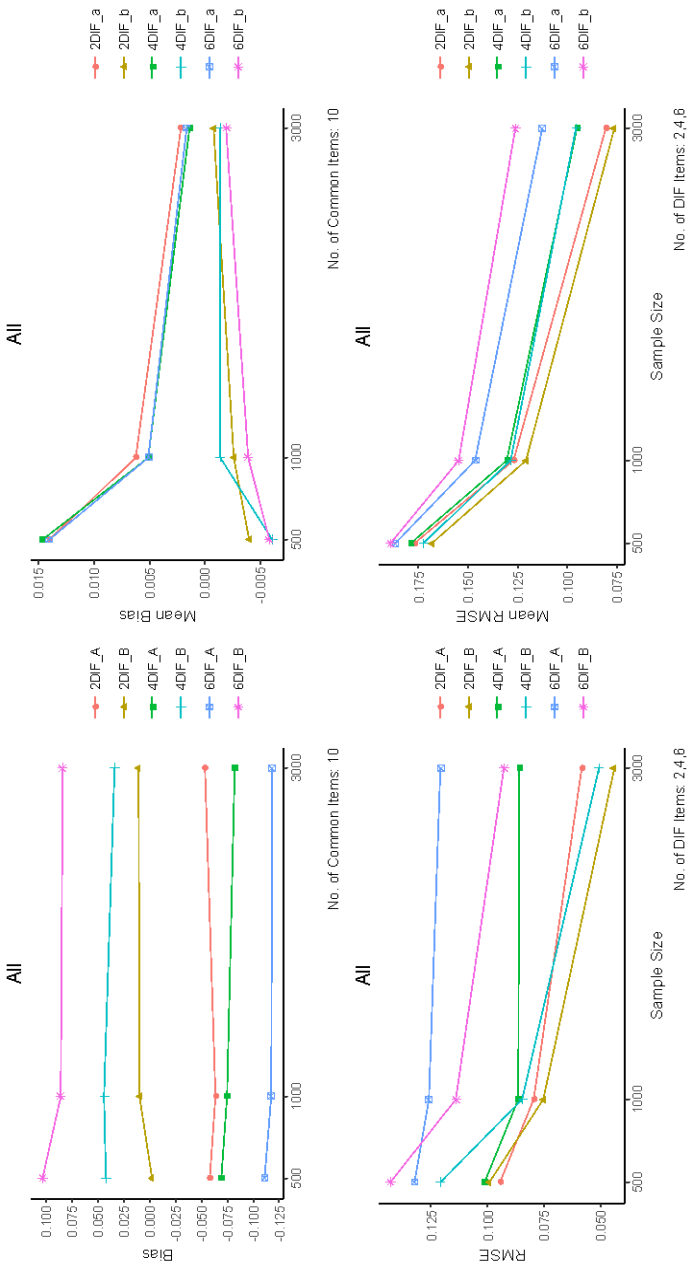

Figure 39 Small Uniform and Nonuniform DIF Favoring Group 2 10 Common Items Linking Constants and Item Parameter Recovery

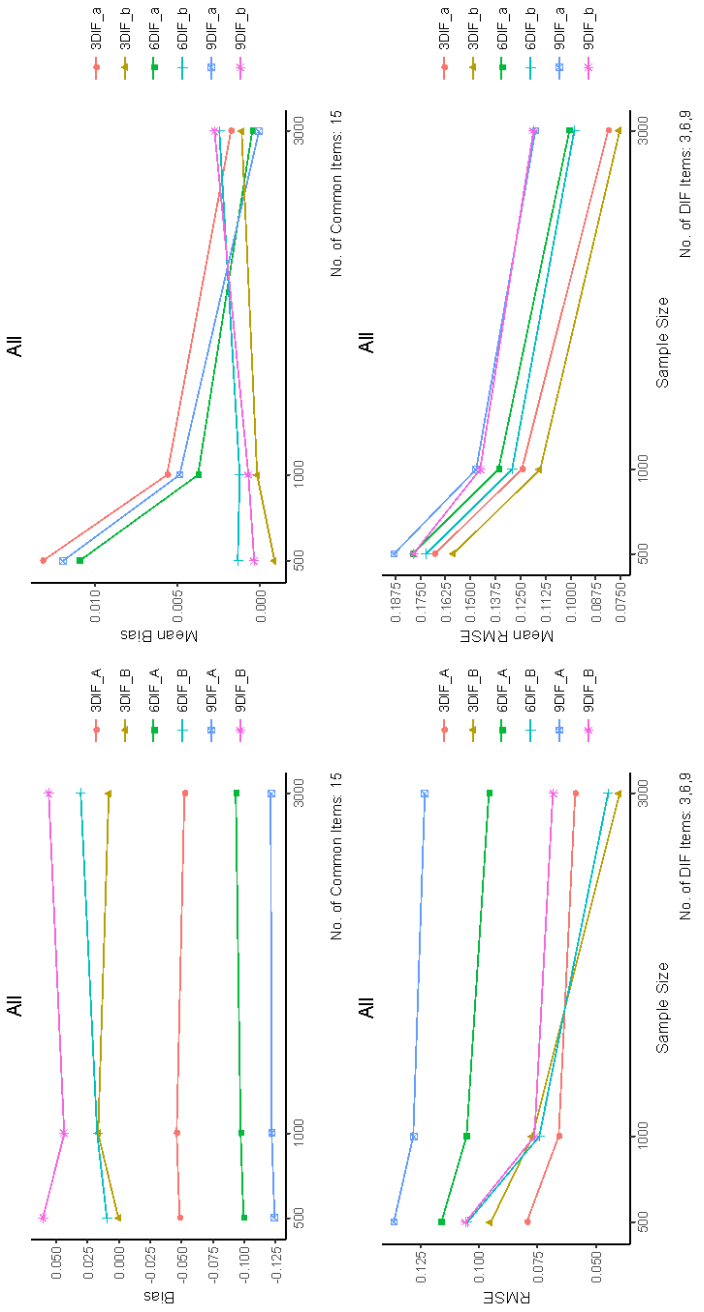

Figure 40 Small Uniform and Nonuniform DIF Favoring Group 2 15 Common Items Linking Constants and Item Parameter Recovery

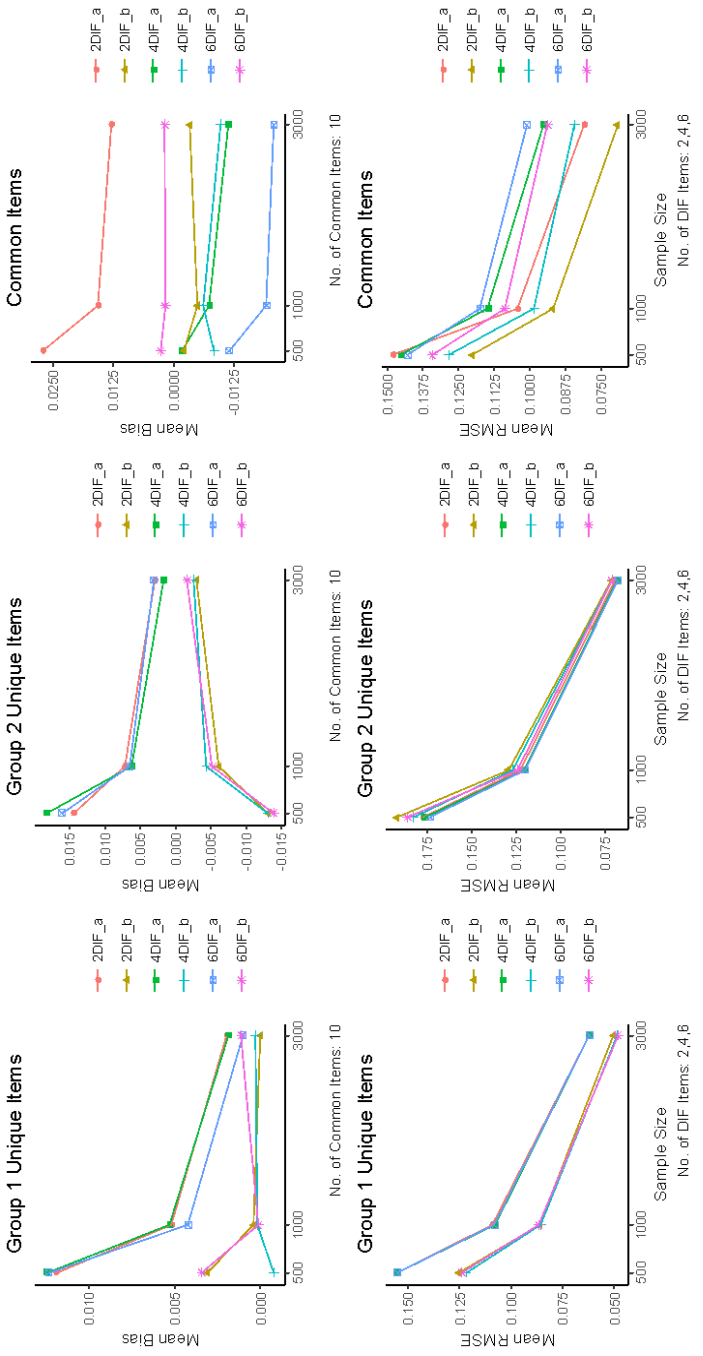

Figure 41 Small Uniform and Nonuniform DIF Favoring Group 2 10 Common Items Item Parameter Recovery by Group

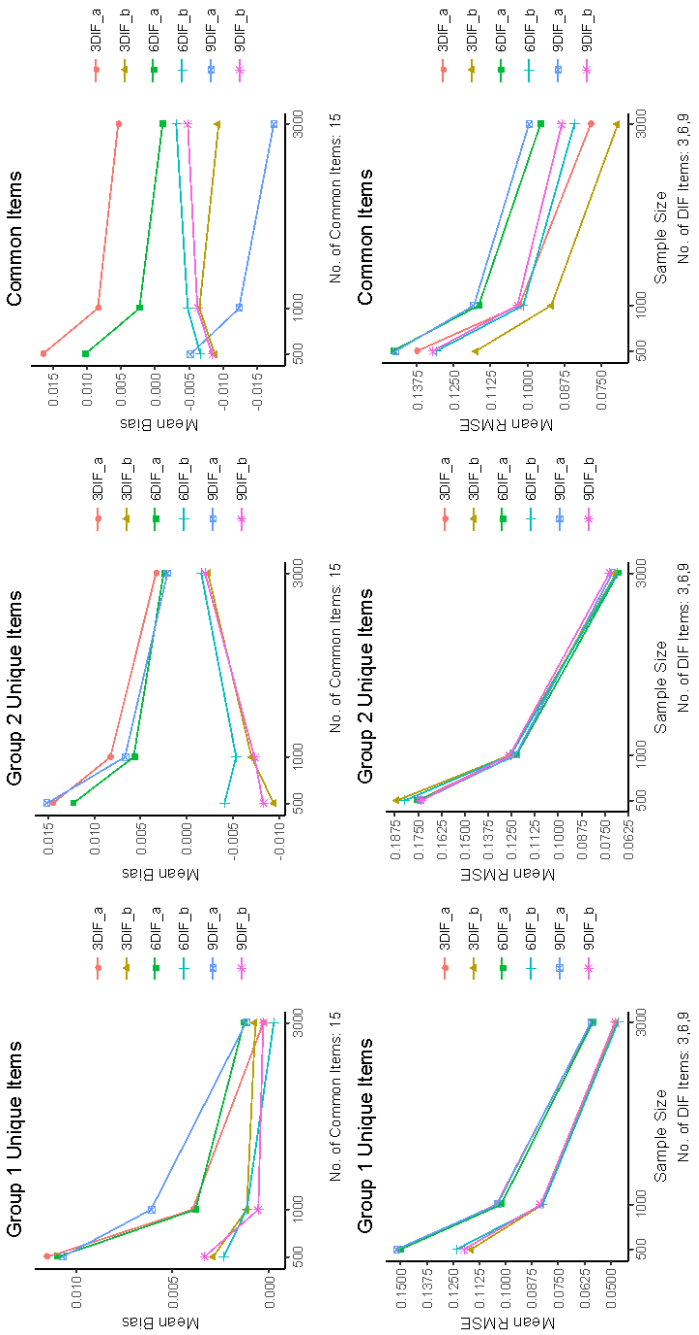

Figure 42 Small Uniform and Nonuniform DIF Favoring Group 2 15 Common Items Item Parameter Recovery by Group

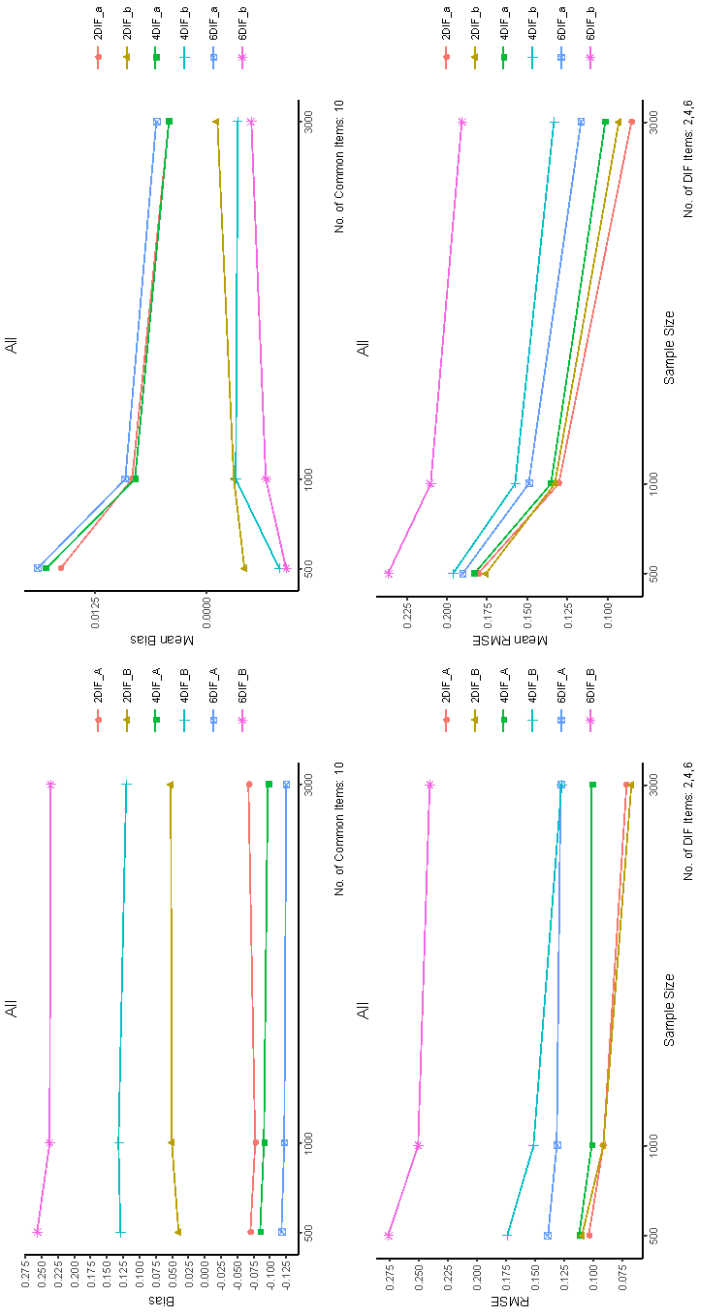

Figure 43 Medium Uniform and Small Nonuniform DIF Favoring Group 2 10 Common Items Linking Constants and Item Parameter Recovery

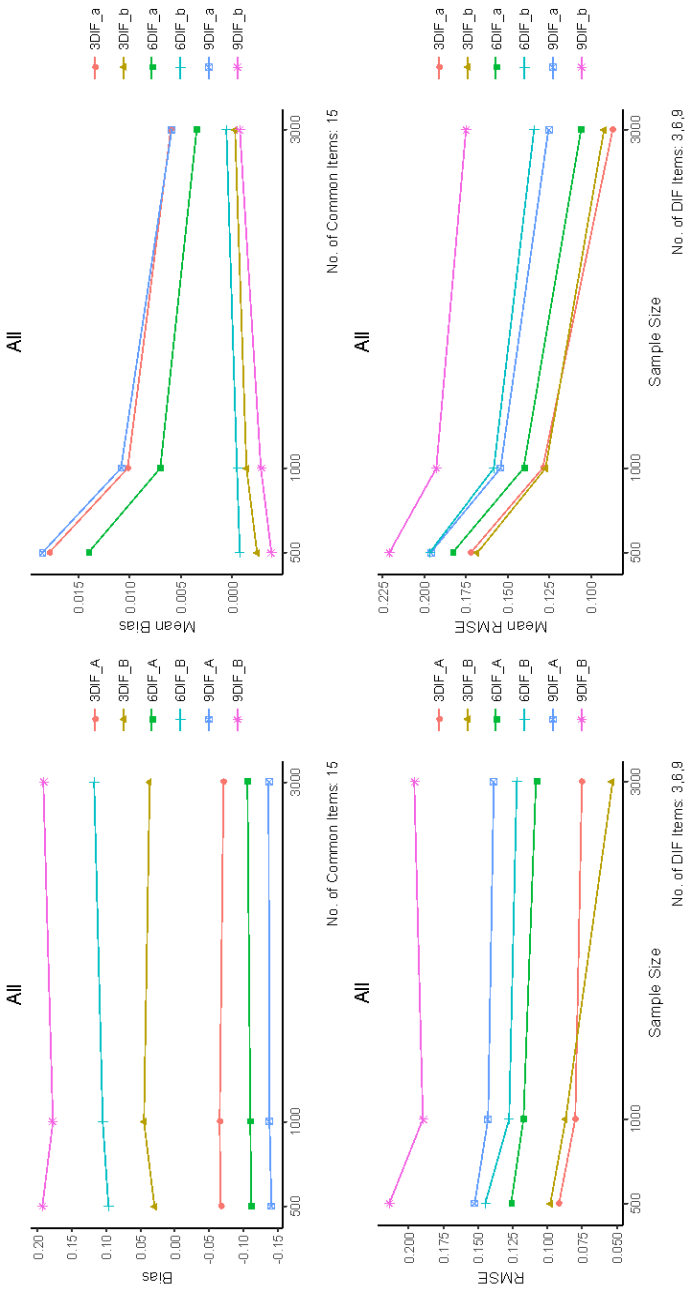

Figure 44 Medium Uniform and Small Nonuniform DIF Favoring Group 2 15 Common Items Linking Constants and Item Parameter Recovery

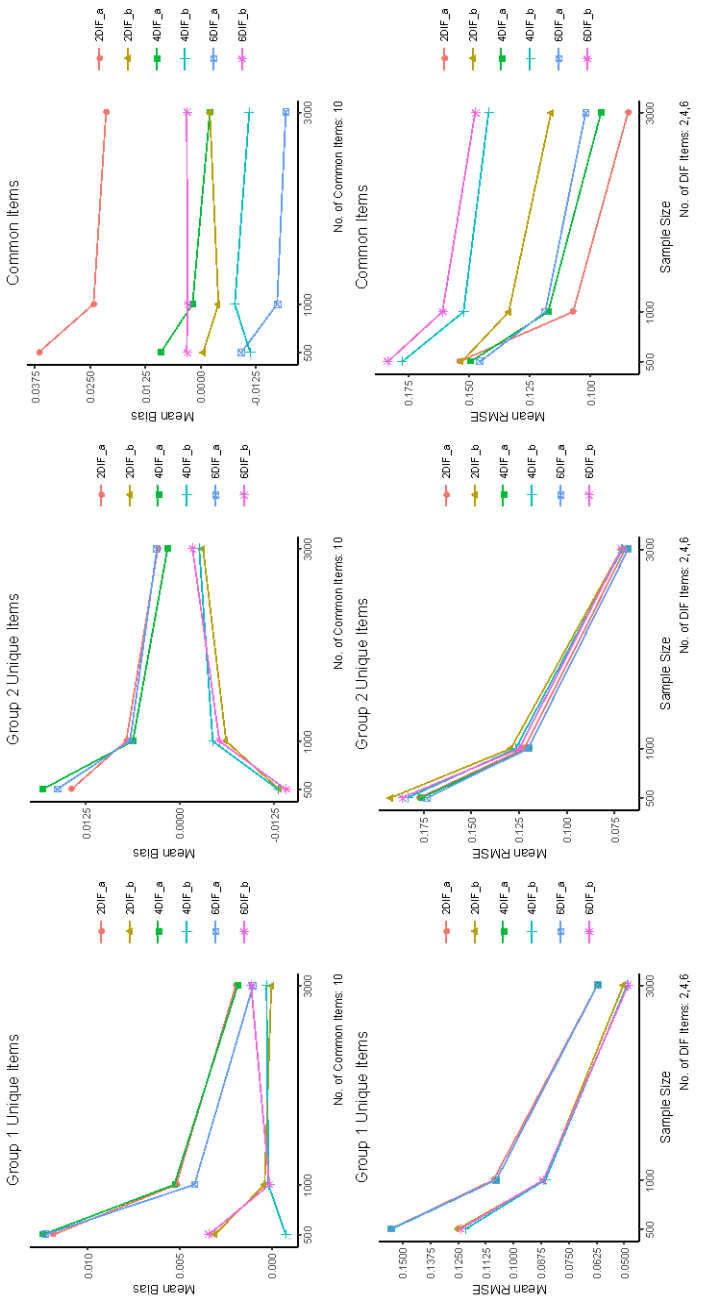

Figure 45 Medium Uniform and Small Nonuniform DIF Favoring Group 2 10 Common Items Item Parameter Recovery by Group

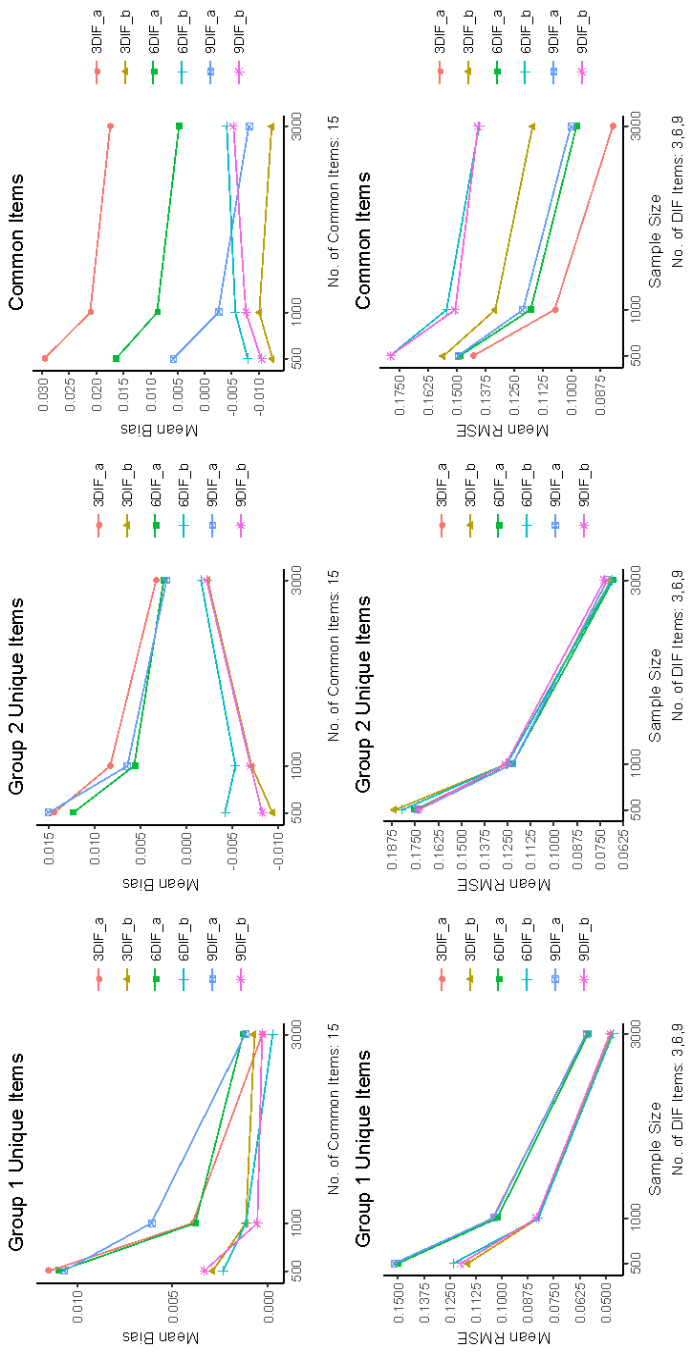

Figure 46 Medium Uniform and Small Nonuniform DIF Favoring Group 2 15 Common Items Item Parameter Recovery by Group

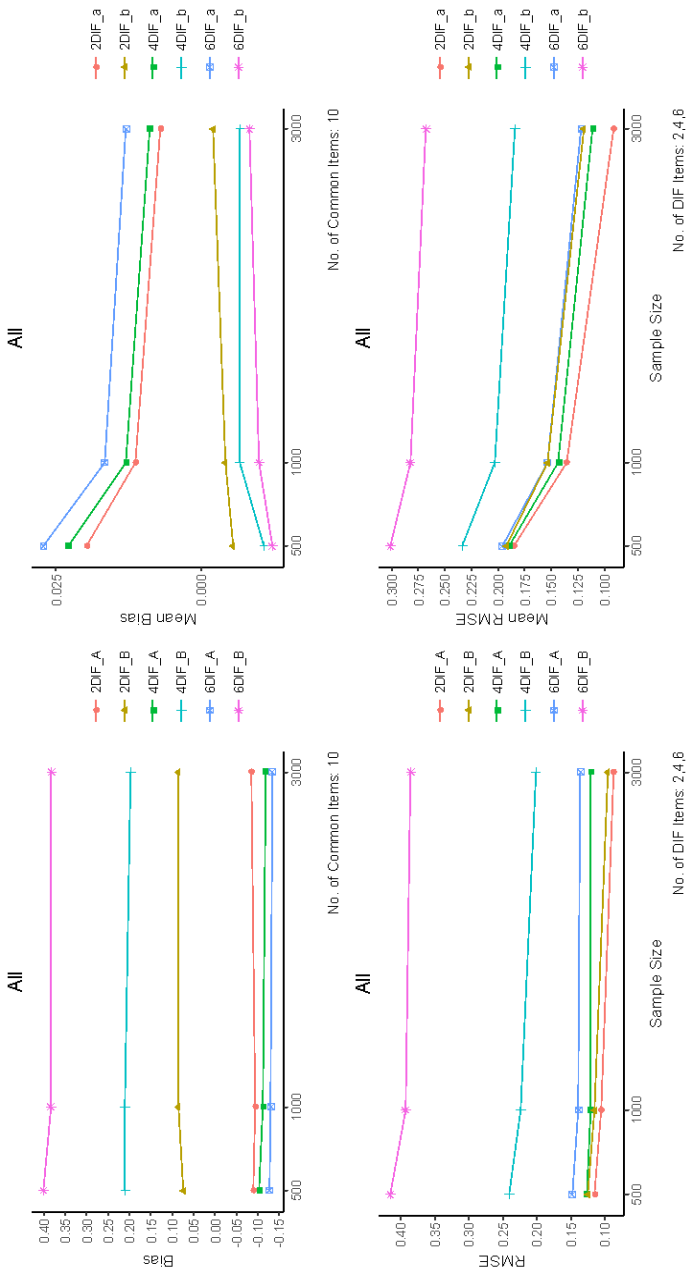

Figure 47 Large Uniform and Small Nonuniform DIF Favoring Group 2 10 Common Items Linking Constants and Item Parameter Recovery

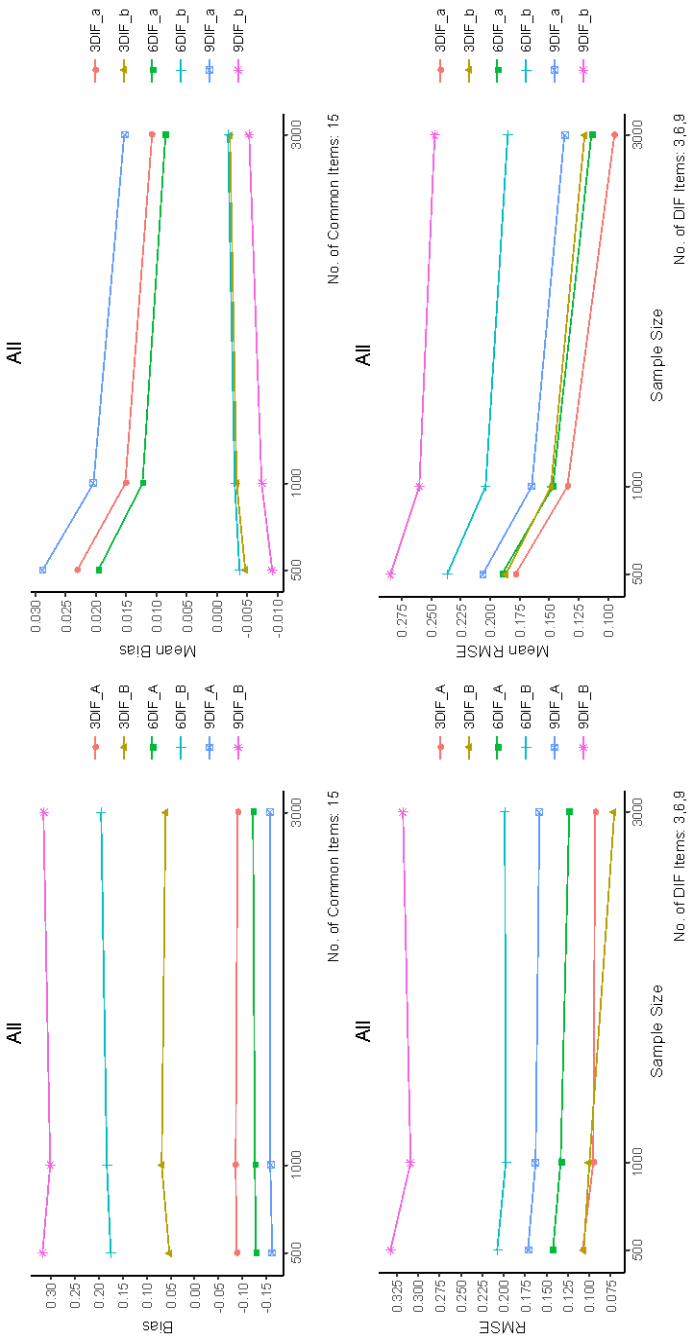

Figure 48 Large Uniform and Small Nonuniform DIF Favoring Group 2 15 Common Items Linking Constants and Item Parameter Recovery

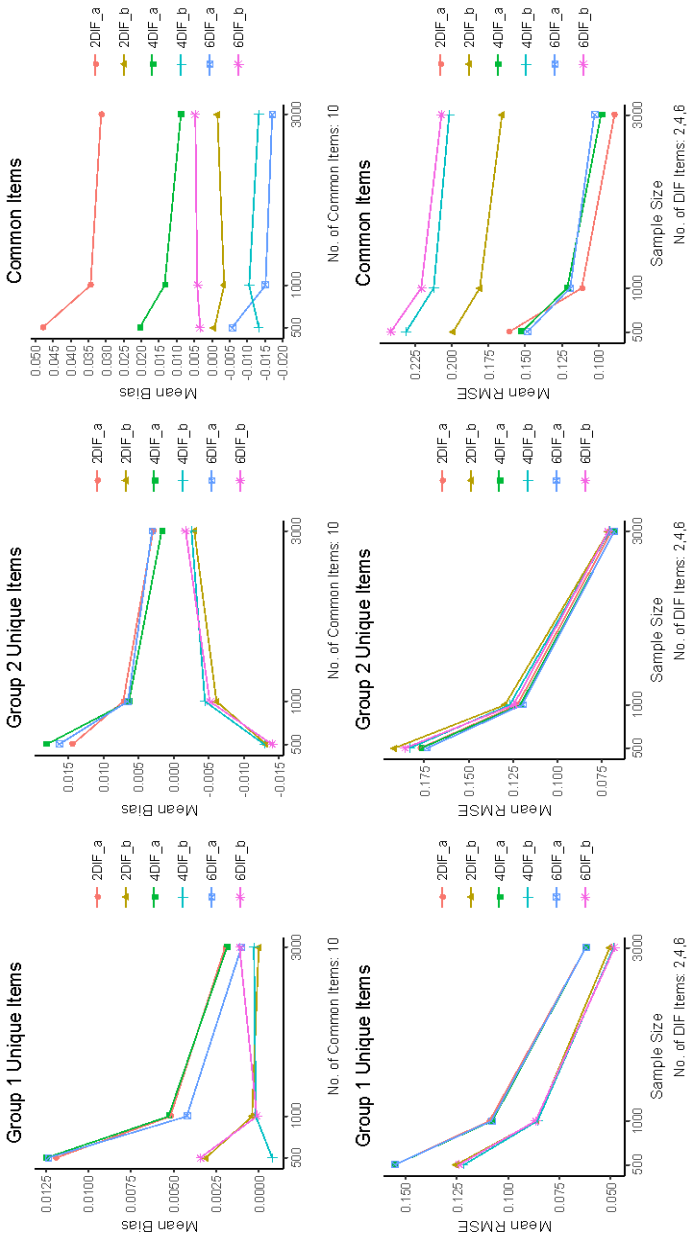

Figure 49 Large Uniform and Small Nonuniform DIF Favoring Group 2 10 Common Items Item Parameter Recovery by Group

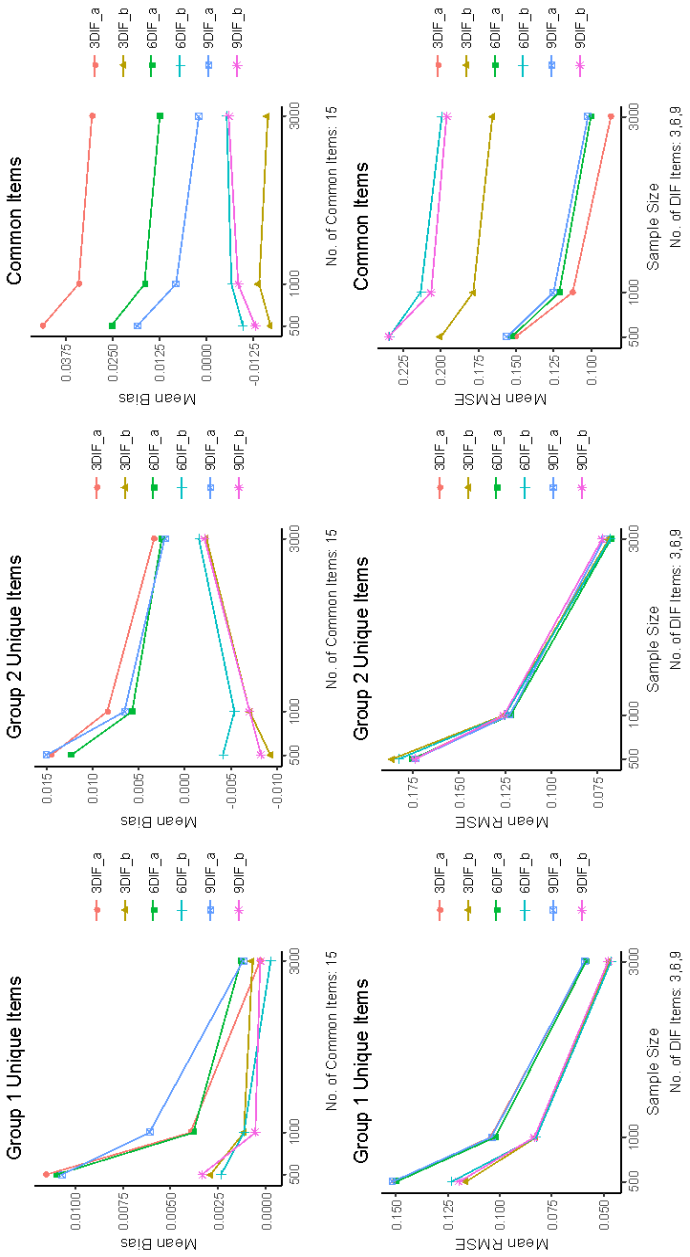

Figure 50 Large Uniform and Small Nonuniform DIF Favoring Group 2 15 Common Items Item Parameter Recovery by Group

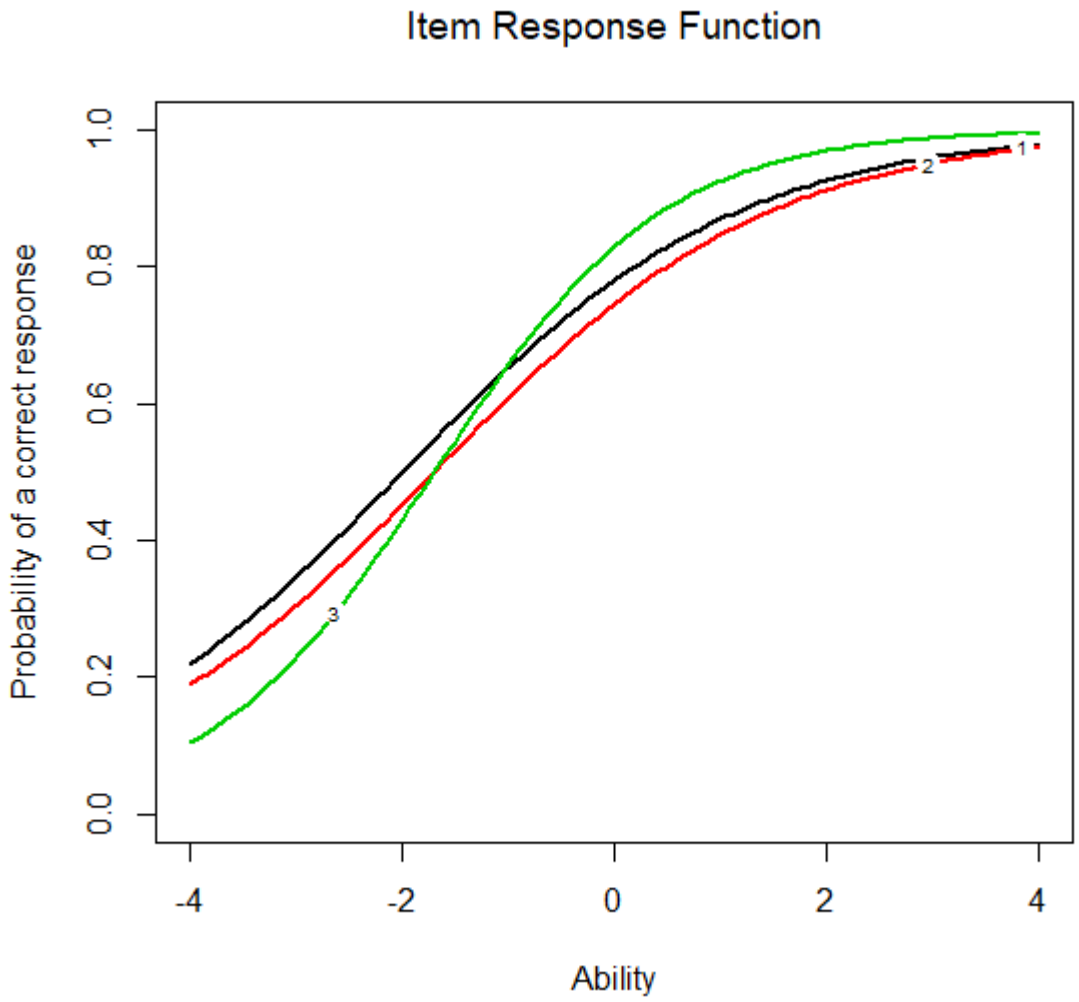

Figure 51 Invariant Item 1: a=0.63, b=-2.00; DIF Item 2: a=0.63, b=-1.70; DIF Item 3: a=0.93, b=-1.70

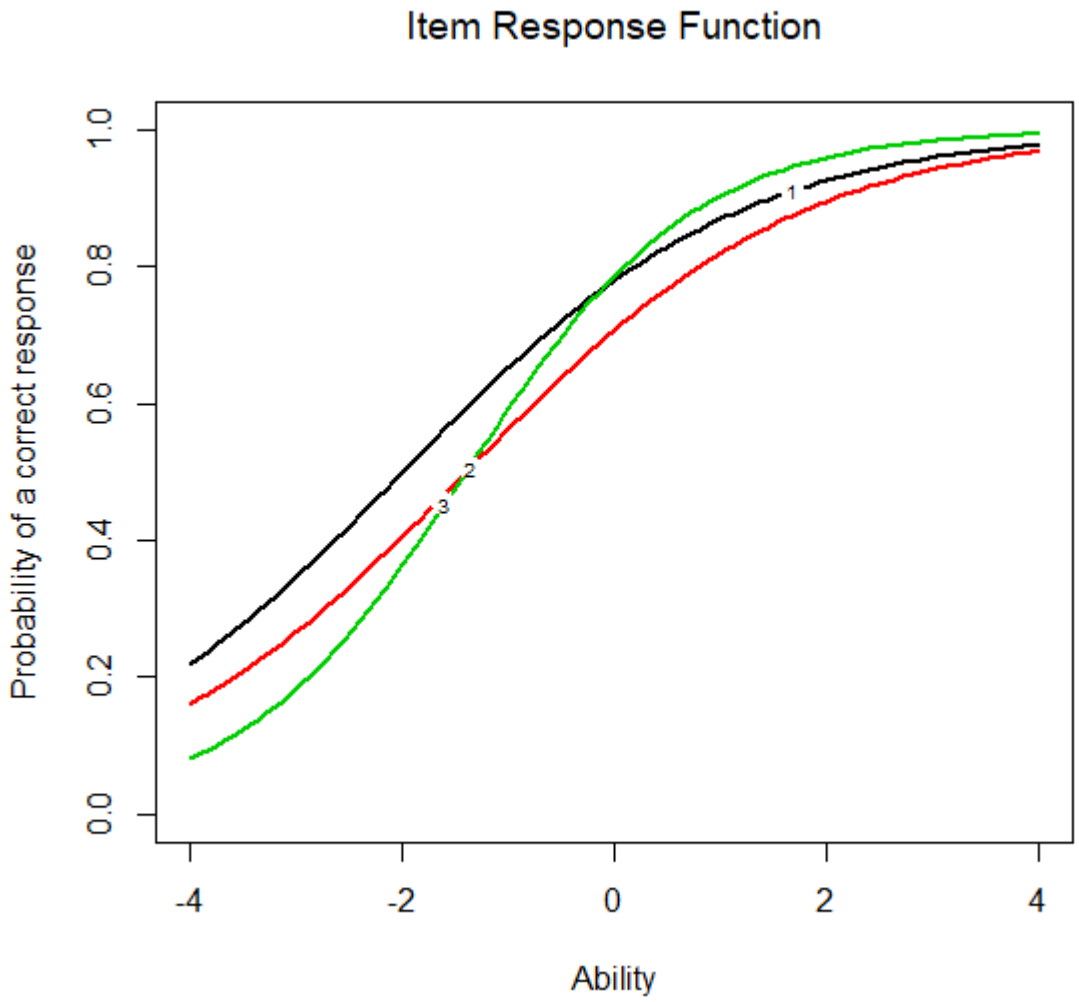

Figure 52 Invariant Item 1: a=0.63, b=-2.00; DIF Item 2: a=0.63, b=-1.40; DIF Item 3: a=0.93, b=-1.40

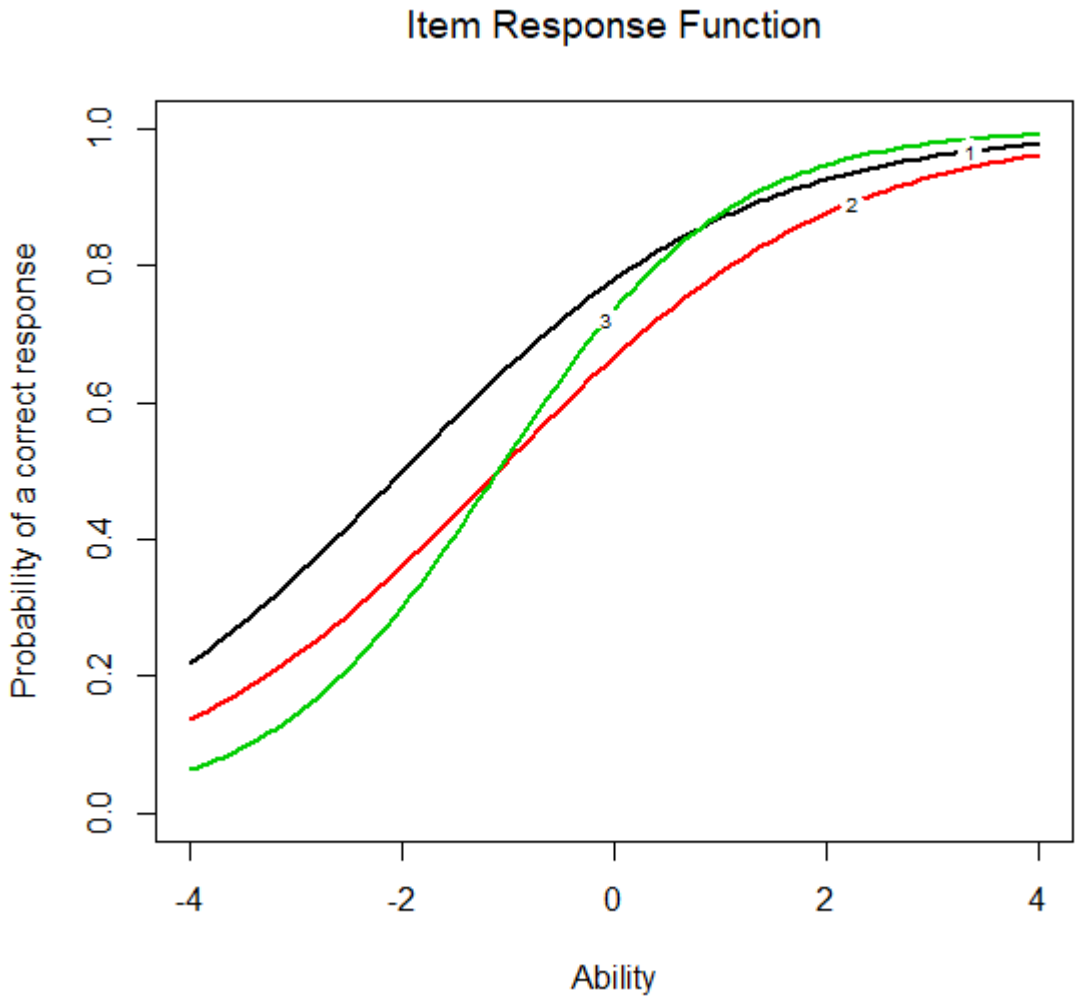

Figure 53 Invariant Item 1: a=0.63, b=-2.00; DIF Item 2: a=0.63, b=-1.10; DIF Item 3: a=0.93, b=-1.10