Regularization Methods for Detecting Differential Item Functioning:

Author: Jing Jiang

BOSTON COLLEGE

Lynch School of Education and Human Development

Department of
Measurement, Evaluation, Statistics, and Assessment

REGULARIZATION METHODS FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING

Dissertation
by

JING JIANG

submitted in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy

May 2019

REGULARIZATION METHODS FOR DETECTING

DIFFERENTIAL ITEM FUNCTIONING

by

Jing Jiang

Dissertation Chair: Zhushan Mandy Li

ABSTRACT

Differential item functioning (DIF) occurs when examinees of equal ability from different groups have different probabilities of correctly responding to certain items. DIF analysis aims to identify potentially biased items to ensure the fairness and equity of instruments, and has become a routine procedure in developing and improving assessments. This study proposed a DIF detection method using regularization techniques, which allows for simultaneous investigation of all items on a test for both uniform and nonuniform DIF. In order to evaluate the performance of the proposed DIF detection models and understand the factors that influence the performance, comprehensive simulation studies and empirical data analyses were conducted. Under various conditions including test length, sample size, sample size ratio, percentage of DIF items, DIF type, and DIF magnitude, the operating characteristics of three kinds of regularized logistic regression models: lasso, elastic net, and adaptive lasso, each characterized by their penalty functions, were examined and compared. Selection of optimal tuning parameter was investigated using two well-known information criteria AIC and BIC, and cross-validation. The results revealed that BIC outperformed other model selection criteria, which not only flagged high-impact DIF items precisely, but also

prevented over-identification of DIF items with few false alarms. Among the regularization models, the adaptive lasso model achieved superior performance than the other two models in most conditions. The performance of the regularized DIF detection model using adaptive lasso was then compared to two commonly used DIF detection approaches including the logistic regression method and the likelihood ratio test. The proposed model was applied to analyzing empirical datasets to demonstrate the applicability of the method in real settings.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Pursuing a Ph.D. has been a truly life-changing experience for me and I still cannot believe this amazing journey is at an end with the completion of this dissertation. There are many people have guided, assisted, and encouraged me to get this far, and the foremost among those people is my advisor and dissertation chair, Dr. Zhushan Mandy Li. I am deeply indebted to her for being constantly supportive, helpful, and caring in numerous ways. She guided me as a supervisor and a friend to help me overcome the problems and difficulties that I encountered as a student and a researcher throughout my doctoral education. She had confidence in me when I doubted myself and brought out the best in me. She was always there, always listened to me, always gave me sincere advice, and always cared about me. I would not be able to grow so fast, and this Ph.D. would not have been achievable without the support and nurturing of her.

I would also like to express my deepest gratitude to my committee members, Dr. Henry Braun and Dr. Ehri Ryu, for spending precious time reading and evaluating my dissertation. Thank you so much for your encouragement throughout the process and your support to make my final defense happen. Your insightful comments and constructive suggestions have been vital to improve the quality of my dissertation.

Additionally, my special thanks go to Dr. Louis Roussos, my mentor during my internship at Measured Progress. He has shared so much wisdom and valuable experience with me, both professionally and personally. His everlasting optimism and encouragement brought me new energy, inspired me to think creatively, and motivated me to work harder. I will always treasure this experience as a unique and an enlightening one.

I have been extremely fortunate to have the opportunity to intern at the Oculus Research team at Facebook. Being surrounded by talented, enthusiastic researchers from a variety of backgrounds is an extraordinary learning experience. The interdisciplinary nature of the research projects has opened my eyes to new ideas and new methods, which helped me figure out where the last piece of the puzzle needs to go in outlining my dissertation, and saved me a lot of time trying fruitlessly to tie everything together. I am therefore thankful to my mentors Dr. Cesare V. Parise and Dr. Raymond King for their valuable advice and unwavering guidance.

My love and gratitude for my family can hardly be expressed in words. I would like to thank my parents and grandparents for raising me up and understanding my choices. I owe many thanks to my mom, Sufang Liu, whose unflinching insistence, endless patience, and generous support have resulted in my achievement. I would not have been what I am today and where I am today without her loving upbringing. I would like to extend my gratitude to my aunt and uncle, who have been greatly supportive ever since I planned to study abroad and have given me a sense of belonging during the past years. I also thank with love to my husband and my friend, Zui Tao, for accompanying me along this winding journey until this point, and in the future.

The last words of acknowledgment I have saved for myself. I would like to thank myself for never giving up, for pushing past my limits, and for always being me.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Y, y	dependent variable / item response
X, x	independent variable / predictor
α, λ, γ	tuning parameter
β	coefficients for the regression model
τ	coefficients for the logistic regression DIF model
ω	weight for the model parameter
ε	error variable
σ^2	variance
$\boldsymbol{\Sigma}$	variance-covariance matrix
θ	latent ability
a	item discrimination
b	item difficulty
c	item pseudo-guessing
d	item intercept
D	scaling constant
π	probability of correct responses
$\boldsymbol{v}$	vector of item parameters
i	subscript for ith item
j	subscript for the jth independent variable
k	subscript for the kth subset of a dataset
n	subscript for the nth observation
p	subscript for pth person
s	subscript for sth observed test score
t	subscript for tth score level
R	subscript for persons in the reference group
F	subscript for persons in the focal group
G	group membership

I	number of items
J	number of independent variables
K	number of subsets of a dataset
N	number of observations
P	number of persons
S	total score for a test
T	total score level for a test
C	subset of a dataset
m	number of free parameters in a model

CHAPTER 1. INTRODUCTION

1.1 Statement of the Problem

Educational and psychological tests such as college admission tests, employment tests, licensure examinations, and mental health inventories are used to measure individuals' latent traits such as intelligence, attitudes, and other abilities or skills, and to distinguish between their trait levels, so as to make personal, social and political decisions regarding placement, advancement, and licensure (Clauser & Mazor, 1998). Considering the widespread usage and deep social implications of various tests and assessments, measurement bias has become an important issue in educational and psychological measurement over the past decades (Millsap & Everson, 1993). It is expected that test results should be comparable across groups, leading to fair comparisons based on these results. However, if some test items maintain an advantage for one group over another, the validity of test-based inferences might be threatened (Kane, 2006). These items are suspected of functioning differentially across groups, while they are exhibiting what researchers refer to as differential item functioning or DIF (Holland & Thayer, 1988; Dorans & Holland, 1993).

In the field of psychometrics, there is a distinction between impact and DIF, where the former refers to the difference between groups in test performance caused by a between-group difference on a valid skill (Ackerman, 1992), and the latter is a statistical property of an item indicating the item might be measuring different traits for individuals from separate groups. For example, the gender gap in the Programme for International Student Assessment test scores (González de San Román & de la Rica, 2016) is a good example of impact. DIF items are of great concern since they are putatively biased

against a particular group, indicating that participants from different demographic or socioeconomic groups have different probabilities of correctly responding to certain items and obtain different test scores, even after they have been matched on a measure of latent ability. That means, the group difference in performance on test items cannot be fully explained by the group difference in the latent construct targeted by the items (Crocker & Algina, 1986).

The study of bias in items and tests, as well as early work on DIF at the Educational Testing Service began from the end of the 1960s and early 1970s (e.g., Cardall & Coffman, 1964; Angoff & Ford, 1973). Since then, "psychometricians hastened to provide definitions of bias in terms of objective criteria, to develop rigorous and precise methods for studying bias, and to consider empirical investigations of test bias" (Berk, 1982). Several approaches for seeking out item bias in psychological and educational assessments were developed during that time period, such as the analysis of variance (ANOVA), with the null hypothesis of no significant interaction effect between the studied items and group membership. However, the group-item interaction in ANOVA was considered as an incomplete criterion since non-significant results cannot successfully rule out the existence of item bias (Osterlind, 1983).

Another early attempt to detect item bias was based on the Chi-square procedures (e.g., Scheuneman, 1979; Marascuilo & Slaughter, 1981; Mellenbergh, 1982), by examining the probabilities of individuals from different groups correctly responding to an item at every ability level. However, Holland and Thayer (1988) pointed out that the early-stage Chi-square tests for DIF detection always tended to reject the null hypothesis when the relevant sample size was large enough, and did not have a parametric measure

of DIF amount exhibited by the studied items. Further, they applied the Mantel-Haenszel (MH) Chi-square procedure (Mantel & Haenszel, 1959) for stratified samples to detecting DIF items. The MH method is possibly still the most widely used DIF detection technique nowadays.

In addition to the MH test, another popular non-parametric DIF detection approach is SIBTEST (Shealy & Stout, 1993), which focuses on psychometric dimensionality in the test rather than identification of aberrant functioning among a set of items. The MH test and SIBTEST are considered non-parametric because they are not based on any probability models. On the other hand, logistic regression (Swaminathan & Rogers, 1990) is widely used in detecting DIF items as a parametric method using likelihood functions, allowing for investigating both uniform DIF and nonuniform DIF effects.

A commonality among the MH, SIBTEST, and logistic regression DIF detection methods is that they do not rely on item response theory (IRT). IRT is a modern test theory grounded on a mathematical model representing the relationship between the latent traits, the properties of items on a test, and individuals' responses to test items. It primarily focuses on the item-level information, compared to classical test theory that focuses on the test-level information. Several DIF detection approaches were developed under the framework of IRT, such as Lord's χ^2 test statistic (Lord, 1980), the likelihood ratio test (Thissen, Steinburg, & Wainer, 1998), and Raju's area measures (Raju, 1988; 1990).

However, the aforementioned DIF detection approaches, including both IRT and non-IRT based techniques, are all conducted at the item level, focusing on analyzing each

item individually, which causes several non-trivial problems. First, placing individuals from different groups on a common metric is a necessary step in these DIF detection procedures since only individuals with a same ability level from the reference and focal groups should be compared. If individuals with different abilities are matched by mistake, the DIF detection results may be unreliable. Usually, individuals are matched according to some observed criteria, such as their test scores. However, test score is considered as an appropriate measure for individuals' latent ability levels only when the common metric consisting of anchor items is invariant across groups. Here, anchor items refers to the items whose parameters are constrained to be identical between groups. Previous studies (e.g., Wang & Yeh, 2003; Wang, 2004; Stark, Chernyshenko, & Drasgow, 2006; Wang, Shih, & Sun, 2012) suggested that, if a set of anchor items was contaminated, test score might not be a fair measure for latent traits, thus the type I error rates were often inflated. In addition, failure to identify invariant anchor items also led to type II errors in testing invariance (Johnson, Meade, & DuVernet, 2009). Obviously, in item-level DIF detection, the assumption that the anchor items (e.g., all items except the studied item) are supposed to be invariant across groups is not guaranteed, therefore it is highly possible to obtain inaccurate DIF detection results.

A number of scale purification procedures were developed to remove DIF items from the anchor set in order to improve the stability of anchor items and reduce the negative impact of non-invariant anchor items in DIF detection. Although there are different variations of scale purification procedures, generally it proceeds by testing each item for DIF using all other items as anchors. The items flagged as DIF items are removed from the anchor, and the studied items are re-evaluated using the remaining

anchor items. The procedure iterates until a same set of anchor items is identified as invariant in consecutive iterations. Previous studies showed that such purification procedures were able to improve DIF detection accuracy (Lautenschlager, Flaherty, & Park, 1994), but the process was tedious to some degree and was not able to handle highly contaminated scales.

Additionally, a significance test is conducted for each item in a typical item-level DIF study, indicating that for the whole test there exist multiple testing issues. Multiple testing increased the possibility of making a type I error at least once (Shaffer, 1995), resulting in incorrectly identifying non-DIF items as DIF items. Therefore, the validity of a test might be threatened if DIF items are falsely identified (Kim & Oshima, 2013). Several adjustment procedures such as the Bonferroni correction (Bonferroni, 1936) and the Holm method (Holm, 1979; Holland & Copenhaver, 1987) were used to control the type I error rate, and the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) was used to control the expected false discovery rates. Specifically, the Bonferroni correction compensates for the type I error increase by testing each statistical hypothesis at a reduced significance level that equals to the desired overall alpha level divided by the total number of hypotheses. The Holm method is a more powerful sequential version of the Bonferroni correction with slightly different threshold levels. The Benjamini-Hochberg procedure controls the false discovery rate by defining a sequential p-value procedure. However, previous research indicated that these adjustment approaches only worked for certain DIF detection procedures. When the DIF magnitude became larger, the power of detecting DIF items correctly was reduced substantially after applying the adjustments in most conditions (Kim & Oshima, 2013).

Chapter 1. Introduction

Simultaneous investigation of all items on a test for DIF seems to be an intuitive solution to address the limitations caused by the item-level DIF detection approaches, which can be realized, for example, by using regularization methods for estimation. Regularization refers to the procedure of introducing additional information such as a penalty for complexity to prevent overfitting and solve ill-posed problems (Hastie, Tibshirani, & Friedman, 2009). However, existing applications of regularization methods in DIF detection were largely based on the Rasch model, so they were only able to detect uniform DIF effects (e.g., Magis, Tuerlinckx, & De Boeck 2015; Schauberger, 2015; Tutz & Schauberger, 2015). In practice, the assumptions of the Rasch model are often unrealistic for test items, since it is rare to have all items with an identical discrimination across the ability continuum. The multi-parameter IRT models such as the two-parameter logistic (2PL) or three-parameter logistic (3PL) models are alternatives to the Rasch model when the assumption of equal discrimination is untenable. In many applications of IRT in the testing industry such as scaling, equating, standard setting, and DIF detection, the 2PL and 3PL models are the most popular psychometric models (San Martín, González, & Tuerlinckx, 2015). In this case, it is desirable to develop a more general DIF detection model that allows for simultaneous detection of both uniform and nonuniform DIF effects under the framework of multi-parameter IRT models using appropriate regularization methods.

In addition, current DIF studies using regularization techniques employed lasso (Tibshirani, 1996), an $\ell 1$ penalty, to perform feature selection and determine DIF effects. Lasso encourages shrinking more coefficients to zero, which may encounter problems when there exist correlated variables. This is because lasso assigns a non-zero value to

one of the variables arbitrarily, and reduces the remaining model coefficients to zero, which provides incomplete information on these correlated variables. Therefore, other penalty functions such as the elastic net (Zou & Hastie, 2005) and the adaptive lasso (Zou, 2006) can be considered in DIF studies.

Moreover, which measure—the observed test score or the estimated IRT ability—is a better proxy of person's latent ability remains questionable. Using test score makes DIF analysis more efficient since the ability estimation process can be skipped, but the results may be misleading since test score is not always a representative measure of latent ability from the IRT perspective. On the other hand, when using ability estimates in the DIF detection model, the accuracy of the analysis results is then relied on the parameter recovery, which might be problematic when the sample size is small and the test length is long (e.g., Drasgow, 1989; Stone, 1992; Sahin & Anil, 2017).

1.2 Purposes and Research Questions

The purpose of this study is to propose a DIF detection model using regularization techniques, which allows for simultaneous investigation of all items on a test for both uniform and nonuniform DIF detection. Comprehensive simulation studies were conducted under various manipulated conditions in order to evaluate the performance of the proposed DIF detection model and investigate the factors that influence the performance. Moreover, the performance of the proposed DIF detection model was compared to two commonly used DIF detection approaches including the logistic regression method and the likelihood ratio test. In addition, the proposed model was used to analyze two empirical datasets in order to demonstrate the applicability of the method

in real settings by comparing the results with the existing DIF detection approaches. Specifically, the following research questions were addressed in this study:

1. Which penalty function yields the best operating characteristics in DIF detection? In order to answer this question, two preliminary questions were studied beforehand.
    a. Which model selection technique selects the optimal tuning parameter in terms of DIF detection performance?
    b. Which of the ability measures, the observed test score or the IRT ability estimate, has a better performance in detecting DIF items?
2. How does each of the manipulated factors impact the number of items flagged correctly and incorrectly as DIF items using the proposed DIF detection model? The manipulated factors and the levels of each factor were:
    - Two test lengths: 20 items, 40 items;
    - Three sample sizes: 1000, 2000, 4000 examinees;
    - Two sample size ratios: 1:1, 4:1 (reference/focal group sizes—500/500, 800/200; 1000/1000, 1600/400; 2000/2000, 3200/800);
    - Two percentages of DIF items: 10%, 20%;
    - Three DIF types: uniform DIF only, nonuniform DIF with drift on the discrimination parameters only, nonuniform DIF with drift on the difficulty and discrimination parameters;
    - Different DIF magnitudes: 0.4, 0.8 for drift on the difficulty parameters, and 0.5, 1.0 for drift on the discrimination parameters.

3. What are the differences between the proposed method and other existing techniques in terms of their DIF detection performance?

1.3 Significance of the Study

DIF analysis is a critical part of developing and evaluating assessments, which has become a routine procedure in practice since it can be used to assess measurement bias and therefore test claims about validity (Martinková et al., 2017). A large number of parametric and nonparametric procedures have been developed to detect DIF effects, with or without the use of IRT. The proposed DIF detection model differs from most commonly used item-level DIF detection approaches as all items on a test can be examined simultaneously using a single coherent model, which avoids the strict assumption that all other items as anchors except the studied item should be free from DIF in the scale, and overcomes the problem caused by multiple testing.

Further, the proposed method allows for detecting not only uniform DIF items but also nonuniform DIF items. Although uniform DIF occurs more often than nonuniform DIF in tests and assessments, DIF does not always occur equally over the ability continuum, and the amount of DIF may vary appreciably across the latent ability continuum in real data (Narayanan & Swaminathan, 1996). Therefore, this study is more practical to the current measurement industry since none of the existing regularization methods addresses nonuniform DIF detection.

In addition, this study investigates the use of different extensions of lasso in order to examine whether these extensions remedy the limitations of the traditional lasso methodology and improve the operating characteristics in DIF detection.

Lastly, since the proposed DIF detection model is a logistic regression model, it has great flexibility in comparing multiple groups, examining categorical or continuous DIF items, and detecting multiple covariates by changing existing variables or adding additional variables in the model.

1.4 Dissertation Organization

Chapter 1 emphasizes the importance of DIF analysis in test development and test validation, followed by a brief overview of the history and development of DIF detection methods, as well as the limitations of existing approaches. Then the research purpose, questions of interest, and significance of the study are described in sequence. This chapter ends with an outline of the dissertation organization.

Before stepping into DIF, Chapter 2 begins with an introduction to IRT, including the fundamental assumptions, popular IRT models and widely used parameter estimation techniques. Next, a detailed review of both theoretical definitions and statistical procedures associated with DIF is provided. At last, in order to propose a new DIF detection model using regularization methods, various concepts including the definition and purpose of regularization, different types of penalty functions, common model selection techniques, and current applications are illustrated.

Chapter 3 explicates the proposed regularized logistic regression DIF detection model, as well as the estimation and evaluation procedures associated with the model. Two preliminary analyses were conducted to help determine the model setting and choose the optimal model selection criterion. In addition, the research design for the

comprehensive simulation studies and the background information of two empirical datasets are provided in this chapter.

Chapter 4 presents the results of two simulation studies and two empirical data analyses, answering three proposed research questions. Finally, Chapter 5 discusses the results, interprets the implications of the findings, and points out potential directions for future studies.

CHAPTER 2. LITERATURE REVIEW

2.1 Item Response Theory

Item response theory (IRT), also known as latent trait theory, is a modern psychological and educational test theory, which establishes a mathematical model of person and item parameters to represent the relationship between individuals' responses to test items and their levels of latent ability. IRT has more or less replaced the role classical test theory (CTT) played in developing and analyzing tests and assessments, due to its potential to address the disadvantages of CTT. CTT produces findings that are both sample-dependent and scale-dependent, leading to serious logical drawback if the measurement performance of an instrument is affected by the sample it is supposed to be measuring and vice versa (Petrillo, Cano, McLeod, & Coon, 2015).

2.1.1 Item Response Theory Models

IRT is commonly used to evaluate how well the entire instrument and individual items perform in measuring person's abilities, skills, attitudes, or other latent traits. Within the IRT framework, many models have been developed for analyzing and scoring different types of instruments (Hambleton & Jones, 1993). For example, the item response can be dichotomous (e.g., multiple-choice item), polytomous (e.g., Likert-type item) or continuous (e.g., slider scale item); the scoring category can be ordered or unordered; the latent trait can be measured within a unidimensional or multidimensional framework; and the relationship between the latent ability and item responses can be modeled using different statistical functions (e.g., the logit model or the normal ogive model).

Chapter 2. Literature Review

Typically, IRT models can be classified into unidimensional and multidimensional models, and this study only focuses on the unidimensional case. Three basic assumptions are required for unidimensional IRT models. First, the unidimensionality assumption assumes that there is one single continuous latent ability variable that accounts for the response behavior on test items. Second, the local independence assumption states that an individual's response to an item is only due to this individual's location on the continuous latent variable and is not related to how they respond to other items. Third, the item response can be modeled by using a mathematical item response function (IRF), which gives the probability that an individual with a given ability level can answer an item correctly.

The 3PL IRT model is the most general model for dichotomous items (Birnbaum, 1968) [1]. Assume that person p ($p = 1, \dots, P$) responds to item i ($i = 1, \dots, I$) on a test, the corresponding response is expressed as $y_{ip} = 1$ if person p answers item i correctly, otherwise $y_{ip} = 0$, the 3PL model can be written as:

$$P\left(y_{ip} = 1 \middle| \theta_p, a_i, b_i, c_i\right) = \pi_{ip} = c_i + (1 - c_i)\frac{exp\left[a_i\left(\theta_p - b_i\right)\right]}{1 + exp\left[a_i\left(\theta_p - b_i\right)\right]} \qquad (2.1)$$

where θ_p represents the ability level of person p, the parameters a_i, b_i, and c_i represent item discrimination, difficulty, and pseudo-guessing parameter correspondingly, and $P\left(y_{ip} = 1 \middle| \theta_p, a_i, b_i, c_i\right)$ represents the probability π_{ip} that person p with ability level of θ_p responds to item i correctly with $y_{ip} \sim \text{Bernoulli}(\pi_{ip})$.

[1] Although the four-parameter logistic IRT model (Barton & Lord, 1981) was developed as an extension of 3PL model by adding an upper-asymptote parameter, it is rarely used in practice.

Alternatively, the normal ogive model is sometimes used as an alternative to the logistic model, which is actually the first IRT model for measuring latent traits (Mosier, 1940; 1941). A mathematical expression of the three-parameter normal ogive model is:

$$P(y_{ip} = 1|\theta_p, a_i, b_i, c_i) = c_i + (1 - c_i)\int_{-\infty}^{a_i(\theta_p - b_i)} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \qquad (2.2)$$

where z is a standardized score involving an examinee's latent trait score, and two item parameters a_i and b_i., The normal ogive model is not very popular compare to the logistic model in practice due to its inconvenience and complexity of computation. The item parameters in the normal ogive model and the logistic model can be connected by using a scaling factor D whose value is set to either 1.0 or 1.7:

$$P(y_{ip} = 1|\theta_p, a_i, b_i, c_i) = \pi_{ip} = c_i + (1 - c_i)\frac{exp[Da_i(\theta_p - b_i)]}{1 + exp[Da_i(\theta_p - b_i)]} \qquad (2.3)$$

With $D = 1.7$, two models yield almost equivalent values of item parameters, and under this situation the model parameters are placed on what is referred to as the "normal metric" (Han, 2013). In this study, since preserving consistent interpretations between two models are not important, the value of *D* was set to 1.0 and the model parameters were placed on the so-called "logistic metric".

The 2PL model (Birnbaum, 1957) and the one-parameter logistic (1PL) model are special cases of the 3PL model. In terms of the 2PL model, there is no pseudo-guessing parameter so that all c_i $(i = 1, \dots, I)$ are restricted to zero, therefore the model in Equation 2.1 becomes:

$$P(y_{ip} = 1|\theta_p, a_i, b_i) = \frac{exp[a_i(\theta_p - b_i)]}{1 + exp[a_i(\theta_p - b_i)]} \qquad (2.4)$$

Additionally, the 1PL model can be obtained by restricting $a_i = 1$ $(i = 1, \dots, I)$:

$$P(y_{ip} = 1|\theta_p, b_i) = \frac{exp(\theta_p - b_i)}{1 + exp(\theta_p - b_i)} \quad (2.5)$$

Although conceptually different, the 1PL model and the Rasch dichotomous model (Rasch, 1960) are the same from a purely mathematical standpoint.

2.1.2 Parameter Estimation

There are several techniques for estimating the item and ability parameters in IRT models, such as joint maximum likelihood estimation (JMLE), marginal maximum likelihood estimation (MMLE) and Bayesian approaches. Conditional maximum likelihood estimation (CLME) is also an alternative for maximum likelihood estimation (MLE), especially for models with a simple sufficient statistic, such as the Rasch model (Andersen, 1970). However, the 2PL and 3PL models do not have sufficient statistics, so CMLE will not be discussed in this section.

All these MLE methods rely on the independence assumption that individuals are independent of each other, and the item responses of a given individual are independent. Therefore, use the 2PL model as an example, the likelihood function for person *p* with a response pattern y_p can be defined as the joint product of probability functions:

$$L(y_p|\theta_p, a_i, b_i) = \prod_{i=1}^{I} P(y_{ip}|\theta_p, a_i, b_i) \quad (2.6)$$

Equation 2.6 can be used to estimate the ability parameters when the item parameters are known. Additionally, the likelihood function for item *i* is defined as:

$$L(y_i|\theta_p, a_i, b_i) = \prod_{p=1}^{P} P(y_{ip}|\theta_p, a_i, b_i) \quad (2.7)$$

And Equation 2.7 can be used to estimate the item parameters when person's abilities are known. Similarly, the likelihood function for the data matrix is defined as:

$$L(y|\theta_p, a_i, b_i) = \prod_{p=1}^{P} \prod_{i=1}^{I} P(y_{ip}|\theta_p, a_i, b_i) \quad (2.8)$$

JMLE (Lord, 1968) treats both item and ability parameter as unknown but fixed. It proceeds by estimating $L(y|\theta_p, a_i, b_i)$ with respect to ability parameter θ_p and item parameters a_i, b_i simultaneously. JMLE can be divided into the MLE of item parameters and the MLE of ability parameters. These two steps are iterated so that the item and ability parameters are estimated back and forth until a convergence criterion is met. One major problem with JMLE is that parameter estimates do not exist for extreme scores of individuals or items (e.g., an individual answers all items correctly or incorrectly, or all individuals answer an item correctly or incorrectly). Also, previous research found that JMLE estimates were statistically inconsistent and biased particularly for short tests or small samples (Andersen, 1973).

Unlike JMLE, MMLE estimates item parameters without having to estimate ability parameters, which is accomplished by integrating over the ability distribution to eliminate these parameters (Bock & Lieberman, 1970). The theoretical marginal likelihood is defined as:

$$L(y|\theta_p, a_i, b_i) = \prod_{p=1}^{P} \int_{-\infty}^{+\infty} \prod_{i=1}^{I} P(y_{ip}|\theta_p, a_i, b_i) f(\theta_p)\, d\theta_p \quad (2.9)$$

Item parameter estimates are obtained by maximizing the marginal likelihood function in Equation 2.9. MMLE is more complicated to implement but the parameter estimates are consistent under the hypothesis of normality of the latent trait. The computational burden

can be reduced by using the expectation–maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), which is efficient in finding the maximum likelihood estimates of parameters in the presence of unobserved random variables. The EM iteration alternates between performing an expectation step, which creates a function for the expectation of the likelihood function based on the initial values or current estimates of model parameters, and performing a maximization step, which updates old parameter estimates and obtains new estimates by maximizing the expected likelihood function in the expectation step.

Ability parameters can be estimated subsequently, using Bayesian procedures (Bock & Mislevy, 1982), which basically entails combining the likelihood function with a prior distribution to estimate the posterior distribution of ability:

$$f(\theta_p|y, a_i, b_i) \propto L(y|\theta_p, a_i, b_i) f(\theta_p) \quad (2.10)$$

In Equation 2.10, $f(\theta_p)$ is the prior distribution representing some prior belief about the ability distribution, and $f(\theta_p|y, a_i, b_i)$ is the posterior distribution of ability for person p. The maximum a posteriori (MAP) and expectation a posteriori (EAP) approaches are often employed to estimate ability parameters. Specifically, the Bayesian MAP estimator is the mode of the posterior distribution, and the EAP estimator equals the mean of the posterior distribution.

2.2 Differential Item Functioning

2.2.1 Definition

An item is biased when individuals from one group are less likely to correctly answer it compared to individuals from another group, which is because some

characteristics of test items or testing situations are not relevant to the test purpose (Zumbo, 1999). Such items exhibit differential item functioning (DIF), a necessary but not sufficient condition for item bias (Clauser & Mazor, 1998), indicating that the probabilities of correctly responding to certain items are unexpectedly different for people from group to group, even after they have been matched on the ability of interest (Holland & Wainer, 1993). In other words, item bias implies DIF; but if an item shows DIF, it is not sufficient to report this item is biased. Items with DIF may reflect measurement bias and lead to discrimination against particular groups. Typically, follow-up analyses such as content analysis, empirical evaluation or other judgmental approaches are required to determine the presence of item bias (Zumbo, 1999). Therefore, since the DIF investigation and detection procedures completely rely on statistical techniques and the analysis results can help flag potentially biased items, DIF is a very important indicator for researchers, educators, and policymakers to examine whether test items display the same statistical properties for individuals from different groups within the population.

2.2.2 Types of DIF

Conceptually, an item displaying DIF or not can be assessed by comparing the item characteristic curves (ICCs) of different groups on this item. ICC is a graphical representation of IRF, showing the relationship between the latent ability level and the probability of correct response. As shown in Figure 2.1, if the reference and focal group ICCs are very close to each other, the item is more likely to be a non-DIF item. Otherwise, an item is considered to display DIF if there is significant difference between the ICCs across groups.

DIF can be classified into uniform DIF and nonuniform DIF. In Figure 2.2, the ICCs do not cross with each other. This type of DIF is defined as uniform DIF, indicating the studied item consistently gives one group an advantage across all ability levels. In Figure 2.3, the ICCs do cross each other. In this case, the item shows nonuniform DIF, indicating that an item gives an advantage to a reference group at one end of the ability continuum while favors the focal group at the other end (Walker, 2011). In IRT, an item showing uniform DIF only varies in the difficulty parameter, while an item displaying nonuniform DIF varies in the discrimination parameter, and possibly varies in the difficulty parameter (Mellenbergh, 1982).

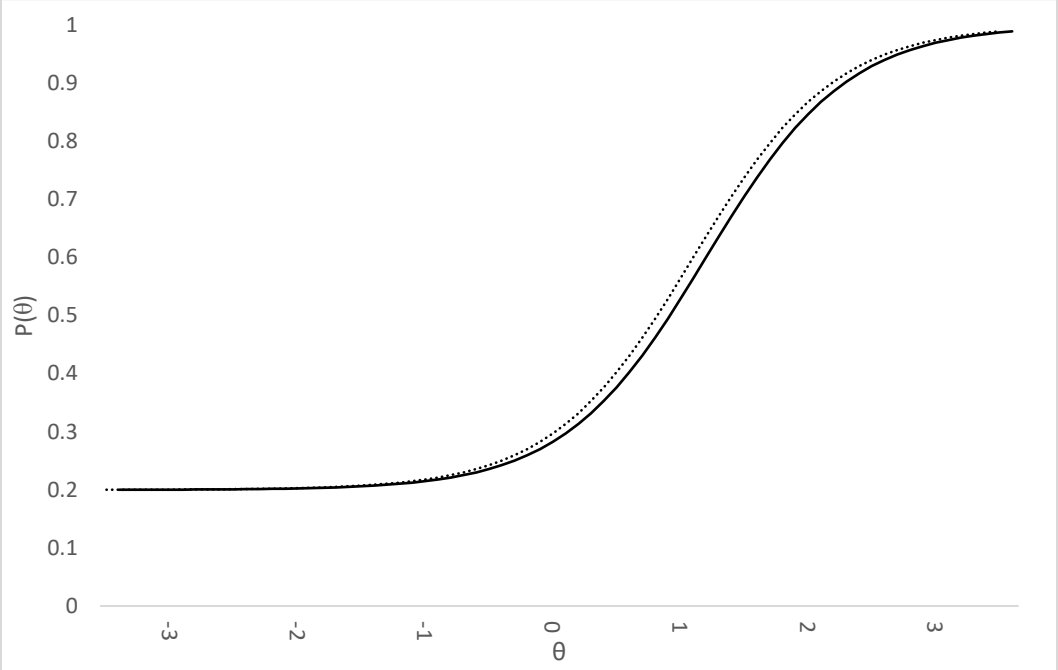

Figure 2.1 The Reference and Focal Group ICCs for a Non-DIF Item

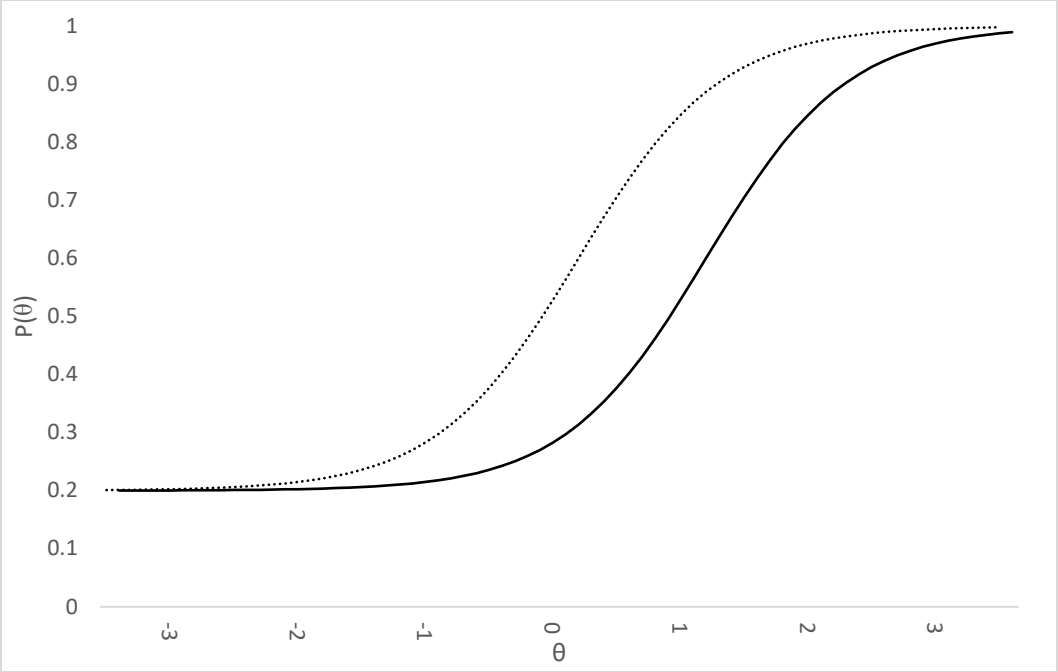

Figure 2.2 The Reference and Focal Group ICCs for a Uniform DIF Item

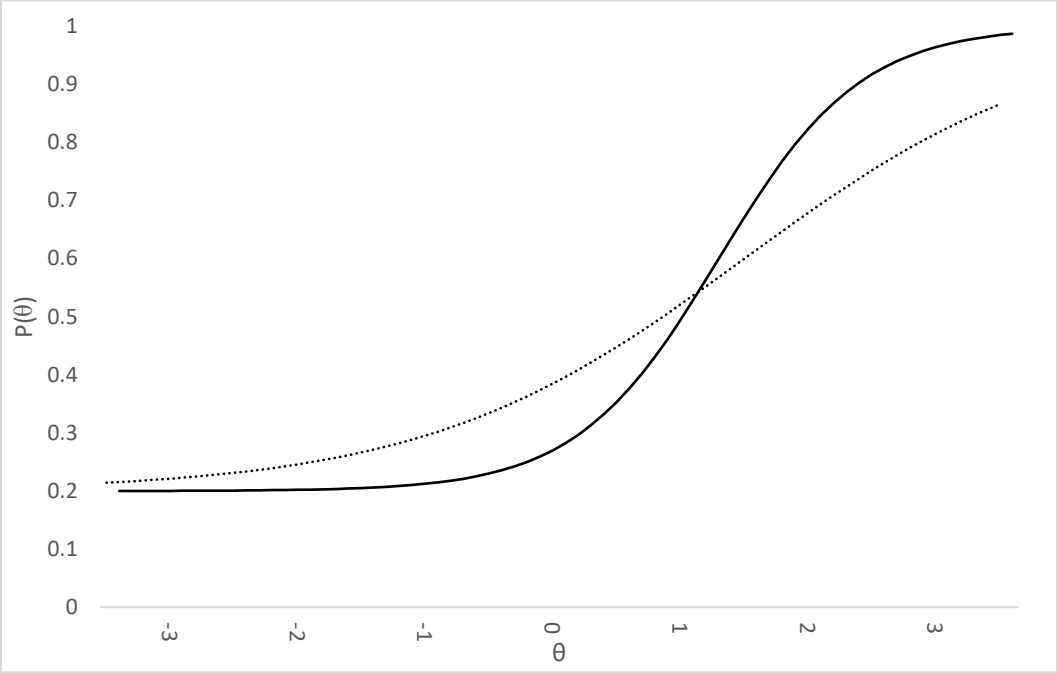

Figure 2.3 The Reference and Focal Group ICCs for a Nonuniform DIF Item

Previous research found that the presence of DIF items had negative impacts on measurement consequences such as inaccurate IRT ability estimates (e.g., Wells, Subkoviak, & Serlin, 2002) and inflated type I error rates in DIF detection (e.g., Li, Brooks, & Johanson, 2012). Therefore, how much difference between item parameters is non-negligible becomes a question of interest. Wells et al. (2002) suggested that a difference of 0.4 in difficulty parameters and a difference of 0.5 in discrimination parameters had a minimal impact on ability estimates, demonstrating the robustness of IRT when the invariance property is violated.

2.2.3 Popular DIF Detection Methods

Many methods were developed to detect DIF over the years. Fundamentally, most of the studies focus on comparing two pre-determined groups—the reference group and the focal group. The Mantel-Haenszel method (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), and SIBTEST (Shealy & Stout, 1993) are popular non-IRT DIF detection methods. IRT-based approaches are also widely used by comparing either the item parameters or the item response functions between groups, such as Lord's χ^2 test (Lord, 1980), the likelihood ratio test (Thissen et al., 1998), and Raju's area measures (Raju, 1988; 1990).

2.2.3.1 The Mantel-Haenszel Method

The Mantel-Haenszel (MH) method is based on analyzing a contingency table, which displays the multivariate frequency distribution of variables of interest. This procedure compares the probabilities of a correct response between the focal and reference group members with the same ability level, and is popular for assessing uniform DIF in dichotomously scored items. The MH statistic determines if the item

responses are independent of group membership after conditioning on the observed scores.

Let t represent the *t*th ($t = 1, \dots, T$) score level, a two-by-two contingency table of a studied item can be constructed as follows:

Table 2.1 A Contingency Table for a Studied Item at the *t*th Score Level

		Group Membership		**Total**
		Reference	**Focal**	
Score on the Studied Item	**1 (correct)**	N_{R1t}	N_{F1t}	N_{1t}
	0 (incorrect)	N_{R0t}	N_{F0t}	N_{0t}
Total		N_{Rt}	N_{Ft}	N_t

According to Table 2.1, at the *t*th score level, $N_{R1t}, N_{R0t}, N_{F1t}, N_{F0t}$ are the observed counts of individuals in each cell, $N_{Rt} = N_{R1t} + N_{R0t}$ and $N_{Ft} = N_{F1t} + N_{F0t}$ are the number of reference and focal group members, $N_{1t} = N_{R1t} + N_{F1t}$ and $N_{0t} = N_{R0t} + N_{F0t}$ are the number of correct and incorrect responses, and $N_t = N_{Rt} + N_{Ft} = N_{1t} + N_{0t}$ represents the total number of individuals.

In sum, there are *T* contingency tables for each studied item. The common odds ratio for the *T* contingency tables is defined in Equation 2.11:

$$odds\ ratio = \frac{\sum_{t=1}^{T} N_{R1t}N_{F0t}/N_t}{\sum_{t=1}^{T} N_{R0t}N_{F1t}/N_t} = \frac{\sum_{t=1}^{T} N_{R1t}N_{F0t}}{\sum_{t=1}^{T} N_{R0t}N_{F1t}} \quad (2.11)$$

When the odds ratio equals 1, the probabilities of correct responses are equal in both groups, indicating there is no association between observed scores and group membership. When the odds ratio is greater than 1, reference group members are more likely to answer the studied item correctly. Otherwise, when the odds ratio is less than 1, the focal group members are more likely to answer the studied item correctly. Therefore,

the null hypothesis is the odds ratio equal to 1, while the alternative hypothesis is the odds ratio unequal to 1.

The MH statistic with a continuity correction can be calculated as:

$$\chi^2_{MH} = \frac{\{|\sum_{t=1}^{T}[N_{R1t} - E(N_{R1t})]| - 0.5\}^2}{\sum_{t=1}^{T} var(N_{R1t})} \quad (2.12)$$

where $E(N_{R1t}) = \frac{N_{Rt}N_{1t}}{N_t}$ and $var(N_{R1t}) = \frac{N_{Rt}N_{Ft}N_{1t}N_{0t}}{N_t^2(N_t-1)}$. The MH statistic approximately follows a Chi-square distribution with one degree of freedom (Mantel & Haenszel, 1959).

2.2.3.2 Logistic Regression

Swaminathan and Rogers (1990) first applied logistic regression to DIF detection, which is a model-based approach. The logistic regression DIF approach has a statistical significance test and a measure of DIF effect size, which can be conceptualized as a link between the contingency table methods and the IRT methods (Clauser & Mazor, 1998). Logistic regression is more general and flexible than the model that underlies the MH procedure, and is more powerful in detecting nonuniform DIF. When using the logistic regression model to detect DIF items, the item response is treated as an outcome variable, while the latent ability, group membership, as well as an interaction between group membership and latent ability are treated as predictors. Equation 2.13 defines the logistic regression DIF model.

$$logit\left[P\left(y_{pg} = 1\right)\right] = \tau_0 + \tau_1 s_p + \tau_2 G_p + \tau_3 s_p G_p \quad (2.13)$$

where s_p is the test score of person p, G_p represents group membership which equals 1 if person p belongs to the reference group and 0 otherwise, and $s_p G_p$ is the product of these two independent variables representing the interaction between ability and group membership. Additionally, $\tau_0, \tau_1, \tau_2, \tau_3$ represent the intercept, the effect for ability, the

effect for group, and the interaction of ability and group for the studied item. τ_0 and τ_1 can be further considered as the counterparts of item difficulty and item discrimination parameters in the IRT framework when $G_p = 0$ (Li, 2014). The MH procedure can be thought of as being based on the logistic regression model if the ability variable is discrete, and no interaction is included in the model (Swaminathan & Rogers, 1990).

The logistic regression coefficients τ_2 and τ_3 are referred to as the DIF parameters, and the detection of DIF items is realized by testing the significance of these two DIF parameters. If the studied item is DIF-free, both τ_2 and τ_3 are equal to 0; if the studied item shows uniform DIF, $\tau_2 \neq 0$ but $\tau_3 = 0$; if the studied item shows nonuniform DIF, $\tau_3 \neq 0$ and τ_2 can be either zero or non-zero. The difference between the minus twice the log-likelihood of the parsimonious model including only τ_0 and τ_1 and the augmented model including $\tau_0, \tau_1, \tau_2, \tau_3$ is associated with a Chi-square distribution with two degrees of freedom. If the null hypothesis of no DIF is rejected, the studied item is flagged as a DIF item and should be reviewed by experts afterwards (Jodoin & Gierl, 2001).

2.2.3.3 SIBTEST

The full name of SIBTEST is simultaneous item bias test, which is a nonparametric procedure that not only estimates the amount of DIF in test items, but also tests whether that amount is different from zero statistically (Bolt, 2000). SIBTEST was developed under the multidimensional IRT framework, which examines DIF at the test level and provides a statistical test to investigate if DIF is present on a test (Narayanan & Swaminathan, 1994). A regression-based correction technique is used in SIBTEST to help match individuals from different groups at the same latent ability level rather than

the same observed score level, and this method showed improved performance in controlling the type I errors (Shealy & Stout, 1993; Roussos & Stout, 1996).

The statistical hypotheses under SIBTEST are $H_0: B = 0$ vs. $H_1: B \neq 0$, where B is a parameter specifying the DIF magnitude as follows:

$$B = \int_{\theta} [P(\theta, R) - P(\theta, F)] f_F(\theta) d\theta \quad (2.14)$$

In Equation 2.14, θ represents the ability level, $P(\theta, R)$ and $P(\theta, F)$ are the probabilities of correct responses on an item in the reference and focal groups respectively, and $f_F(\theta)$ is the density function for θ in the focal group. Thus, B is a weighted expected score difference between the reference and focal group members of a same ability level on the studied item. B can be approximated using the observed test score s ($s = 0, \ldots, S$) on a subset of anchor items:

$$\hat{B} = \sum_{s=0}^{S} p_s (\bar{Y}_{Rs} - \bar{Y}_{Fs}) \quad (2.15)$$

where p_s represents the proportion of individuals from the pooled reference and focal groups getting score *s* on the anchor items, and $\bar{Y}_{Rs}$ and $\bar{Y}_{Fs}$ are the mean scores for all reference group and focal group members obtaining score *s* on the anchor items respectively.

However, when $f_F(\theta) \neq f_R(\theta)$ meaning that two latent distributions are different, $\hat{B}$ might be easily inflated and biased. A modified SIBTEST was proposed by Shealy and Stout (1993) employing a regression-based correction technique to obtain a corrected version of $\bar{Y}_{Rs}$ and $\bar{Y}_{Fs}$, $\bar{Y}^*_{Rs}$ and $\bar{Y}^*_{Fs}$, which are the adjusted mean scores for reference and focal group members matched on the estimated true score correspondingly. Therefore,

$$\hat{B}^* = \sum_{s=0}^{S} p_s(\bar{Y}_{Rs}^* - \bar{Y}_{Fs}^*) \quad (2.16)$$

The SIBTEST statistic is then constructed based on the new estimate $\hat{B}^*$:

$$SIB = \frac{\hat{B}^*}{\hat{\sigma}(\hat{B})} \quad (2.17)$$

where $\hat{\sigma}(\hat{B}) = \left[\sum_{s=0}^{S} p_s^2 \left(\frac{\hat{\sigma}^2(Y|s,R)}{N_{Rs}} + \frac{\hat{\sigma}^2(Y|s,F)}{N_{Fs}}\right)\right]^{\frac{1}{2}}$, N_{Rk} and N_{Fk} represent the number of individuals in the reference and focal groups with score *k* on the anchor items correspondingly. Since the test statistic *SIB* has an asymptotic distribution of $N(0,1)$ under the null hypothesis of no DIF, the null hypothesis is rejected if *SIB* exceeds the $\frac{100(1-\alpha)}{2}$ percentile point of the standard normal distribution using a nondirectional hypothesis test (Shealy & Stout, 1993).

2.2.3.4 Lord's χ^2 test

Lord's (1980) χ^2 test was proposed under the IRT framework, testing the difference in item parameters between groups. Generally, using the Lord's method to detect DIF items has three steps: first, estimate item parameters for the combined reference and focal groups, and standardize on the difficulty estimates; second, fix the pseudo-guessing parameters at the values obtained from the concurrent estimation, re-estimate the item discrimination and difficulty parameters for each group separately, and standardize again on the difficulty estimates; last, compare the discrimination and difficulty parameters obtained from the separate estimation using the χ^2 statistic demonstrated in Equation 2.18.

$$\chi^2 = \boldsymbol{v}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{v}_i \quad (2.18)$$

where

$$\boldsymbol{v_i} = \boldsymbol{v_{iR}} - \boldsymbol{v_{iF}^*} = \left(\hat{a}_{iR}, \hat{b}_{iR}, \hat{c}_{iR}\right)' - \left(\hat{a}_{iF}^*, \hat{b}_{iF}^*, \hat{c}_{iF}^*\right)' = \left(\hat{a}_{iR} - \hat{a}_{iF}^*, \hat{b}_{iR} - \hat{b}_{iF}^*, \hat{c}_{iR} - \hat{c}_{iF}^*\right)'$$

$$\boldsymbol{\Sigma_i} = \boldsymbol{\Sigma_{iR}} + \boldsymbol{\Sigma_{iF}^*}$$

$\hat{a}_{iR}, \hat{b}_{iR}, \hat{c}_{iR}$ represent the discrimination, difficulty, and pseudo-guessing parameters for item *i* correspondingly in the reference group, $\hat{a}_{iF}^* = \frac{\hat{a}_{iF}}{A}, \hat{b}_{iF}^* = A \times \hat{b}_{iF} + B, \hat{c}_{iF}^* = \hat{c}_{iF}$ represent the transformed item parameters in the focal group where *A* and *B* are the equating coefficients placing the focal group item parameters on the metric of the reference group (Kim, Cohen, & Kim, 1994), $\boldsymbol{\Sigma_{iR}}$ is the variance-covariance matrices of $\boldsymbol{v_{iR}}$, and $\boldsymbol{\Sigma_{jF}^*}$ is the transformed variance-covariance matrices of $\boldsymbol{v_{iF}^*}$.

Under the hypothesis that there is no difference between item parameters, the observed value of χ^2 follows a Chi-square distribution. For the 2PL model, χ^2 has two degrees of freedom, while for the 3PL model, χ^2 has three degrees of freedom. If the difference between item parametesr is significantly different from zero, the item is flagged as a DIF item.

2.2.3.5 The Likelihood Ratio Test

The likelihood ratio test compares the ratio of two nested models, with the hypothesis that the parameter estimates are invariant between groups. It computes the difference between the minus twice the log-likelihood of the constrained model and the free model, where the constrained model usually has fewer parameters than the free model. The test statistic of the likelihood ratio test can be calculated as:

$$G^2 = -2\,ln\left(\frac{L_{constrained}(\theta)}{L_{free}(\theta)}\right) = -2\,ln\,L_{constraied}(\theta) + 2\,ln\,L_{free}(\theta) \qquad (2.19)$$

Specifically, the parameters estimates for all test items are set equal for both reference and focal groups in the constrained model. However, in the free model, parameter estimates for all item except the studied items are constrained to be equal in the reference and focal groups (Thissen et al., 1998), that is, only item parameters of the studied items are separately estimated. Therefore, the metric used in the likelihood ratio test is based on all items except the studied items (Cohen, Kim, & Wollack, 1996). G^2 approximately follows a Chi-square distribution with degrees of freedom equal to the difference between the number of free parameters in two models.

2.2.3.6 Raju's Area Measures

Raju (1988; 1990) proposed two area measures—the exact signed area and the unsigned area between two IRFs for IRT models. For the 3PL model, let P_R and P_F represent the IRFs for item *i* in the reference and focal groups respectively:

$$P_R = c_{iR} + (1 - c_{iR}) \frac{exp[Da_{iR}(\theta - b_{iR})]}{1 + exp[Da_{iR}(\theta - b_{iR})]} \quad (2.20)$$

$$P_F = c_{iF} + (1 - c_{iF}) \frac{exp[Da_{iF}(\theta - b_{iF})]}{1 + exp[Da_{iF}(\theta - b_{iF})]} \quad (2.21)$$

The signed and unsigned areas between P_R and P_F are defined as follows:

$$SA = \int_{-\infty}^{+\infty} (P_R - P_F)\, d\theta \quad (2.22)$$

$$UA = \int_{-\infty}^{+\infty} |P_R - P_F|\, d\theta \quad (2.23)$$

However, since the area between two IRFs is infinite when the lower asymptotes are unequal for the 3PL model, Raju's method works only when $c_{iR} = c_{iF}$. For the 2PL

model, the signed and unsigned area measures for item *i* are calculated in a different way (Raju, 1988):

$$SA = b_{iF} - b_{iR} \tag{2.24}$$

$$UA = \begin{cases} |b_{iF} - b_{iR}|, & if\ a_{iF} = a_{iR} \\ |H_i|, if\ a_{iF} \neq a_{iR} \end{cases} \tag{2.25}$$

where

$$H_i = \frac{2(a_{iF} - a_{iR})}{D a_{iF} a_{iR}} \ln\left\{1 + \exp\left[\frac{D a_{iF} a_{iR}(b_{iF} - b_{iR})}{a_{iF} - a_{iR}}\right]\right\} - (b_{iF} - b_{iR})$$

In order to test whether the signed or unsigned area measure is significantly different from zero, $Z_i(SA)$ and $Z_i(UA)$ are defined as:

$$Z_i(SA) = \frac{b_{iF} - b_{iR}}{[var(b_{iF}) + var(b_{iR})]^{\frac{1}{2}}} \tag{2.26}$$

$$Z_i(UA) = \frac{H_i}{[var(H_i)]^{\frac{1}{2}}} \tag{2.27}$$

and these two test statistics both follow a standard normal distribution.

2.2.4 Anchor Selection and Scale Purification Procedures

A variety of methods was developed to select anchor items and build a common metric for DIF assessment. In addition to selecting the anchor items based on expert review, other common anchor selection methods include equal-mean-difficulty (EMD), all others as anchors (AOAA), and constant-item (CI). These methods do not involve any iterative or purification procedure, so they are easy and quick strategies for empirically selecting anchors.

Under the EMD approach, the mean item difficulties are constrained to be equal across groups. The EMD method is true under the assumption that either the test does not

contain any DIF items, or there exist several DIF items but the DIF amounts are same—some items favor one group, and other items favor another group (Wang, 2004; Wang et al., 2012). Some popular psychometric programs such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), ConQuest (Wu, Adams, Wilson, & Haldane, 2007), and PARSCALE (Muraki & Bock, 1997) employ the EMD method in DIF detection. However, this method works in very limited conditions, since previous research showed that if a test contains multiple DIF items, the EMD method functioned appropriately only when the difference in the mean item difficulties between the reference and focal groups approaches zero (Wang, 2004).

In terms of AOAA, all other items except the item currently tested for DIF are treated as anchors. Like the EMD approach, there is a prerequisite for this method—all items should be DIF-free or the studied item is the only DIF item in the test (Wang et al., 2012). It means that containing more than one DIF item is a great threat to AOAA. Moreover, each item tested for DIF has a different anchor, indicating that each item is tested with a slightly different common metric, which might be problematic in practice.

The CI approach selects a subset of items to establish a pre-determined common metric so that other items can be assessed for DIF. Previous studies found that the longer the anchor, the lower the type I error and the higher the power of detecting DIF items (e.g., Wang & Yeh, 2003; Wang, 2004). In practice, four anchor items are usually enough (Thissen et al., 1988). CI method yielded good control over type I error and achieved reasonable power when there were more than 30% of DIF items (Wang & Yeh, 2003).

Anchor selection approaches involving iterative procedures were also frequently used by researchers (e.g., Woods, 2009; Kopf, Zeileis, & Strobl, 2013). The iterative

scale purification procedure aims to remove DIF items from the common metric, and the major steps include: (1) calibrate item parameters separately for each group, link and place the reference and focal groups on a common metric; (2) assess each item for DIF using all items as anchors; (3) relink group metrics using only those items identified invariant in the previous step; (4) reassess each item for DIF; (5) repeat steps 3 and 4 until same items are identified invariant in consecutive iterations (Wang et al., 2012). Previous studies showed that the scale purification procedures made a huge improvement in DIF detection over EMD, AOAA, as well as CI methods (Lautenschlager et al., 1994).

2.3 Regularization Methods

2.3.1 Introduction

Assume that there exists a basic model

$$Y = f(X) + \varepsilon \tag{2.28}$$

where the response vector Y is given as a function f of the predictor vector X, with normally distributed errors ε having a mean of zero and variance of σ_ε^2. An estimate $\hat{f}$ of the underlying relationship f can be obtained using different modeling techniques such as linear regression or logistic regression, and the predictions made by $\hat{f}$ at particular values of X should approximate the true values given by Y as well as possible. In this case, the expected squared prediction error at X can be defined as $E\left[\left(Y - \hat{f}(X)\right)^2\right]$, which can be further decomposed into bias and variance components as shown in Equation 2.29.

$$\begin{aligned} E\left[\left(Y - \hat{f}(X)\right)^2\right] &= \left(E\left[\hat{f}(X)\right] - f(X)\right)^2 + E\left[\left(\hat{f}(X) - E\left[\hat{f}(X)\right]\right)^2\right] + \sigma_\varepsilon^2 \\ &= \text{Bias}^2 + \text{Variance} + \text{Irreducible Error} \end{aligned} \tag{2.29}$$

Equation 2.29 is usually called the bias-variance composition. Bias and variance are two sources of error that prevent a model from effective generalization—bias is an error introduced by approximating a real-life problem, and variance refers to the amount of error by which the predicted values may change if the model is estimated using a different dataset (James, Witten, Hastie, & Tibshirani, 2013). The third term is the noise term that cannot be reduced by any model. There is a tradeoff between bias and variance that comes with model complexity. Generally, complex models have high variance and low bias, while simple models have high bias and low variance. In this study, both analytic methods and simulation studies were used to find a good balance between bias and variance so that the final model minimizes the prediction errors and yields the optimal operating characteristics in terms of hit rates and false alarm rates.

The bias-variance decomposition forms the conceptual foundation of regularization. Regularization is a technique that helps solve overfitting problem by explicitly controlling for model complexity. For example, in linear regression, the ordinary least squares (OLS) solution gives the best linear unbiased estimator, but it often has poor prediction and generalization since it relies too much on the training data. Regularization has the benefit of reducing variance and improving predictive accuracy compared to OLS, by introducing bias into the regression model. It identifies the preferred level of model complexity by adding a penalty term to the objective function to be minimized. The most common penalties are the $\ell 1$ and $\ell 2$ penalties, where the $\ell 1$ penalty equals the absolute value of the magnitude of coefficients, and the $\ell 2$ penalty equals the square of the magnitude of coefficients.

2.3.2 Different Types of Penalties

Assume that there is a dataset consisting of *N* observations and *J* variables; *n* is used to index the observations from 1 to *N* ($n = 1, \dots, N$), and *j* is used to index the predictor variables from 1 to *J* ($j = 1, \dots, J$). Let x_{nj} represent the value of the *j*th variable for the *n*th observation and y_n represent the outcome variable for the *n*th observation, therefore $\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \dots, (\boldsymbol{x_N}, y_N)\}$ represent the observed data where $\boldsymbol{x_n} = (x_{n1}, x_{n2}, \dots, x_{nJ})$ is a vector of length *J*. A linear regression model can be expressed as:

$$y_n = \beta_0 + \sum_{j=1}^{J} \beta_j x_{nj} + \varepsilon \tag{2.30}$$

where $\beta_0, \beta_1, \dots, \beta_J$ are known as model coefficients. The OLS procedure obtains estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_J$ for model coefficients, such that the loss function—the sum of squared errors defined in Equation 2.31 is minimized:

$$SSE_{OLS} = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \tag{2.31}$$

where $\hat{y}_n = \hat{\beta}_0 + \sum_{j=1}^{J} \hat{\beta}_j x_{nj}$ is the prediction for y_n based on the observed values of $x_{n1}, x_{n2}, \dots, x_{nJ}$.

The lasso regression (Tibshirani, 1996) adds an $\ell 1$ penalty term to the loss function in Equation 2.31:

$$SSE_{lasso} = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 + \lambda \sum_{j=1}^{J} |\beta_j| = SSE_{OLS} + \lambda \sum_{j=1}^{J} |\beta_j| \tag{2.32}$$

where $\lambda \geq 0$ is a tuning parameter, controlling the strength of the $\ell 1$ penalty term. When $\lambda = 0$, the OLS estimates are obtained since Equation 2.31 and Equation 2.32 are identical. A large value of λ shrinks many model's coefficient estimates towards zero and returns a very sparse solution, meaning that $\hat{\beta}_j = 0$ for most j ($j = 1, \dots, J$). It is because imposing a lasso penalty corresponds to assuming a Laplace prior on model coefficients, which expects many zero-valued coefficients, and a small subset to be larger and nonzero (Friedman, Hastie, & Tibshirani, 2010).

The ridge regression (Hoerl & Kennard, 1970) uses an $\ell 2$ penalty. Therefore,

$$SSE_{ridge} = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 = SSE_{OLS} + \lambda \sum_{j=1}^{J} \beta_j^2 \qquad (2.33)$$

Using an $\ell 2$ penalty also shrinks the estimated coefficients but never sets them to zero exactly. Again when $\lambda = 0$, the OLS estimates are obtained. The ridge penalty is ideal when there are many correlated predictors because ridge regression tends to shrink coefficients of correlated predictors towards each other, leading to small but nonzero values. It is different from lasso regression which tends to pick one of correlated variables and ignore the rest. Lasso regression performs shrinkage and variable selection simultaneously and produces simpler and more interpretable models by setting a subset of model coefficients equal to zero, which is a major advantage over ridge regression.

Further, the elastic net (Zou & Hastie, 2005) combines the $\ell 1$ and $\ell 2$ penalties:

$$\begin{aligned} SSE_{elastic\ net} &= \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 + \lambda \left[\alpha \sum_{j=1}^{J} |\beta_j| + (1-\alpha) \sum_{j=1}^{J} \beta_j^2 \right] \\ &= SSE_{OLS} + \lambda \left[\alpha \sum_{j=1}^{J} |\beta_j| + (1-\alpha) \sum_{j=1}^{J} \beta_j^2 \right] \end{aligned} \qquad (2.34)$$

In Equation 2.34, λ is a tuning parameter serving the same purpose that some of the coefficients should be set to zero exactly, and α is another tuning parameter that combines the $\ell 1$ and $\ell 2$ penalty terms together. When α equals 0 or 1, the ridge or lasso regression are obtained correspondingly. The elastic net regression not only results in sparse solutions and performs variable selection like the lasso regression, but also takes the advantage of the ridge regression which performs well with highly correlated variables. However, the elastic net is computationally expensive since a grid search is required to determine the appropriate values of α and λ.

Under the linear regression setting, least squares estimation is a special case of maximum likelihood. Maximum likelihood estimation (MLE) is a more general and flexible approach, which can be used to fit many linear and nonlinear models (James et al., 2013). As with the case of the penalized sum of squared errors in Equations 2.32-2.34, a penalty term can be added to the maximum likelihood function. Therefore, rather than maximizing the log-likelihood function, a penalized version of log-likelihood is maximized. For example, when estimating a logistic regression model:

$$logit[P(y_n = 1)] = ln\left[\frac{P(y_n = 1)}{1 - P(y_n = 1)}\right] = \beta_0 + \sum_{j=1}^{J} \beta_j x_{nj} \qquad (2.35)$$

where $y_n (i = 1, \dots, N)$ is a binary variable with two categories 0 and 1, x_{nj} $(n = 1, \dots, N; j = 1, \dots, J)$ represents the value of the *j*th variable for the *n*th observation, and $\beta_0, \beta_1, \dots, \beta_J$ are the $J + 1$ parameters to be estimated. The likelihood function is:

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{n=1}^{N} \pi(\boldsymbol{x_n}; \beta_0, \boldsymbol{\beta})^{y_n} [1 - \pi(\boldsymbol{x_n}; \beta_0, \boldsymbol{\beta})]^{1-y_n} \qquad (2.36)$$

where $\pi = P(y_n = 1)$ indicates the probability of $y_n = 1$, $\boldsymbol{x_n} = (x_{n1}, x_{n2}, \dots, x_{nJ})$ represents the observations of J predictors corresponding to y_n, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)$. The log-likelihood function is then defined as the natural log of Equation 2.36:

$$l(\beta_0, \boldsymbol{\beta}) = ln[L(\beta_0, \boldsymbol{\beta})] = \sum_{n=1}^{N} (y_n \, ln \, \pi(\boldsymbol{x_n}; \beta_0, \boldsymbol{\beta}) + (1 - y_n) \, ln[1 - \pi(\boldsymbol{x_n}; \beta_0, \boldsymbol{\beta})])$$
$$= \sum_{n=1}^{N} \left\{ y_n \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{nj} \right) - ln \left[1 + exp \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{nj} \right) \right] \right\} \quad (2.37)$$

The loss function for logistic regression is defined as the negative log-likelihood:

$$-l(\beta_0, \boldsymbol{\beta}) = -\sum_{n=1}^{N} \left\{ y_n \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{nj} \right) - ln \left[1 + exp \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{nj} \right) \right] \right\} \quad (2.38)$$

Therefore, the estimates of model parameters in a logistic regression model can be obtained by minimizing the penalized loss function in Equation 2.39 or maximizing the penalized likelihood in Equation 2.40:

$$-l(\beta_0, \boldsymbol{\beta}) + \lambda \cdot \eta(\boldsymbol{\beta}) \quad (2.39)$$

$$l(\beta_0, \boldsymbol{\beta}) - \lambda \cdot \eta(\boldsymbol{\beta}) \quad (2.40)$$

where $\eta(\boldsymbol{\beta}) = \alpha \sum_{j=1}^{J} |\beta_j| + (1 - \alpha) \sum_{j=1}^{J} \beta_j^2$ refers to the lasso ($\alpha = 1$), the ridge ($\alpha = 0$) or the elastic net penalty ($0 < \alpha < 1$), and α and λ are two tuning parameters.

When applying the regularization methods to DIF detection, typically a set of variables containing item and person characteristics as well as parameters indicating DIF effects are included in the model. The identification of DIF items can be realized by examining whether the DIF parameter estimates equal to zero or not. Since the ridge penalty is not able to shrink the estimated coefficients to exactly zero, it does not have the ability to differentiate DIF items and non-DIF items. Therefore, the ridge penalty was not

considered in this study. In terms of other penalties, Knight and Fu (2000) have shown that for fixed J and $\boldsymbol{\beta}$, the lasso solution is asymptotically consistent and asymptotically normal as $N \to \infty$. Moreover, as a generalization of the traditional lasso, the elastic net enjoys similar asymptotic properties as the lasso penalty (De Mol, De Vito, & Rosasco, 2009), but outperforms lasso when J is much bigger than N (Zou & Hastie, 2005).

From the perspective of model selection consistency, according to previous studies, in a standard lasso (e.g., Meinshausen & Bühlmann, 2006; Zou, 2006) or elastic net procedure (e.g., Yuan & Lin, 2007; Jia & Yu, 2010), there are cases where a given value of tuning parameter leading to optimal estimation accuracy ends up with inconsistent selection of variables. The adaptive lasso (Zou, 2006) was developed to address the limitations by introducing weights to the penalty on each coefficient in the traditional lasso procedure, which can be solved using the same algorithm for solving the lasso:

$$\eta(\boldsymbol{\beta})_{adaptive\ lasso} = \sum_{j=1}^{J} \omega_j |\beta_j|, \qquad \omega_j = \frac{1}{|\hat{\beta}_j^{ini}|^{\gamma}} \tag{2.41}$$

In Equation 2.41, $\hat{\beta}_j^{ini}$ is an initial estimate of the coefficients β_j, usually obtained from ridge regression, and $\gamma > 0$ defines the weighted vector ω_j for $|\beta_j|$. The adaptive lasso procedure yields consistent estimates because large model coefficients are penalized less than small coefficients. Besides, the adaptive lasso retains the same advantages of the lasso penalty which shrinks some of the coefficients to exactly zero, thus performs a selection of attributes with regularization.

2.3.3 Model Selection Techniques

An important issue remains unsolved is "which tuning parameter value is optimal", since given different tuning parameter values, a set of candidate models will be obtained. Often the goal of model selection is to choose a model for future prediction, and it is natural to measure the predictive accuracy by estimating the prediction error (Barbieri & Berger, 2004).

Particularly in this study, the tuning parameter also determines the number of items identified as DIF or non-DIF items. Further, it determines if all DIF items are correctly detected and if there are non-DIF items that are falsely identified as DIF items. Obviously, when the tuning parameter is too small, it is difficult to obtain zero estimates for model parameters. In this case, not only DIF items will show DIF effects, but also many non-DIF items will show DIF effects due to nonzero estimates of the corresponding DIF parameters. To conclude, a too small tuning parameter value will increase the possibility of misidentifying non-DIF items as DIF items. On the other hand, a too large tuning parameter value sets too many DIF parameters to zero, thus too few items will be flagged as DIF items.

The selected tuning parameter value needs to balance both model fit as indicated by minimizing the prediction error and model complexity as indicated by the number of zero and nonzero DIF parameter estimates. Therefore, a method is required to determine which tuning parameter value yields a good model fit as well as an optimal differentiation between DIF items and non-DIF items.

2.3.3.1 Cross-Validation

Cross-validation (CV) is a popular technique for evaluating a model's performance in order to tackle overfitting problems and assess how the model can generalize to a new dataset. CV involves randomly dividing a data sample into two complementary subsets, performing the proposed analysis on one subset which is usually called the training data, and validating the analysis on the other subset which is usually called the testing data. However, the validation results may be very different depending on how the dataset is divided.

In order to reduce variability, CV is performed multiple times using different divisions, and the validation results are averaged to give an estimate of the model's predictive performance. For example, the *K*-fold CV divides the data sample into *K* subsets $C_k(k = 1, \dots, K)$ of approximately equal size, fits the model on *K*-1 subsets by removing one of the subsets, and computes the prediction error of the fitted model using the observations in the left-out subset. This procedure is repeated *K* times, and each subset is used once as the left-out testing set, resulting in *K* estimates of prediction error. The average of *K* estimates is calculated as a final measure of model performance. Typical choices of *K* are 5 and 10.

Let $N_k(k = 1, \dots, K)$ represent the number of observations in the subset C_k. For a linear regression model, the mean squared error is used to estimate the prediction error, and the *K*-fold CV estimate of prediction error can be calculated as follows:

$$CV_{linear} = \sum_{k=1}^{K} \frac{N_k}{N} \sum_{n \in C_k} \frac{\left\{ y_n - \left(\hat{\beta}_0^{(-k)} + \sum_{j=1}^{J} \hat{\beta}_j^{(-k)} x_{nj} \right) \right\}^2}{N_k} \qquad (2.42)$$

For a logistic regression model, the deviance defined as minus twice the log-likelihood on the left-out data can be calculated to estimate the prediction error (Friedman et al., 2010). Based on Equation 2.37, the *K*-fold CV estimate of prediction error can be computed as follows:

$$CV_{logistic} = -2\sum_{k=1}^{K}\frac{N_k}{N}\sum_{n\in C_k}\left\{y_n\left(\hat{\beta}_0^{(-k)} + \sum_{j=1}^{J}\hat{\beta}_j^{(-k)}x_{nj}\right) - ln\left[1 + exp\left(\hat{\beta}_0^{(-k)} + \sum_{j=1}^{J}\hat{\beta}_j^{(-k)}x_{nj}\right)\right]\right\} \quad (2.43)$$

In Equations 2.42 and 2.43, $\hat{\beta}_0^{(-k)}, \hat{\beta}_1^{(-k)}, \ldots, \hat{\beta}_J^{(-k)}$ are the estimates based on the specified training data only (e.g., *K-1* subsets except C_k). The best model is the one with the minimum CV estimate, indicating that the best choice of tuning parameter is the one that minimizes the prediction error.

2.3.3.2 Information Criteria

Although CV is useful to compute the out-of-sample prediction error directly, the process is usually tedious since repeated model fits are required, and it is less practical when the data is sparse. Information criterion is much simpler in terms of computation and is used as an alternative model selection technique in practice. An information criterion estimates the out-of-sample prediction error by adjusting the error based on the training data in order to account for the bias due to overfitting (James et al., 2013). Therefore, all observed data are used to train the model, and different techniques are applied to adjust the training error.

This study considered two commonly used information criteria including the Akaike information criterion (AIC; Akaike, 1974), and the Bayesian information criterion (BIC; Schwarz, 1978). The AIC is defined as:

$$AIC = -2l_m + 2m \quad (2.44)$$

where m represents the model size which equals the number of free parameters in the model, l_m is the log-likelihood of the model evaluated at the maximum likelihood estimates. Particularly, the term $-2l_m$ is a measure of lack of model fit, and the term $2m$ can be interpreted as a penalty for increasing the size of the model so as to enforce parsimony in the number of parameters. Therefore, the optimal model selected by AIC is the one with the minimum AIC value. Akaike's approach achieves an important objective called the asymptotic efficiency (Shibata, 1976). Asymptotic efficiency essentially minimizes the prediction error and maximizes the predictive accuracy, meaning that AIC is able to find the best approximating model (Aho, Derryberry, & Peterson, 2014). Additionally, Stone (1977) proved that the asymptotic equivalence of choosing model by CV and AIC when MLE was used within each model, indicating that CV and AIC have similar performance in model selection.

BIC selects a model that minimizes

$$BIC = -2l_m + m \log N \tag{2.45}$$

where *N* corresponds to sample size. Although closely related to AIC, BIC was derived in a fully Bayesian framework, aiming to select a model that maximizes the posterior model probability. The BIC procedure has a well-known property of consistency, meaning that the probability of selecting the true model converges to 1 when the sample size increases (Yang, 2005). Moreover, previous research indicated that when *n* became larger, BIC tended to favor models of lower dimension than those chosen by AIC (Koehler & Murphree, 1988).

AIC, BIC, and *K*-fold CV usually select different tuning parameter values, so that different models are obtained yielding different DIF detection results. Several operating

characteristics such as hit rates and false alarm rates are calculated based on these models, and the model yielding the best operating characteristics (e.g., high hit rates and well-controlled false alarm rates) is finally selected.

2.3.4 Applications in DIF Detection

Magis et al. (2015) proposed a lasso DIF method based on logistic regression, allowing for simultaneous detection of uniform DIF for all test items using a single model:

$$logit\left[P\left(y_{ip}=1\right)\right]=\beta_{0i}+\beta_1 s_p+\beta_{2i}G_p \tag{2.46}$$

In Equation 2.46, y_{ip} is the response of person p ($p=1,\dots,P$) to item i ($i=1,\dots,I$), β_{0i} ($i=1,\dots,I$) represents item difficulty, β_1 represents the effect for test score s_p and is constrained to be identical across items, β_{2i} ($i=1,\dots,I$) are the DIF parameters representing uniform DIF effect under the setting of the Rasch models, and G_p is the group membership indicator where $G_p=1$ when person p is in the reference group and $G_p=0$ otherwise. The vector $\boldsymbol{\beta}=(\beta_{01},\dots,\beta_{0I},\beta_1,\beta_{21},\dots,\beta_{2I})$ collects all model parameters, and the corresponding parameter estimates $\widehat{\boldsymbol{\beta}}(\lambda)$ can be obtained by maximizing the penalized log-likelihood with an $\ell 1$ penalty on parameters $\beta_{21},\dots,\beta_{2I}$:

$$\widehat{\boldsymbol{\beta}}(\lambda)=arg\,max\left\{\boldsymbol{l}(\boldsymbol{\beta})-\lambda\sum_{i=1}^{I}|\beta_{2i}|\right\} \tag{2.47}$$

A higher value of λ shrinks more DIF parameters towards zero. In their study, λ was selected by comparing several methods including *K*-fold CV ($K=3,5,10$), AIC, BIC, and another weighted information criterion which is a combination of AIC and BIC. The main advantage of this method is its flexibility as the assumption of an invariant anchor is

no longer required, and the multiple testing issue is also addressed. However, the authors used test scores as a proxy for ability, which might be less accurate compared to using the estimated person ability from IRT modeling. It is because from the IRT point of view, the individuals with a same test score may have different levels of ability if those items vary in their discriminations and difficulties.

Tutz and Schauberger (2015) proposed a similar penalization approach by using the examinees' estimated latent proficiency level $\hat{\theta}_p$ instead of the test score s_p in Equation 2.46. The DIF model then becomes:

$$logit\left[P\left(y_{ip}=1\right)\right]=\beta_{0i}+\beta_1\hat{\theta}_p+\beta_{2i}G_p \qquad (2.48)$$

Again, penalized MLE was used for parameter estimation, and BIC was used to evaluate model performance.

To conclude, the purpose of DIF analysis is to determine whether an item displays DIF or not. As discussed in Section 2.2.3, in a traditional item-level DIF analysis, an item is usually assessed through a hypothesis test. Hypothesis testing problems can be examined from the model selection perspective—the acceptance or rejection of the null hypothesis (e.g., no DIF) is considered in terms of the selection of a more appropriate model when one is nested inside the other (Cubedoo & Oller, 2002). In regularization methods, DIF parameters for all items are included in a single model; whether an item displays DIF or not is indicated by whether the corresponding DIF parameter estimate is zero or not. The regression coefficients for DIF parameters are determined by the selected tuning parameter values, which are chosen such that the model has the best prediction for a future dataset.

2.4 Summary

This chapter started with an introduction to IRT, including common IRT models and parameter estimation techniques, which are important components of the research design in this study. IRT examines the relationship between the probability of correct response and the level of latent ability both mathematically and graphically using IRT models and ICCs respectively, which is a useful tool for assessing DIF.

Section 2.2 described how to identify different types of DIF according to different patterns of ICCs and item parameter estimates between groups. Additionally, various DIF detection techniques were discussed in this Section. The MH method, logistic regression, and SIBTEST are popular non-IRT approaches, while Lord's χ^2 test, the likelihood ratio, and Raju's area measures are commonly used IRT-based DIF methods. These methods are all conducted at the item level, focusing on analyzing each item individually. As mentioned in Chapter 1, there are several limitations of the item-level approaches, such as the strict assumption of an invariant anchor, and the issue caused by multiple testing. Although these limitations can be partially addressed by applying certain anchor selection and scale purification procedures, simultaneous investigation of all item on a test for DIF seems to be a better solution to address these limitations. Simultaneous DIF detection can be realized using either the regularization methods for estimation or the generalized linear mixed models (e.g., Kamata, 2001; Swanson, Clauser, Case, Nungester, & Featherman, 2002; Van den Noorgate & De Boeck, 2005; Binici, 2007; Acar, 2012), and this study focused on the regularization methods only. To conclude, Figure 2.4 summarized all aforementioned DIF detection methods.

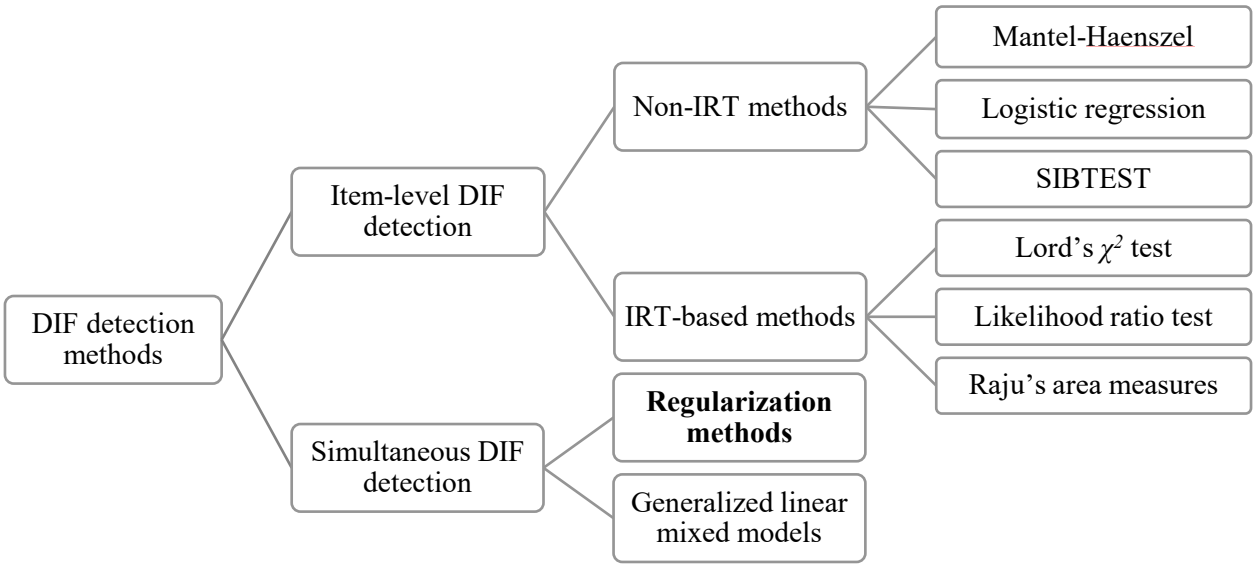

Figure 2.4 Summary of DIF Detection Methods

Additionally, Section 2.3 reviewed the basic concepts of regularization methods including different penalty functions, model selection techniques, as well as current applications of regularization methods in DIF detection.

Next, in Chapter 3, a new DIF detection model using regularization techniques will be proposed. Further, comprehensive simulation studies will be designed to evaluate the performance of the proposed model. Empirical datasets will also be used to assess the model's applicability in real settings.

CHAPTER 3. RESEARCH DESIGN

3.1 A General Framework for DIF Detection

3.1.1 DIF Detection Model

A DIF detection model for simultaneous detection of both uniform and nonuniform DIF effects is defined as follows:

$$logit(\pi_{ip}) = logit[P(y_{ip} = 1)] = \beta_{0i} + \beta_{1i}\theta_p + \beta_{2i}G_p + \beta_{3i}\theta_p G_p \quad (3.1)$$

y_{ip} is the binary response of person p $(p = 1, \dots, P)$ on item i $(i = 1, \dots, I)$;

$\pi_{ip} = P(y_{ip} = 1)$ is the probability of $y_{ip} = 1$ and $y_{ip} \sim \text{Bernoulli}(\pi_{ip})$;

β_{0i} $(i = 1, \dots, I)$ is the intercept for item i, which can be referred to as the counterpart of item difficulty in the IRT framework;

θ_p is the latent ability for person p $(p = 1, \dots, P)$;

β_{1i} represents the effect for ability which can also be considered as the counterpart of item discrimination in the IRT framework;

β_{2i} is the DIF parameter representing uniform DIF effect for item i;

G_p is the group membership indicator where $G_p = 1$ when person p is in the reference group and $G_p = 0$ otherwise assuming there exist two mutually exclusive groups;

β_{3i} is the DIF parameter representing nonuniform DIF effect for item i; and $\theta_p G_p$ is the product of two variables θ_p and G_p.

Item i shows uniform DIF when $\hat{\beta}_{2i} \neq 0$ and $\hat{\beta}_{3i} = 0$, and item i shows nonuniform DIF when $\hat{\beta}_{3i} \neq 0$ regardless of the value of β_{2i}.

Equation 3.1 can be re-expressed as follows:

$$logit(\pi_{ip}) = \boldsymbol{\beta_0 I_i} + \boldsymbol{\beta_1 I_i} \cdot \theta_p + \boldsymbol{\beta_2 I_i} \cdot G_p + \boldsymbol{\beta_3 I_i} \cdot \theta_p G_p = \boldsymbol{\beta x_p} \quad (3.2)$$

$\boldsymbol{\beta_0} = (\beta_{01}, \dots, \beta_{0I}), \boldsymbol{\beta_1} = (\beta_{11}, \dots, \beta_{1I}), \boldsymbol{\beta_2} = (\beta_{21}, \dots, \beta_{2I}), \boldsymbol{\beta_3} = (\beta_{31}, \dots, \beta_{3I})$, and $\boldsymbol{I_i} =$ (0,0, … ,1, … ,0,0) with 1 at position i and 0 otherwise;

$\boldsymbol{\beta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}) = (\beta_{01}, \dots, \beta_{0I}, \beta_{11}, \dots, \beta_{1I}, \beta_{21}, \dots, \beta_{2I}, \beta_{31}, \dots, \beta_{3I})$ represents the vector of parameters that need to be estimated;

$\boldsymbol{x_p} = (1, \dots, 1; \theta_p, \dots, \theta_p; G_p, \dots, G_p; \theta_p G_p, \dots, \theta_p G_p)^{\boldsymbol{T}}$ is a $4I \times 1$ vector including four unique elements $1, \theta_p, G_p, \theta_p G_p$, where each element is repeated I times successively in the vector $\boldsymbol{x_p}$.

One question remained unsolved in this model is the choice of θ_p. As stated in Section 2.3.4, both the observed total score and the IRT ability estimate were used as a proxy for θ_p in previous research (e.g., Magis et al., 2015; Tutz & Schauberger, 2015). However, no study has compared their performance in DIF detection yet. Therefore, a preliminary study is recommended to determine which of the two ability measures has a better performance in DIF detection by comparing their operating characteristics. The results will be presented in Section 3.1.4.2.

3.1.2 Parameter Estimation

The parameter estimates $\widehat{\boldsymbol{\beta}}$ can be found by maximizing Equation 2.40 so that

$$\widehat{\boldsymbol{\beta}} = arg\,max\{l(\boldsymbol{\beta}) - \lambda \cdot \eta(\boldsymbol{\beta})\} \quad (3.3)$$

In Equation 3.3, $l(\boldsymbol{\beta})$ is the log-likelihood of the model, which can be expressed as:

$$l(\boldsymbol{\beta}) = ln\left\{\prod_{p=1}^{P}\prod_{i=1}^{I}\pi(\boldsymbol{x_p};\boldsymbol{\beta})^{y_{ip}}\left[1-\pi(\boldsymbol{x_p};\boldsymbol{\beta})\right]^{1-y_{ip}}\right\}$$
$$= \sum_{p=1}^{P}\sum_{i=1}^{I} y_{ip}\, ln\,\pi(\boldsymbol{x_p};\boldsymbol{\beta}) + (1-y_{ip})\, ln\left[1-\pi(\boldsymbol{x_p};\boldsymbol{\beta})\right]$$
$$= \sum_{p=1}^{P}\sum_{i=1}^{I}\left\{y_{ip}(\boldsymbol{\beta x_p}) - ln\left[1+exp(\boldsymbol{\beta x_p})\right]\right\} \quad (3.4)$$

$\eta(\boldsymbol{\beta})$ is a penalty function that penalizes specific structures in the parameter vector $\boldsymbol{\beta}$, and λ ($\lambda \geq 0$) is the tuning parameter. The traditional maximum likelihood estimates are obtained when $\lambda = 0$; a large value of λ shrinks many model coefficients towards zero, and the nonzero estimates indicate potential DIF effects.

Three types of penalty functions were considered in this study. First, based on Equation 3.1, the lasso penalty can be defined as:

$$\eta(\boldsymbol{\beta})_{lasso} = \sum_{h=2}^{3}\sum_{i=1}^{I}|\beta_{hi}| \quad (3.5)$$

In Equation 3.5, only DIF parameters $(\boldsymbol{\beta_2}, \boldsymbol{\beta_3}) = (\beta_{21}, \dots, \beta_{2I}, \beta_{31}, \dots, \beta_{3I})$ are penalized, while other model parameters $(\boldsymbol{\beta_0}, \boldsymbol{\beta_1}) = (\beta_{01}, \dots, \beta_{0I}, \beta_{11}, \dots, \beta_{1I})$ are unaffected in parameter estimation.

The second type of penalty function is the elastic net:

$$\eta(\boldsymbol{\beta})_{elastic\ net} = \alpha\sum_{h=2}^{3}\sum_{i=1}^{I}|\beta_{hi}| + (1-\alpha)\sum_{h=2}^{3}\sum_{i=1}^{I}{\beta_{hi}}^2 \quad (3.6)$$

where α ($0 \leq \alpha \leq 1$) is another tuning parameter in addition to λ, and only DIF parameters $(\boldsymbol{\beta_2}, \boldsymbol{\beta_3})$ are penalized. As discussed in Section 2.3.2, the lasso and ridge penalties are special cases of the elastic net by constraining α to 1 and 0 respectively. However, the ridge penalty is not applicable in DIF detection, so $\alpha = 0$ was not

considered. Also, $\alpha = 1$ was not considered either in order to avoid duplicate analysis results. Therefore, when performing a grid search over the tuning parameter α, the parameter range was set to $0 < \alpha < 1$. And in this study, α was selected from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$.

Lastly, the adaptive lasso penalty was also studied:

$$\eta(\boldsymbol{\beta})_{adaptive\ lasso} = \sum_{h=2}^{3}\sum_{i=1}^{I} \omega_{hi}|\beta_{hi}|\,, \omega_{hi} = \frac{1}{\left|\hat{\beta}_{hi}^{ini}\right|^{\gamma}} \tag{3.7}$$

where $\hat{\beta}_{hi}^{ini}$ is an initial estimate for β_{hi} obtained from ridge regression estimates, and $\gamma > 0$ defines the weighted vector ω_{hi} for $|\beta_{hi}|$. In this study, γ was selected from $\{0.5, 1, 2\}$ according to Zou's (2006) research. Again, only DIF parameters $(\boldsymbol{\beta_2}, \boldsymbol{\beta_3})$ were penalized.

In sum, there is one tuning parameter using the lasso penalty, and two tuning parameters using the elastic net or the adaptive lasso. Model selection techniques are applied to find the optimal tuning parameter λ^* for the lasso method, or the optimal tuning parameter pairs $\{\alpha^*, \lambda^*\}$ for the elastic net and $\{\gamma^*, \lambda^*\}$ for the adaptive lasso.

3.1.3 Model Selection

Section 2.3.3 briefly introduced several model selection techniques including AIC, BIC, and *K*-fold CV which can be used to determine the tuning parameter estimate(s) and the corresponding DIF detection model. Specifically,

$$AIC = -2l\left(\widehat{\boldsymbol{\beta}}\right) + 2 \cdot \widehat{df} \tag{3.8}$$

$$BIC = -2l\left(\widehat{\boldsymbol{\beta}}\right) + log(P \cdot I) \cdot \widehat{df} \tag{3.9}$$

where $l(\widehat{\boldsymbol{\beta}})$ represents the log-likelihood of the parameter vector $\widehat{\boldsymbol{\beta}}$ which can be calculated according to Equation 3.4, *P* is the total number of participants, *I* is the total number of items, and $\widehat{df}$ is the total degrees of freedom equal to the number of nonzero coefficients in the model. Given a set of candidate models, the preferred model is the one with the minimum AIC or BIC estimate.

In terms of *K*-fold CV, the data sample is divided into K folds $C_k(k = 1, \dots, K)$ of approximately equal size. Assume that each fold includes $N_k(k = 1, \dots, K)$ participants, therefore $\sum_{k=1}^{K} N_k = P$. The split of data is made across participants and groups, so that each fold includes participants from both reference and focal groups. According to Equation 2.43, the preferred model is the one with the minimum CV estimate, which can be computed as:

$$CV = -2 \cdot \sum_{k=1}^{K} \frac{N_k}{P} \sum_{p \in C_k} \sum_{i=1}^{I} \{y_{ip}(\widehat{\boldsymbol{\beta}}\boldsymbol{x_p}) - ln[1 + exp(\widehat{\boldsymbol{\beta}}\boldsymbol{x_p})]\} \qquad (3.10)$$

As discussed in Section 2.3.3, different model selection methods may yield different DIF detection results since they usually select different tuning parameter values. However, it is preferable to use one model selection technique consistently throughout the simulation study so that the analysis results can be compared properly. Therefore, a preliminary analysis will be conducted beforehand in order to compare several model selection methods in terms of their DIF detection performance. The four model selection methods considered in this study include AIC, BIC, 5-fold CV, and 10-fold CV. The results will be presented in Section 3.1.4.1.

3.1.4 Preliminary Analyses

3.1.4.1 Comparison of Model Selection Techniques

The purpose of the first preliminary analysis is to determine which model selection technique selects the optimal tuning parameter in terms of DIF detection performance. A DIF detection model derived from Equation 3.1 assessing uniform DIF only was used in this preliminary study, which not only simplifies the analysis procedure but also achieves aforementioned research purpose. The model can be written as:

$$logit(\pi_{ip}) = \beta_{0i} + \beta_{1i}\theta_p + \beta_{2i}G_p \tag{3.11}$$

The IRT ability was used as a proxy for θ_p, estimated using EAP under the 2PL model (see Equation 2.4). The 2PL model can be re-expressed as follows:

$$P(y_{ip} = 1|\theta_p, a_i, b_i) = \frac{exp[a_i(\theta_p - b_i)]}{1 + exp[a_i(\theta_p - b_i)]}$$
$$= \frac{exp[-(a_i\theta_p + d_i)]}{1 + exp[-(a_i\theta_p + d_i)]} = P(y_{ip} = 1|\theta_p, a_i, d_i) \tag{3.12}$$

where d_i is a intercept parameter. Specifically in this study, a_i $(i = 1, \dots, 20)$ were generated following a log-normal distribution with a mean of 0 and a standard deviation of 0.5; d_i $(i = 1, \dots, 20)$ were generated from a standard normal distribution of $N(0, 1)$.[2] These generating distributions were chosen because previous research has shown that they are able to give a reasonable approximation to observed empirical distributions of item parameter estimates (e.g., Du Toit, 2003; Houts & Cai, 2016). Table 3.1 summarizes the generated item discrimination and intercept parameters for 40 items.

[2] The mirt R package (Chalmers, 2012) was used for all IRT analyses in this study. The default model in the mirt R package uses $-(a_i\theta_p + d_i)$ instead of $a_i(\theta_p - b_i)$ in the IRT model. In order to make the estimation procedures more consistent, the intercept parameters d_i rather than the difficulty parameters b_i were generated in all simulation studies. The difficulty parameters b_i was then computed by $b_i = -d_i/a_i$.

Table 3.1 Generated Item Parameters

No.	a_i	d_i	No.	a_i	d_i
Item 01	0.809	-0.918	**Item 21**	0.505	-1.228
Item 02	1.037	-1.466	**Item 22**	0.975	0.185
Item 03	0.802	-0.938	**Item 23**	1.276	-1.311
Item 04	0.490	-0.917	**Item 24**	0.923	-0.415
Item 05	0.655	-1.103	**Item 25**	1.109	-1.148
Item 06	0.256	-1.392	**Item 26**	1.107	-0.698
Item 07	1.006	-1.050	**Item 27**	1.642	0.094
Item 08	0.521	-0.909	**Item 28**	1.339	-0.483
Item 09	1.192	-1.214	**Item 29**	0.881	0.716
Item 10	0.983	-2.131	**Item 30**	0.952	0.037
Item 11	1.531	-0.826	**Item 31**	0.874	0.263
Item 12	1.011	0.423	**Item 32**	0.821	0.234
Item 13	0.920	-0.973	**Item 33**	1.130	-0.637
Item 14	0.634	0.981	**Item 34**	1.719	-0.218
Item 15	1.423	-0.515	**Item 35**	1.133	0.082
Item 16	0.811	-0.376	**Item 36**	1.168	-1.051
Item 17	0.755	-0.412	**Item 37**	0.785	0.733
Item 18	2.140	-0.763	**Item 38**	1.042	-1.345
Item 19	1.061	0.397	**Item 39**	0.595	-0.060
Item 20	0.992	-1.748	**Item 40**	1.444	-1.147

Two test lengths were examined including 20 items and 40 items. For a 20-item test, item parameters of Items 01-20 in Table 3.1 were used to simulate item responses, while for a 40-item test, item parameters of all 40 items were used. Additionally, three sample sizes of 1000, 2000, and 4000 were considered, and the reference and focal groups were assumed to have the same number of examinees. The latent ability distribution was set to $N(0,1)$ for both groups. Moreover, DIF magnitudes on item difficulty parameters (b-DIF magnitude) of 0.4 and 0.8 were studied, corresponding to the differences in item difficulties between two groups. The difficulty parameters were increased by 0.4 or 0.8 to generate item responses for the focal group, indicating that the DIF items are more difficult for the focal group. Here, only unidirectional drift was considered since it represents a worst scenario in which the parameter drift has a

maximum effect during parameter estimation (Wells et al., 2002). Additionally, two *b*-DIF magnitudes were selected because the value of 0.4 represents the negligible DIF magnitude (e.g., Clauser, Mazor, & Hambleton, 1993; Donoghue, Holland, & Thayer, 1993; Penfield, 2001) while the value of 0.8 is useful to quantify the effect of DIF magnitude on power and type I error rate (Magis & De Boeck, 2012). The percentage of DIF items was set to 10% and 20%, and the first 10% or 20% of test items were selected as DIF items. In sum, there were in total $2 \times 3 \times 2 \times 2 = 24$ conditions. For each condition, 100 replications were generated.

The lasso penalty was used in this preliminary analysis. According to Equation 3.11, the parameter estimates $\widehat{\boldsymbol{\beta}}$ can be obtained as follows:

$$\widehat{\boldsymbol{\beta}} = arg\,max\left\{l(\boldsymbol{\beta}) - \lambda \sum_{i=1}^{I} |\beta_{2i}|\right\} \qquad (3.13)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}) = (\beta_{01}, \dots, \beta_{0I}, \beta_{11}, \dots, \beta_{1I}, \beta_{21}, \dots, \beta_{2I})$ represent the vector of parameters and only DIF parameters $\boldsymbol{\beta_2} = (\beta_{21}, \dots, \beta_{2I})$ need to be penalized, $l(\boldsymbol{\beta})$ is the log-likelihood of the model, and λ is the tuning parameter. An item displaying DIF or not is indicated by the corresponding DIF parameter estimate: $\hat{\beta}_{2i} = 0$ indicates item *i* is a non-DIF item while $\hat{\beta}_{2i} \neq 0$ indicates item *i* is a DIF item.

Four model selection techniques including AIC, BIC, 5-fold CV and 10-fold CV were compared under each condition. The hit rates and false alarm rates were recorded as two outcome measures, which can be calculated according to Table 3.2.

Table 3.2 Definition of Hit Rates and False Alarm Rates

		The item is actually a:	
		DIF item	non-DIF item
The item is detected as:	DIF item	N_1	N_2
	non-DIF item	N_3	N_4

Assume there are in total $N_1 + N_2 + N_3 + N_4$ items, the hit rate equals $N_1/(N_1 + N_3)$, and the false alarm rate equals $N_2/(N_2 + N_4)$. Specifically, the hit rate indicates the proportion of DIF items that are correctly flagged as DIF items, and the false alarm rate indicates the proportion of non-DIF items that are incorrectly flagged as DIF items.

The full simulation study was implemented in R (R Development Core Team, 2013). The IRT ability was estimated using the mirt R package (Chalmers, 2012), and the penalized estimation was conducted using the glmnet R package (Friedman et al., 2010). Figure 3.1 shows the relationship between λ and AIC (left y-axis)/BIC (right y-axis) under one selected condition (20 items; 2000 examinees in each group; *b*-DIF magnitude equals 0.8; and 20% DIF items on a test). The filled black dots in Figure 3.1 indicate the optimal λ values selected by AIC and BIC where $\lambda^*_{AIC} = 0.00037$, and $\lambda^*_{BIC} = 0.00104$.

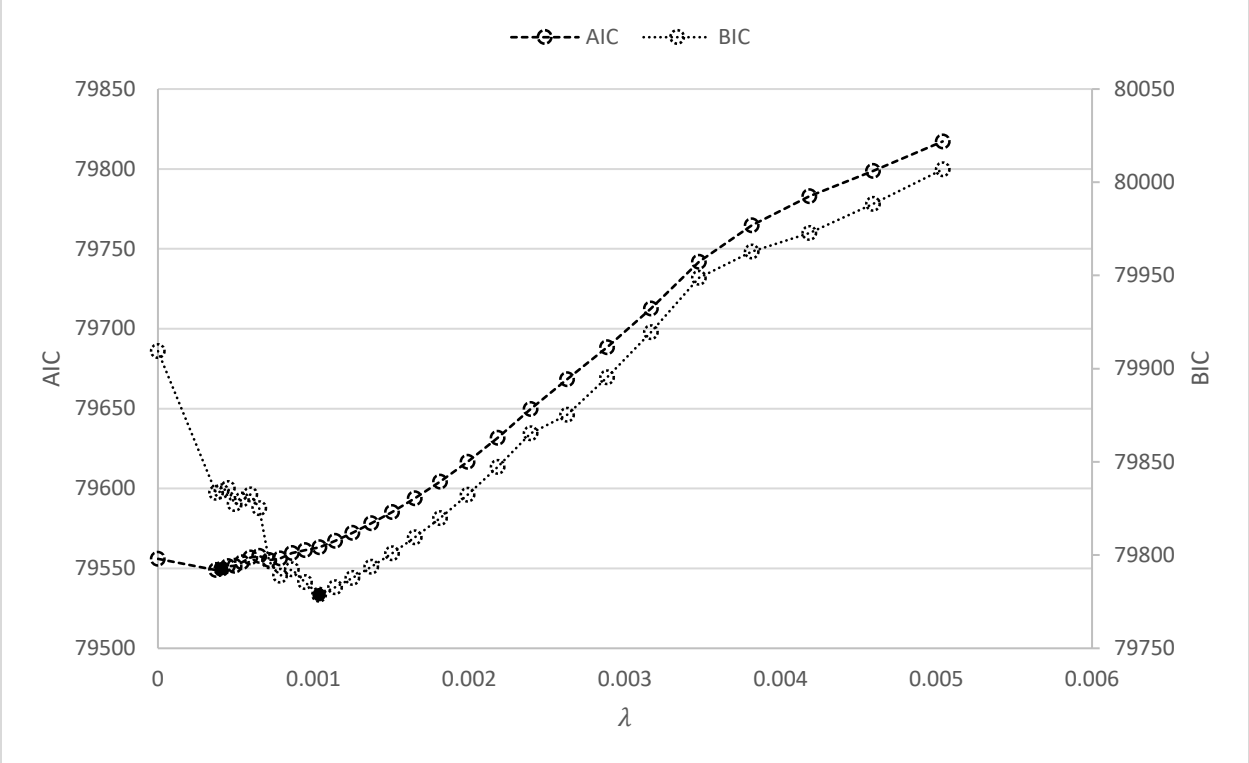

Figure 3.1 An Example of AIC and BIC Values for a Series of λs (the filled black dots indicate the optimal λ values under AIC and BIC)

Figure 3.2 shows the relationship between λ and the average deviance across all validation folds under the same condition using 5-fold CV and 10-fold CV, and the filled black dots indicate the optimal λ values selected by two methods where $\lambda^*_{CV5} = \lambda^*_{CV10} = 0.00049$.

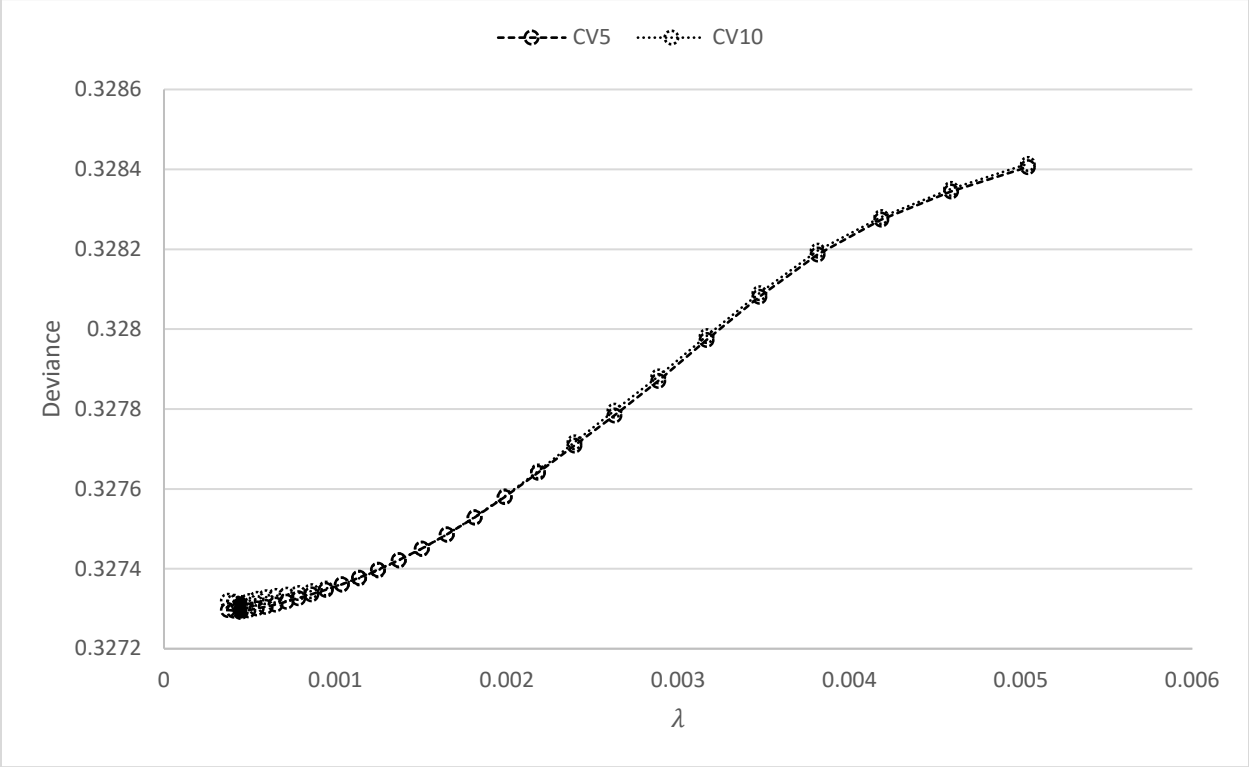

Figure 3.2 An Example of Average Deviance Values for a Series of λs using 5-fold and 10-fold CV (the filled black dots indicate the optimal λ values under CV5 and CV10)

In addition, Figure 3.3 demonstrates the regularization paths of DIF parameter estimates for a series of $log(\lambda)$ values. The red and yellow solid lines represent the DIF parameter estimates of four DIF items and sixteen non-DIF items correspondingly. We can see that as $log(\lambda)$ increases, some DIF parameters quickly shrink to zero indicating a small DIF effect; while other DIF parameters require large values of λ to reach 0

indicating a large DIF effect. As expected, all non-DIF items approach to 0 earlier than the DIF items as λ increases.

The black dotted vertical lines in Figure 3.3 demonstrate the DIF detection results. Since CV5 and CV10 selected the same λ in this particular example, only three vertical lines were plotted for four model selection methods. For those items whose path lines intersect with the vertical line, they are flagged as DIF items. For example, twelve items are flagged as DIF items using AIC including four true DIF items and eight false diagnosed DIF items. CV5 and CV10 yield exactly the same detection results as AIC does. Four items are flagged as DIF items using BIC which are all true DIF items. It means in this example, all model selection criteria can successfully identify the true DIF items, but AIC and CVs have too many false alarms due to smaller λs. BIC is more conservative leading to a better differentiation between DIF items and non-DIF items compared other criteria.

Moreover, Figure 3.3 also shows that when λ ranges from the 0.00095 to 0.00263, that is, $log(\lambda)$ ranges from -6.965 to -5.941, all true DIF items can be correctly identified without any false alarms, indicating that the penalized regularization methods provide a robust solution for DIF detection.

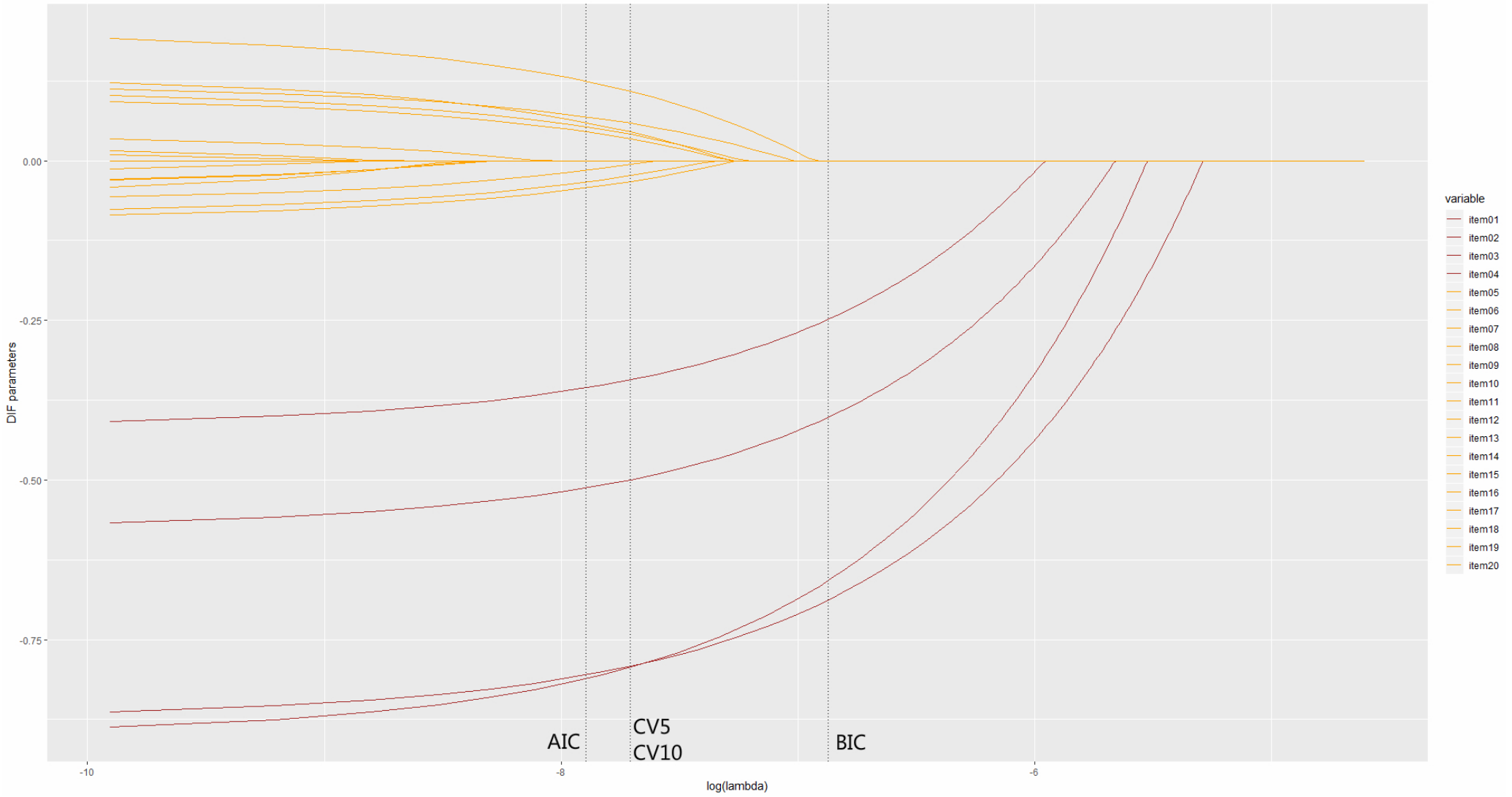

Figure 3.3 Regularization Paths of DIF Parameter Estimates for a Series of $log(\lambda)$ Values

Figure 3.4 summarizes the hit rates using four model selection criteria. Generally, the hit rates decrease when at least one of the four scenarios happens: 1) the sample size becomes smaller, 2) the test becomes longer, 3) the percentage of DIF items on a test becomes larger, and 4) the *b*-DIF magnitude becomes smaller.

Also, Figure 3.4 shows that AIC selects DIF models with the greatest hit rates, and most hit rates are greater than 0.9. The 5-fold CV and 10-fold CV have almost the same hit rates which are greater than 0.8 and are slightly smaller than the hit rates of AIC under most conditions. Models selected by BIC demonstrate the lowest hit rates among four criteria: when *b*-DIF magnitude equals 0.8, most hit rates are greater than 0.9 and only one of them is around 0.8; however, when *b*-DIF magnitude equals 0.4, the average hit rate is greater than 0.7, and the differences in average hit rates between AIC and BIC, CV5 and BIC, CV10 and BIC are 0.22, 0.14 and 0.16 correspondingly.

Therefore, changing DIF magnitude has the greatest impact on hit rates compared to changing other manipulated factors such as test length, group size, or percentage of DIF items on a test: when *b*-DIF magnitude equals 0.8, the hit rates are close to 1.0 under most conditions no matter which model selection criterion is used, but when *b*-DIF magnitude equals 0.4, the hit rates are much lower. However, previous research examined that when *b*-DIF magnitude was around 0.4, the DIF items had a minimal impact on equating and ability estimation (Wells et al., 2002). This indicates that small *b*-DIF magnitudes (e.g., equal to or smaller than 0.4) are undetectable and they are of less concern in practice. In other words, items having negligible DIF is the primary reason for low hit rates in this study. Thus, in terms of the ability of correctly detecting true DIF items, all of the four model selection criteria are acceptable.

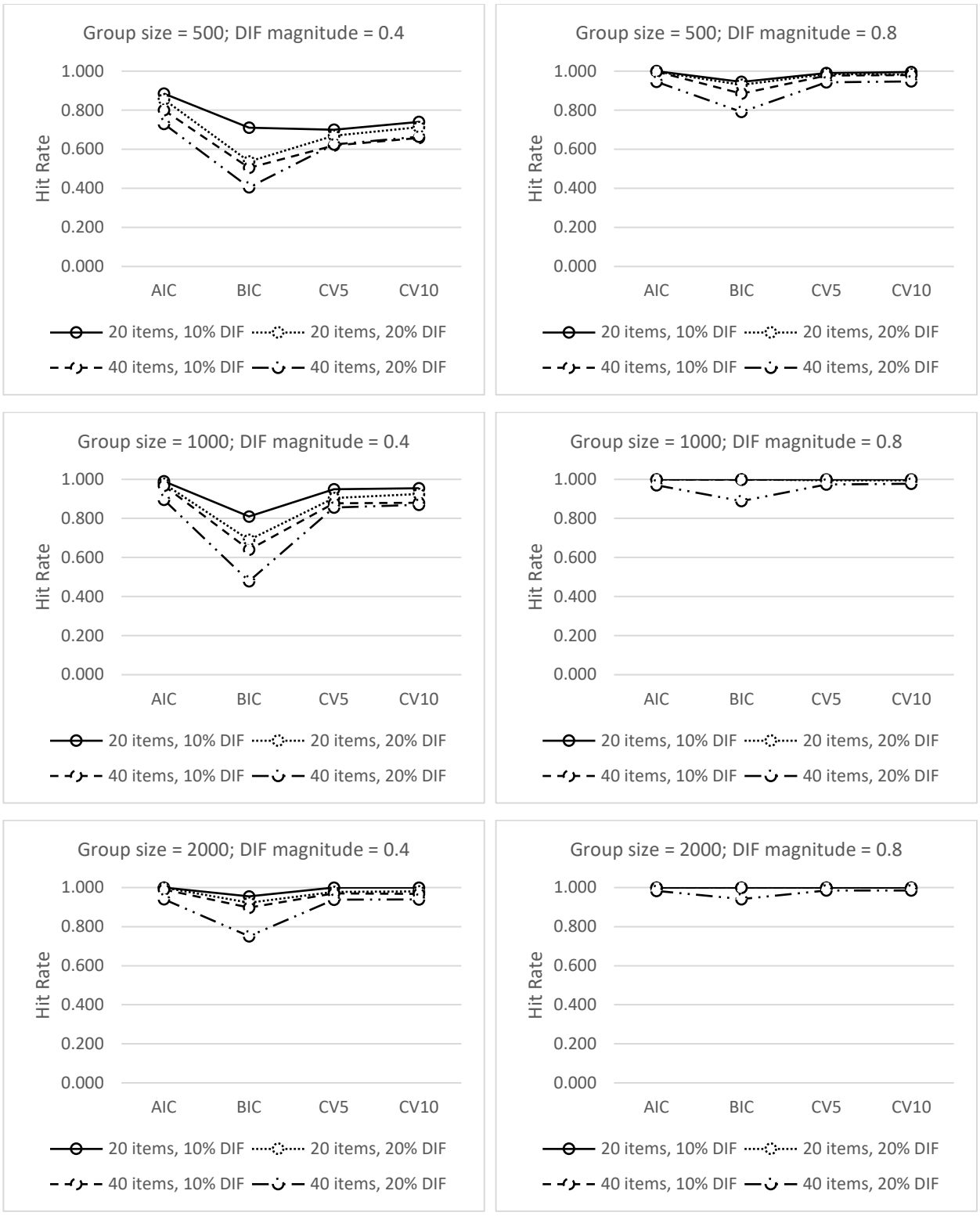

Figure 3.4 Hit Rates for Four Model Selection Criteria

Figure 3.5 summarizes the false alarm rates using four model selection criteria. We can see that group sample size and test length have little impact on the false alarm rates. However, if there are more DIF items on a test or the *b*-DIF magnitude becomes larger, the false alarm rates slightly increase regardless of model selection methods. Particularly, BIC yields well-controlled false alarm rates which are smaller than 0.05 under most conditions. The other three model selection criteria have similar false alarm rates but most values are greater than 0.3.

These findings are consistent with the example showed in Figure 3.3, indicating that BIC is more conservative than AIC and CV in finding the optimal tuning parameters, resulting in a satisfying balance between hit rates and false alarm rates. BIC is able to not only flag high-impact DIF items, but also prevent over-identification of DIF items with few false alarms. Therefore, BIC will be used as the model selection criterion for all future analyses.

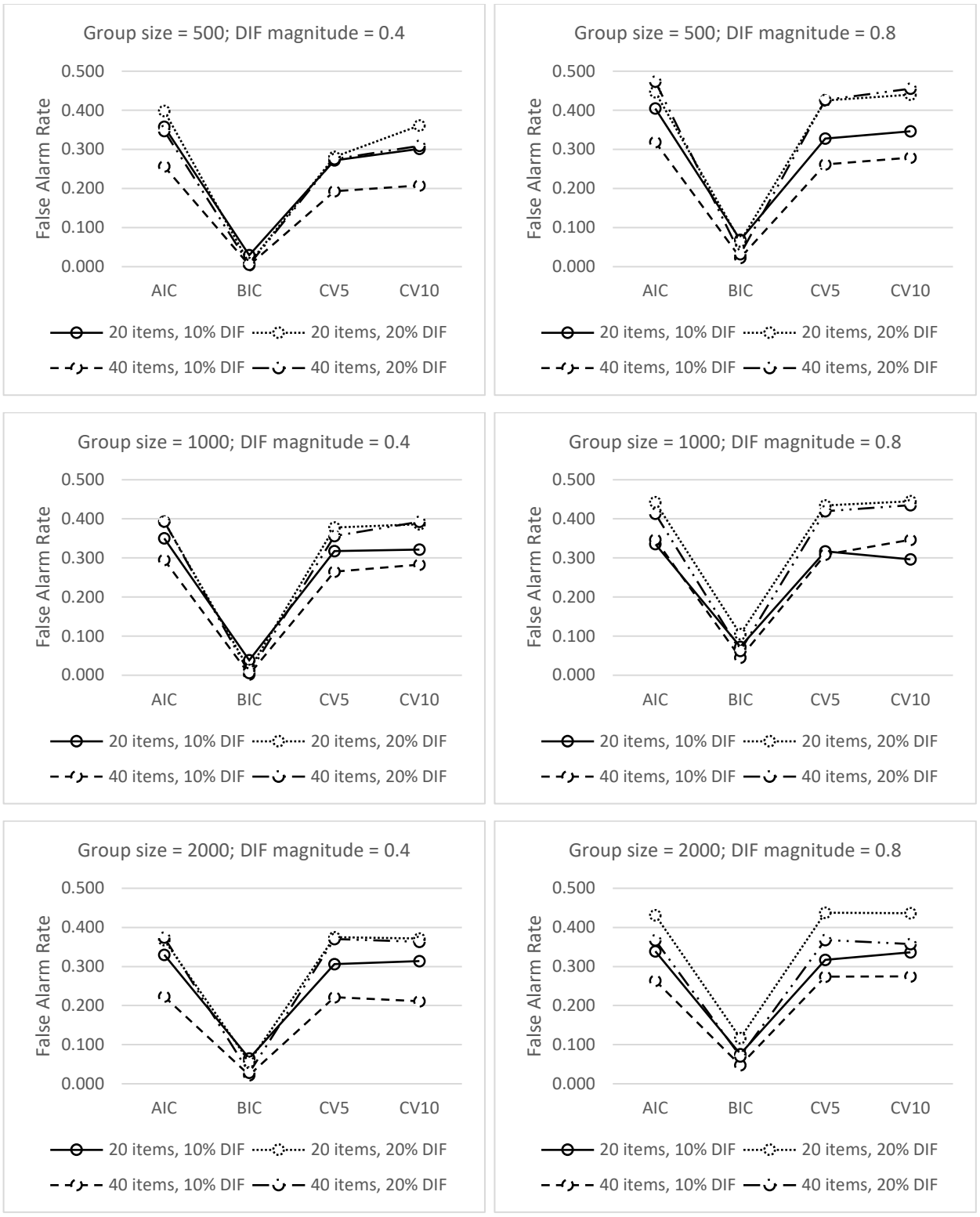

Figure 3.5 False Alarm Rates for Four Model Selection Criteria

3.1.4.2 Comparison of Using Test Score and Ability Estimate as a Proxy for Ability

The purpose of the second preliminary analysis is to compare which of the ability measures, the observed test score or the IRT ability estimate, has a better performance in DIF detection. Simulated datasets were generated to perform DIF analyses based on Equation 3.12. Item parameters displayed in Table 3.1 were used to generate simulated datasets. Same manipulated conditions (24 conditions with 2 test length, 3 group sizes, 2 percentages of DIF items, and 2 DIF magnitudes) were studied in this preliminary analysis, and the latent ability distribution was set to $N(0,1)$ for both groups. For each condition, 100 replications were generated.

The DIF detection model in Equation 3.11 was used in this analysis, where θ_p can either be the observed total score for person p or be estimated using EAP under the 2PL model. Estimated model parameters $\widehat{\boldsymbol{\beta}}$ were obtained according to Equation 3.13. The tuning parameter λ was selected using BIC based on the findings in Section 3.1.4.1. The hit rates and false alarm rates were recorded as outcome measures. This simulation study was implemented in R (R Development Core Team, 2013).

Figure 3.6 demonstrates the hit rates using two different ability measures. Under all manipulated scenarios, using the IRT ability estimate yields higher hit rates than using the total score, and the largest difference is about 0.3. Specifically, as group size increases, the hit rates of both ability measures increase, while as test length or percentage of DIF items increases, the hit rates of two measures decrease under most situations. Similar to the previous results, two DIF detection models with different ability measures are not sensitive to small *b*-DIF magnitude but perform well in predicting items with non-negligible DIF.

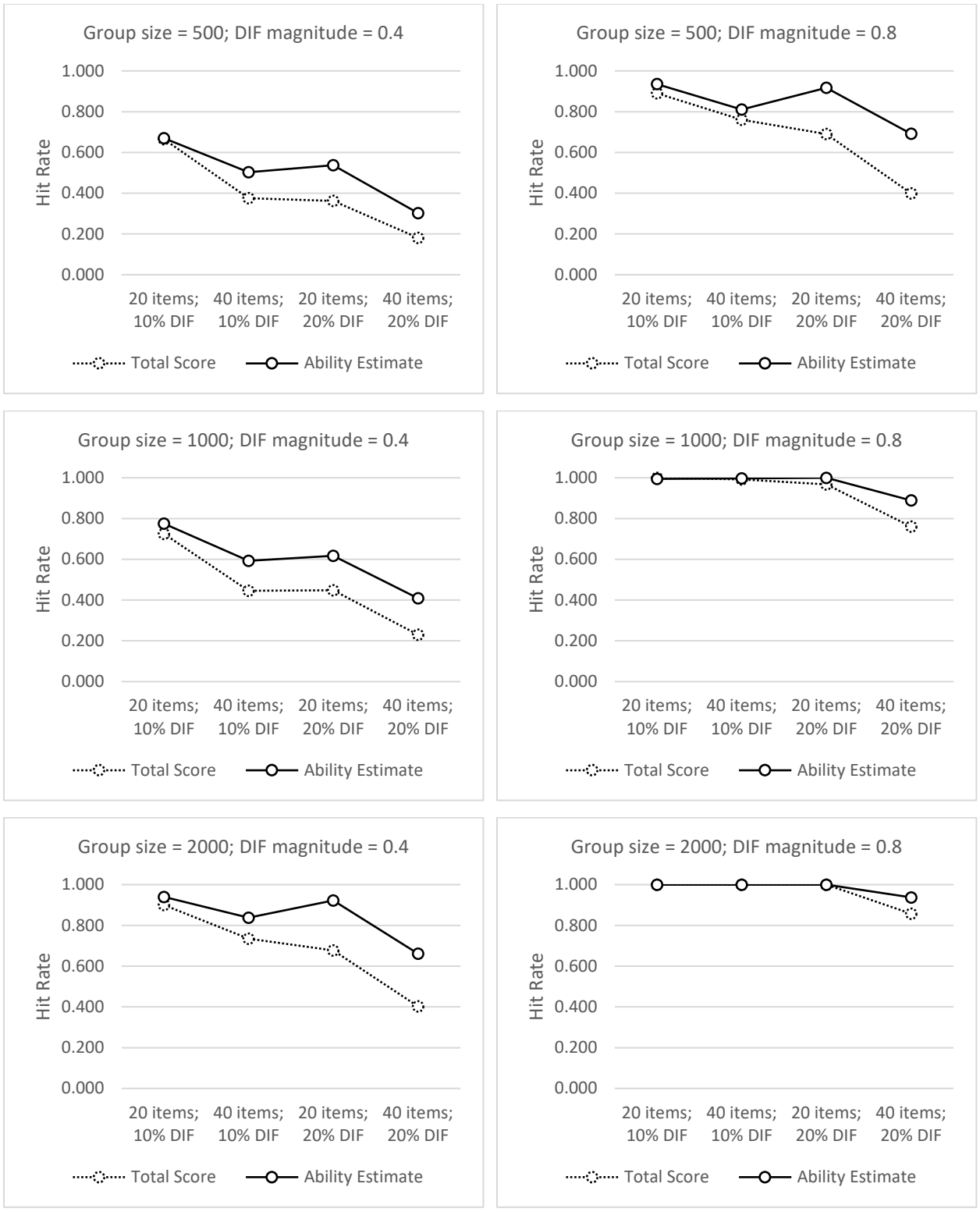

Figure 3.6 Hit Rates of using Two Different Ability Measures

The false alarm rates are displayed in Figure 3.7. Specifically, according to the left-bottom and right-top subplot, using the IRT ability estimate yields larger false alarm rates than using the observed score under several conditions. On the contrary, according to the right-bottom subplot, using the total score yields higher false alarm rates. Under all other conditions, the false alarm rates of both measures are similar to each other. The highest false alarm rate of using the IRT ability estimate is controlled under 0.1, while the highest false alarm rate of using the total score is close to 0.2. Generally, both ability measures yield acceptable false alarm rates under most conditions.

In conclusion, using the IRT ability estimate as a proxy for person's latent ability outperforms using the total score in terms of their DIF detection performance by taking both hit rates and false alarm rates into consideration. Therefore, the IRT ability estimate will be employed as the ability measure in the DIF detection model for all future analyses.

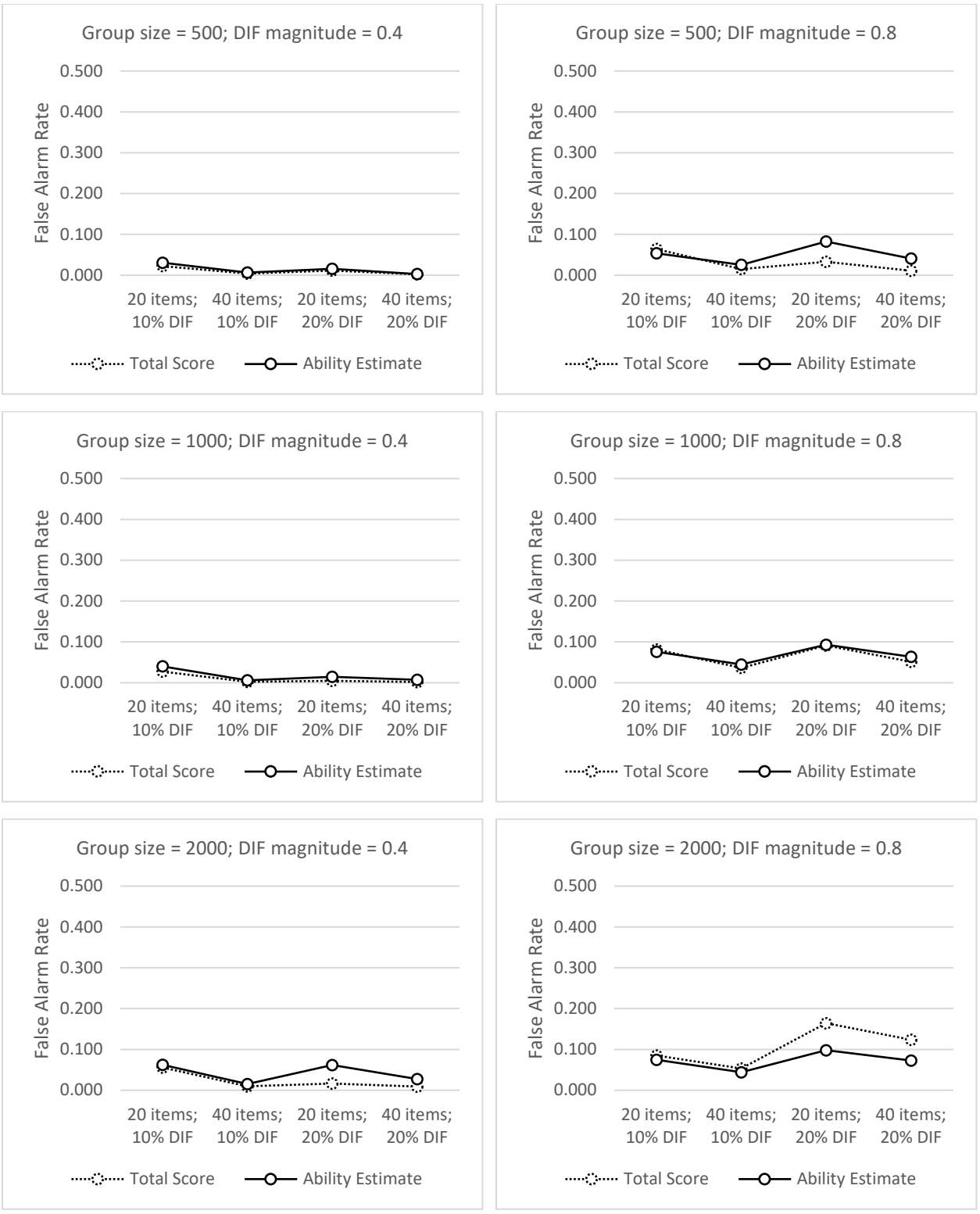

Figure 3.7 False Alarm Rates of using Two Different Ability Measures

3.2 Simulation Studies

3.2.1 Simulation Study One

The first simulation study aims to answer the first research question—which penalty function yields the best operating characteristics in DIF detection, and the second research question—how does each of the manipulated factors impact the number of items flagged correctly and incorrectly as DIF items using the proposed DIF detection model.

Simulated datasets consisting of *P* persons and *I* items were generated from the 2PL model showed in Equation 3.12. Similar to previous preliminary analyses, the item parameters in Table 3.1 were used to generate data where $a_i \sim lognormal(0, 0.5^2)$ and $d_i \sim N(0,1)$, and the latent ability distribution was set to $N(0,1)$ for both groups.

The manipulated factors and the levels of each factor used in this simulation study included:

- Test length: 20 items, 40 items
- Sample size: 1000, 2000, 4000 examinees
- Sample size ratio: 1:1, 4:1 (reference/focal group size—500/500, 800/200; 1000/1000, 1600/400; 2000/2000, 3200/800)
- Percentage of DIF items: 10%, 20%
- DIF type: uniform DIF only (*b*-DIF), nonuniform DIF with drift on discrimination parameter only (*a*-DIF), nonuniform DIF with drift on both difficulty and discrimination parameter (*a*-DIF and *b*-DIF)
- DIF magnitude: 0.4, 0.8 for drift on b_i (*b*-DIF magnitude); 0.5, 1.0 for drift on a_i (*a*-DIF magnitude)

There were in total $2\ (\text{test lengths}) \times 3\ (\text{sample sizes}) \times 2\ (\text{sample size ratios}) \times 2\ (\text{percentages of DIF items}) \times\ 3\ (\text{DIF types}) \times\ 2\ (\text{DIF magnitudes}) = 144$ conditions, among which 48 were uniform DIF conditions, and 96 were non-uniform DIF conditions.

The general DIF model (see Equation 3.1) was used to detect both uniform and nonuniform DIF items. According to Section 3.1.4.2, θ_p was estimated using EAP under the 2PL model. Model parameter estimates $\widehat{\boldsymbol{\beta}}$ were obtained using penalized MLE based on Equation 3.3, and three forms of penalty functions were studied including the traditional lasso (see Equation 3.5), the elastic net (see Equation 3.6), and the adaptive lasso (see Equation 3.7). The tuning parameters were selected using BIC according to Section 3.1.4.1.

For each condition, 100 replications were generated. The hit rates indicating the proportion of DIF items that are correctly flagged as DIF items, and the false alarm rates indicating the proportion of non-DIF items that are incorrectly flagged as DIF items were recorded as the outcome measures. Specifically, under uniform DIF conditions, $\hat{\beta}_{2i} = 0$ indicates item *i* is a non-DIF items while $\hat{\beta}_{2i} \neq 0$ indicates item *i* is a DIF item, while under nonuniform DIF conditions, $\hat{\beta}_{3i} = 0$ indicates item *i* is a non-DIF items while $\hat{\beta}_{3i} \neq 0$ indicates item *i* is a DIF item.

However, in reality, a person's ability θ_p may be over- or under-estimated when the fitted model is not exactly correct. For example, when a test includes multiple-choice items, there is a chance that examinees will answer items correctly by guessing. But in practice and in some measurement models, it is often assumed that the effects of guessing are negligible, and the item and ability parameter might be inaccurately estimated in this

case. Additional analyses were conducted to represent such a scenario—the item responses were generated from the 3PL model incorporating guessing effects as follows:

$$P\left(y_{ip}=1\middle|\theta_p,a_i,d_i,c_i\right)=c_i+(1-c_i)\frac{exp\left[-\left(a_i\theta_p+d_i\right)\right]}{1+exp\left[-\left(a_i\theta_p+d_i\right)\right]} \qquad (3.14)$$

where c_i was set to 0.2 for all items, and the person's ability estimates ($\hat{\theta}_p$) were obtained using a different IRT model—the 2PL model. Meanwhile, all other settings remained unchanged. Under this design, an additional source of misfit—the biased ability estimates—was introduced to the proposed DIF model in Equation 3.1. And the purpose is to investigate how the ability parameter estimates impact the DIF detection results.

The full simulation study was implemented in R (R Development Core Team, 2013). The ability estimation was conducted using the mirt R package (Chalmers, 2012), and the penalized estimation was conducted using the glmnet R package (Friedman et al., 2010).

3.2.2 Simulation Study Two

The second simulation study aims to address the third research question—what are the differences between the proposed method and other existing techniques—the logistic regression method and the likelihood ratio test, in terms of their DIF detection performance.

Logistic regression was selected from three aforementioned non-IRT approaches because it is the starting point for the regularized DIF detection model, and it is of great interest to see the differences in DIF detection using two types of logistic regression models. Among the IRT-based DIF methods, the application of Raju's method is most restrictive since the integrals (see Equations 2.22 and 2.23) used to generate the area

measures are not able to yield finite results if the guessing parameter in the 3PL model is unequal across groups. In addition, although previous research indicated that the test statistics χ^2 and G^2 computed from Lord's χ^2 test and the likelihood ratio test were asymptotically the same (Kim & Cohen, 1995), the accuracy of Lord's χ^2 test was heavily dependent on the accuracy in estimating the variance and covariance matrix for the item parameter estimates (Thissen & Wainer, 1982). Considering these aspects, the likelihood ratio test was used as another comparison method in this simulation study. Specifically, the logistic regression DIF detection was realized using the difR R package (Magis, Beland, & Raiche, 2013), and the likelihood ratio test was implemented using the mirt R package (Chalmers, 2012).

The DIF detection methods were compared by means of receiver operating characteristic (ROC) curves under the following conditions:

- Condition 1: 40 items, 20% DIF items, group sizes of 1000 in both groups, uniform DIF with *b*-DIF magnitude equal to 0.8
- Condition 2: 40 items, 20% DIF items, group sizes of 1600 and 400 in the reference and focal groups, uniform DIF with *b*-DIF magnitude equal to 0.8
- Condition 3: 40 items, 20% DIF items, group sizes of 1000 in both groups, nonuniform DIF with *a*-DIF magnitude equal to 1.0
- Condition 4: 40 items, 20% DIF items, group sizes of 1600 and 400 in the reference and focal groups, nonuniform DIF with *a*-DIF magnitude equal to 1.0
- Condition 5: 40 items, 20% DIF items, group sizes of 1000 in both groups, nonuniform DIF with *a*-DIF magnitude equal to 1.0 and *b*-DIF magnitude equal to 0.8

- Condition 6: 40 items, 20% DIF items, group sizes of 1600 and 400 in the reference and focal groups, nonuniform DIF with *a*-DIF magnitude equal to 1.0 and *b*-DIF magnitude equal to 0.8

The reason for using ROC curves is that the DIF detection performance of the logistic regression and the likelihood ratio test depends on the selection of significance level (e.g., Cohen et al., 1996; Narayanan & Swaminathan, 1996), while the performance of regularization methods relies on the selection of tuning parameters. This makes the comparison in terms of hit rates and false alarm rates more complicated because comparing hit rates without controlling for false alarm rates is meaningless.

Therefore, the ROC curves were suggested in previous research, illustrating the ability of a DIF detection method in classifying DIF items and non-DIF items (Magis et al., 2015). The ROC curves can be created by plotting the false alarm rates on the x-axis against the hit rates on the y-axis. Therefore, the hit rates and false alarm rates were calculated by increasing the tuning parameter values for the regularization methods and by decreasing the significance levels for the logistic regression approach and the likelihood ratio test. For each of the selected condition, 100 replications were generated, and the average ROC curve was finally computed for each DIF detection method under each of the conditions. The full simulation study was implemented in R (R Development Core Team, 2013).

3.3 Empirical Data Analyses

The empirical data were from the Trends in International Mathematics and Science Study (TIMSS), which is an international comparative study of student

achievement, measuring fourth- and eighth-grade students' knowledge and skills on mathematics and science. The most recent TIMSS data collection was conducted in 2015, and about 60 countries consisting of more than 580,000 students participated in TIMSS 2015 (Mullis, Martin, Foy, & Hooper, 2016).

This study used the newest TIMSS 2015 data of eighth-grade students in the United States, and the items were tested for DIF between boys and girls. Considering sample size and test length, for mathematics data, those students assigned the fifth booklet were finally selected, while for science data, those students assigned the second booklet were selected (the booklet number was randomly selected). In addition, the analyses only considered the items that were dichotomously scored, and those students with missing or incomplete responses were excluded from the analyses. Accordingly, the analyses of 18 mathematics items were carried out on 1197 students with a distribution of 595 boys and 602 girls, and the analyses of 20 science items were conducted on 1290 students consisting of 687 boys and 603 girls.

Typically, DIF studies are conducted to find out DIF items in a test, and the results are used to investigate the potential sources of DIF. In this study, the main purpose of empirical data analyses is to compare the DIF detection results of the proposed DIF model with other commonly used DIF methods in order to see if there are any similarities or differences between them. Specifically, three DIF methods used in the second simulation study were used again to analyze the empirical datasets.

CHAPTER 4. RESULTS

4.1 Simulation Study 1

The first simulation study evaluates the DIF detection performance of different regularized logistic regression methods with three types of penalty functions including the lasso penalty, the elastic net penalty, and the adaptive lasso penalty by comparing the operating characteristics under 144 conditions, which addresses the first research question—which penalty function yields the best operating characteristics in DIF detection. Moreover, since the 144 conditions were created by combining six manipulated factors including test length, sample size, sample size ratio, percentage of DIF items, DIF type, and DIF magnitude in different ways, the second research question—how does each of the manipulated factors impact the number of items flagged correctly and incorrectly as DIF items—is answered by comparing the hit rates and false alarm rates at different levels of each manipulated factor.

The results are presented separately according to uniform and nonuniform DIF conditions in the following sections. Each section starts with the results for data generated from the 2PL model, with hit rate results preceding false alarm rate results, followed by discussions on 3PL-generated datasets.

4.1.1 Uniform DIF Conditions

For uniform DIF conditions, there only exist differences in the difficulty parameters between the reference and focal groups. Figures 4.1 and 4.2 demonstrate the hit rates using three types of penalty functions in parameter estimation with *b*-DIF magnitude equal to 0.4 and 0.8 respectively.

The adaptive lasso has slightly lower hit rates compared to the lasso and the elastic net only when *b*-DIF magnitude is 0.8 and group sizes are 1000/1000, 2000/2000 and 3200/800 (see the left-center, left-bottom and right-bottom subplots in Figure 4.2). However, in these conditions, the differences in hit rates between any two methods are smaller than 0.02, which are negligible. In all other conditions, we can see that the adaptive lasso significantly outperforms other penalty functions. Therefore, in uniform DIF detection, generally the adaptive lasso has the best performance in correctly identifying true DIF items, regardless of group size, test length, percentage of DIF items, and *b*-DIF magnitude. In addition, the elastic net has the worst performance, although the average difference in hit rates between the lasso and the elastic net is only 0.029. It indicates that adding the $\ell 2$ penalty term to the loss function does not improve the model's ability to detect uniform DIF items.

In terms of the impact of each manipulated factor on hit rates, both figures indicate that the hit rates of all three procedures increase when the sample size increases from small (1000 examinees) to large (4000 examinees), independent of sample size ratio, test length, percentage of DIF items and *b*-DIF magnitude. The hit rates are consistently higher if the reference and focal group sizes are equal when controlling for other manipulated factors. Moreover, generally the hit rates decrease when the test length increases and/or the percentage of DIF items increases under the conditions of matching group size and *b*-DIF magnitude. That is, the hit rates are higher for 20-item conditions compared to 40-item conditions, and are higher for 10% DIF conditions compared to 20% DIF conditions, with the exception of *b*-DIF magnitude equal to 0.8, group size equal to 1600/400, and number of test items equal to 20 (see the right-center subplot in

Figure 4.2). Lastly, the average hit rates of large uniform DIF conditions (see Figure 4.2) are two to three times the average hit rates of small uniform DIF conditions (see Figure 4.1) for all three methods.

Specifically, the adaptive lasso works well in identifying items with large uniform DIF effects, and the average hit rate of the 24 conditions in Figure 4.2 is 0.813. Furthermore, when the reference and focal group have the same group size, the average hit rate is 0.894; however, for unbalanced group sizes (the reference and focal group ratio equals 4:1), the average hit rate is 0.733, indicating that one needs more examinees in the setting to obtain a larger hit rate. On the other hand, weak uniform DIF identification is much harder and the average hit rate of the 24 conditions in Figure 4.1 is only 0.371, indicating that the items with small magnitudes of uniform DIF are hard to be detected using the adaptive lasso methodology, especially when the sample size is small.

Figures 4.3 and 4.4 demonstrate the false alarm rates under exactly the same conditions corresponding to Figures 4.1 and 4.2. The false alarm rates are well-controlled for all methods under all conditions, and the average false alarm rates for the lasso, the elastic net, and the adaptive lasso are 0.010, 0.007 and 0.006 respectively. Therefore, the adaptive lasso not only has the best performance in identifying true DIF items, but also minimizes the possibility of incorrectly flagging non-DIF items as DIF items.

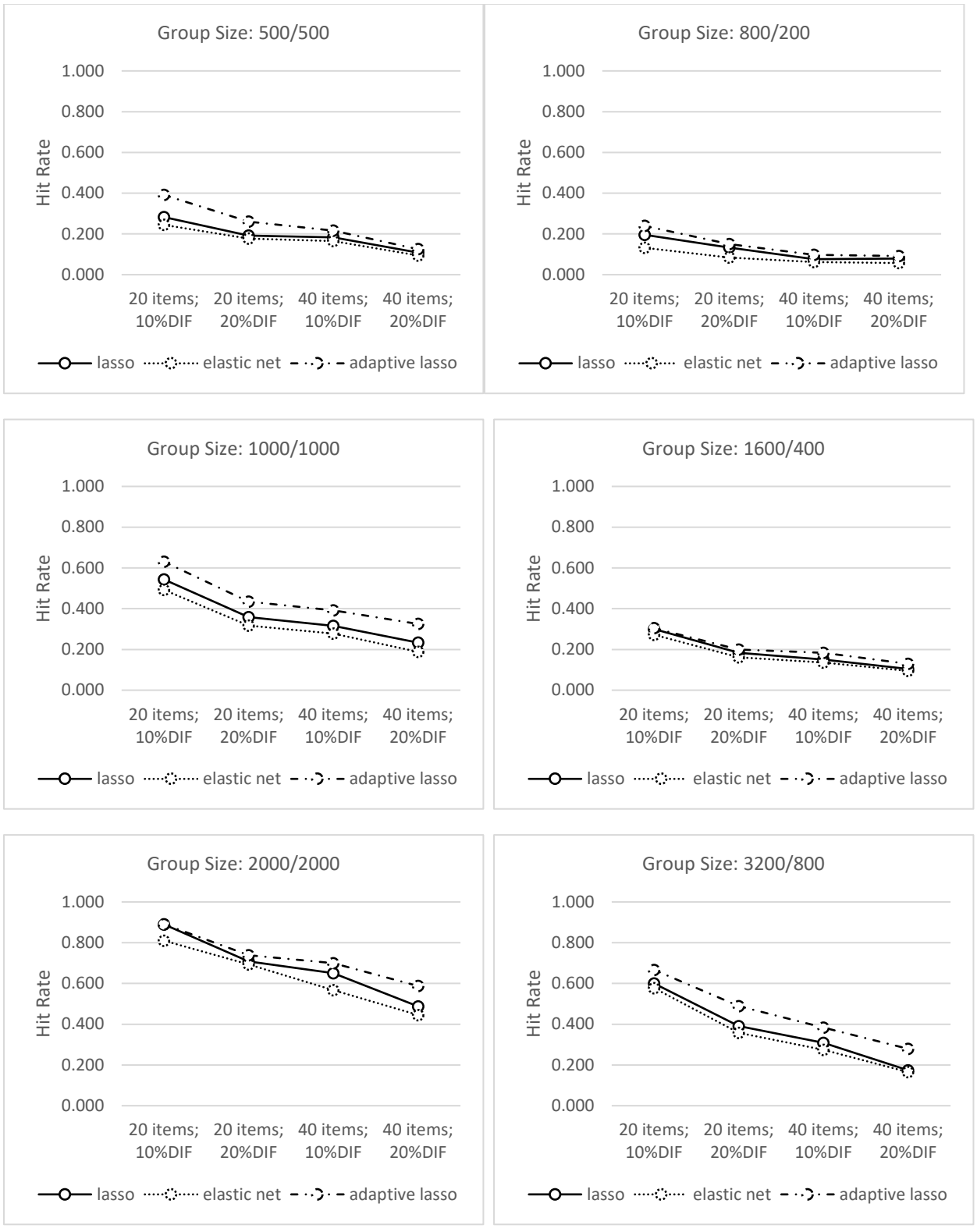

Figure 4.1 Hit Rates for Three Penalty Functions under Uniform DIF Conditions with *b*-DIF Magnitude Equal to 0.4

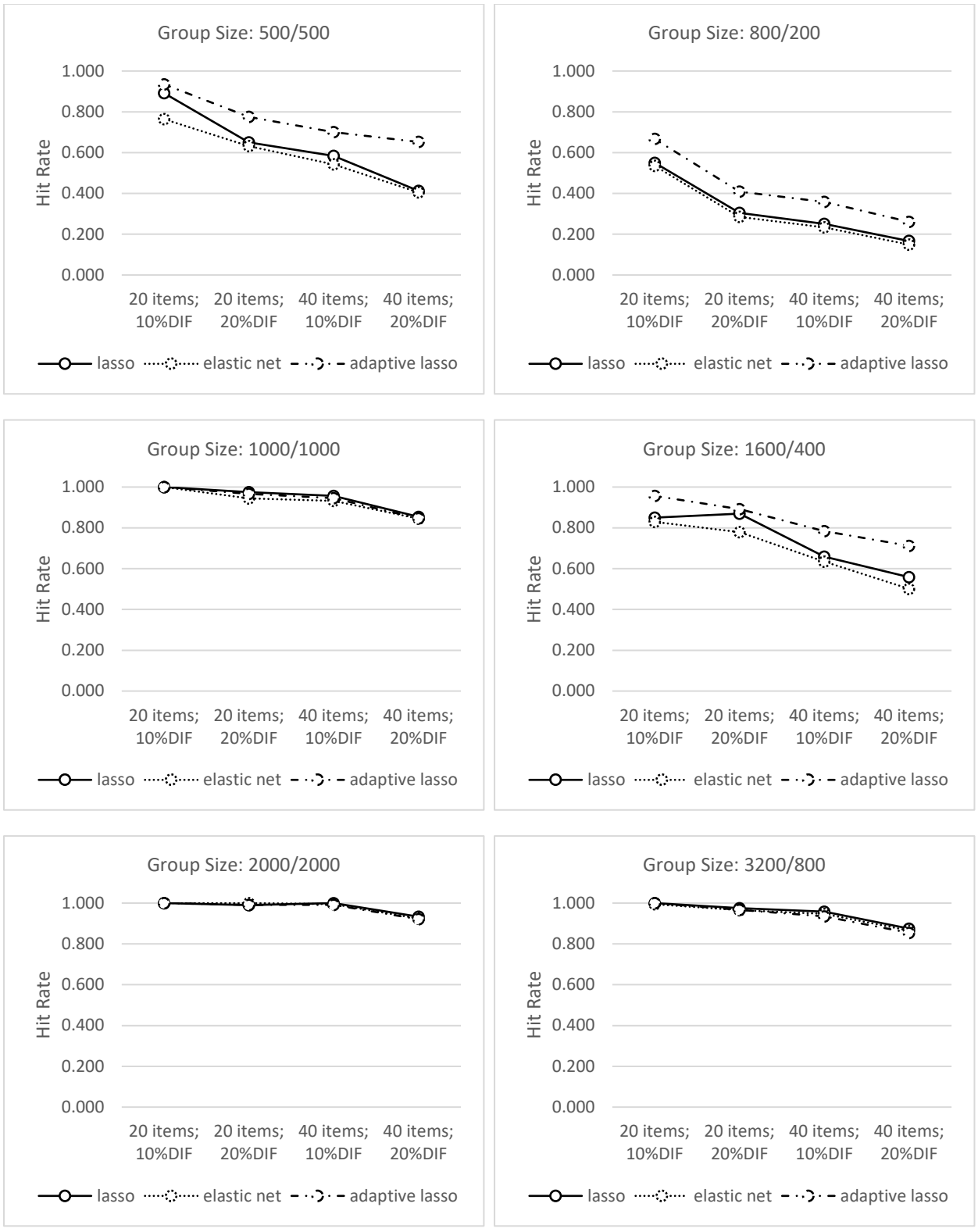

Figure 4.2 Hit Rates for Three Penalty Functions under Uniform DIF Conditions with *b*-DIF Magnitude Equal to 0.8

Figure 4.3 False Alarm Rates for Three Penalty Functions under Uniform DIF Conditions with *b*-DIF Magnitude Equal to 0.4

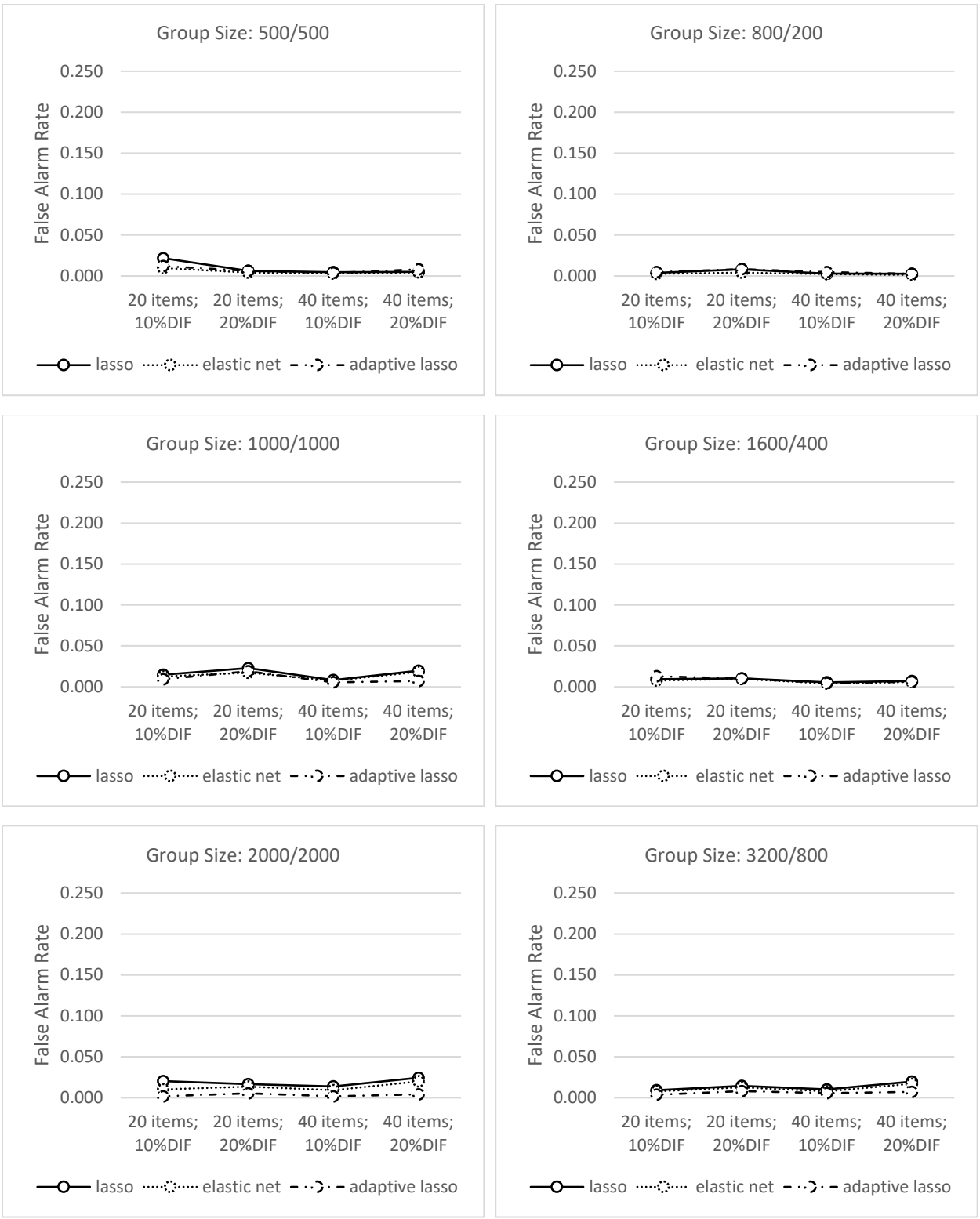

Figure 4.4 False Alarm Rates for Three Penalty Functions under Uniform DIF Conditions with *b*-DIF Magnitude Equal to 0.8

However, as mentioned in Section 3.2.1, in reality person's abilities are not always incorrectly estimated. Therefore, additional simulation analyses were conducted by generating item responses from the 3PL model instead of the 2PL model, and the hit rates and false alarm rates (y-axes) for 2PL- and 3PL-generated datasets under 48 uniform DIF conditions (x-axes) are displayed in Figure 4.5.

We can see that the accuracy of latent ability estimates has a great impact on DIF detection performance. When θ_p (see Equation 3.1) are estimated using an incompatible IRT model, many DIF items are unable to be detected, resulting in serious declines in hit rates. Specifically, the average differences in hit rates between the 2PL and 3PL cases for the lasso, the elastic net, and the adaptive lasso are 0.238, 0.227, and 0.295 respectively, indicating that the regularized logistic regression DIF model is not robust to biased ability estimates.

The false alarm rates decrease as well since it is more difficult for all items including both true DIF items and non-DIF items, to be detected as DIF items. The average false alarm rates for three methods are declined by 0.006, 0.005 and 0.003 correspondingly when using 3PL-generated datasets.

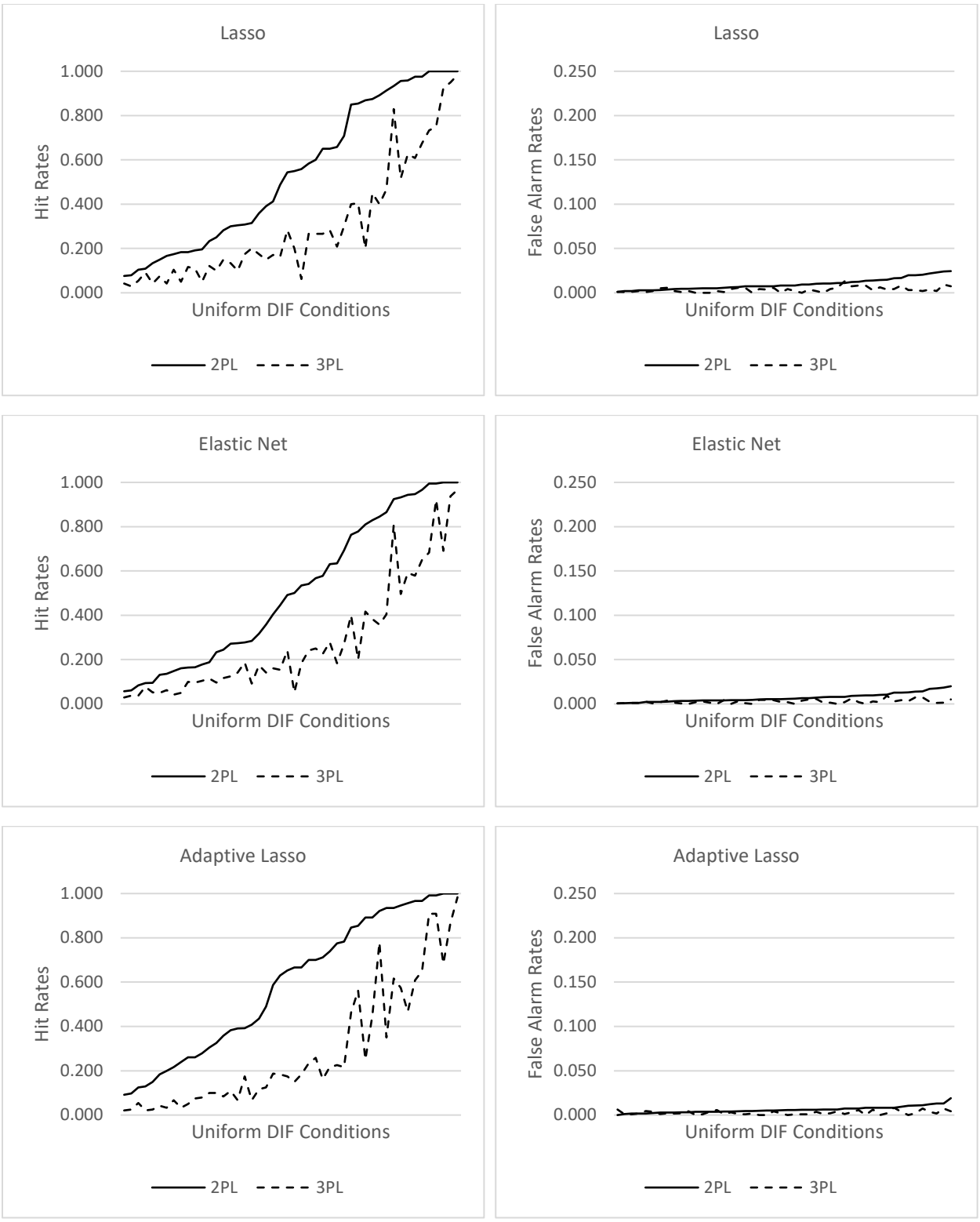

Figure 4.5 Comparison of Hit Rates and False Alarm Rates for 2PL- and 3PL-generated Datasets under 48 Uniform DIF Conditions

4.1.2 Nonuniform DIF Conditions

As stated in Section 2.2.2, an item displaying nonuniform DIF varies in the discrimination parameter, and possibly varies in the difficulty parameter (Mellenbergh, 1982). Therefore, the nonuniform DIF conditions are further divided to nonuniform DIF conditions with drift on the discrimination parameters (*a*-DIF) only, and nonuniform DIF conditions with drift on both discrimination and difficulty parameters (*a*-DIF and *b*-DIF).

4.1.2.1 Nonuniform DIF Conditions with Drift on the Discrimination Parameters Only

Figures 4.6 and 4.7 show the hit rates using three penalties in the regularized logistic regression DIF model with two levels of *a*-DIF magnitude equal to 0.5 and 1.0 respectively. The adaptive lasso has a much better performance in detecting nonuniform DIF items across all conditions compared to the lasso and elastic net approaches. Similar to uniform DIF conditions, the lasso slightly outperforms the elastic net, and the average difference in hit rates between them is 0.025.

Under all conditions, the hit rates increase when the total sample size increases. Moreover, the hit rates are always higher when the reference and focal group size ratio equal to 1 compared to unbalanced group sizes controlling for other factors. In terms of the impact of test length on hit rate, when there exist 20% DIF items, the hit rates decrease when the number of test items increases from 20 items to 40 items; however, when only 10% of test items are DIF items, the pattern is not consistent. Moreover, according to the figures, it seems that there is no relationship between the percentage of DIF items and hit rates since the trends are inconsistent under different conditions.

Similar to the findings in Section 4.1.1, the adaptive lasso has a better performance in identifying items with larger nonuniform DIF effects. The average hit rate

is 0.795 according to Figure 4.7 (*a*-DIF magnitude equal to 1.0) and is 0.421 according to Figure 4.6 (*a*-DIF magnitude equal to 0.5). Specifically, when the reference and focal group size ratio is 1:1, the average hit rates for correctly detecting DIF items with large and small nonuniform DIF are 0.915 and 0.543; when the group size ratio is 4:1, the average hit rates are 0.675 and 0.300 respectively.

Generally, the hit rates and the false alarm rates increase or decrease at the same time when changing the simulation scenarios. As shown in Figures 4.8 and 4.9, the false alarm rates are well-controlled under most conditions, and the average false alarm rates for the lasso, the elastic net, and the adaptive lasso methods are 0.042, 0.036 and 0.020 respectively. Although the average false alarm rates seem to be acceptable for all three methods, in certain conditions the false alarm rates are greater than 0.1 (see the left-bottom subplot in Figure 4.8 and the right-bottom subplot in Figure 4.9) and even around 0.2 (see the left-center and left-bottom subplots in Figure 4.9) if the lasso or elastic net penalty is used. However, the false alarm rates are always below 0.05 when using the adaptive lasso penalty and the largest value is 0.049 (see the right-bottom subplot in Figure 4.9). Again, the adaptive lasso has the best performance in identifying true DIF items and minimizes the possibility of misidentifying non-DIF items compared to the other two penalties.

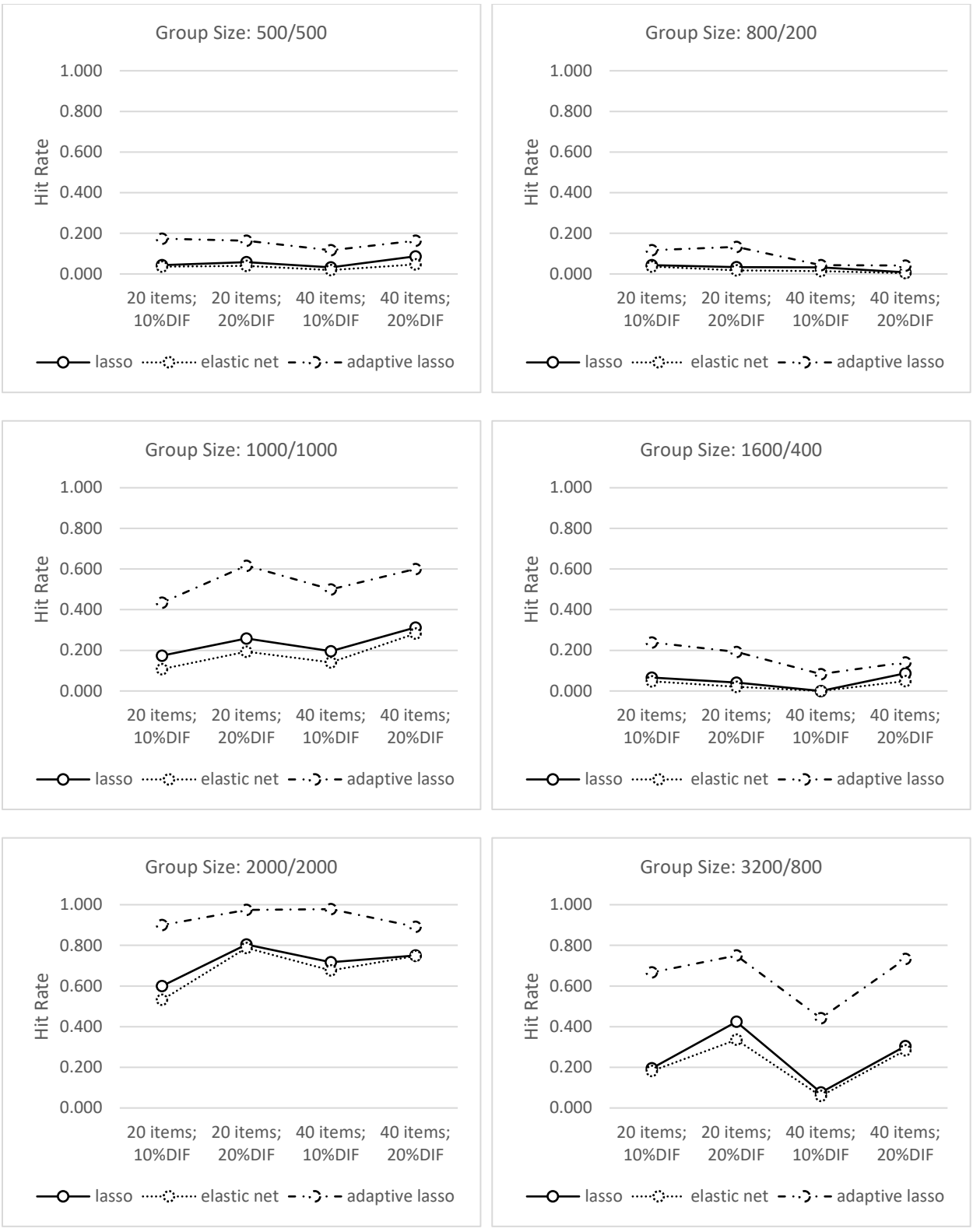

Figure 4.6 Hit Rates for Three Penalty Functions under Nonuniform DIF Conditions with *a*-DIF Magnitude Equal to 0.5

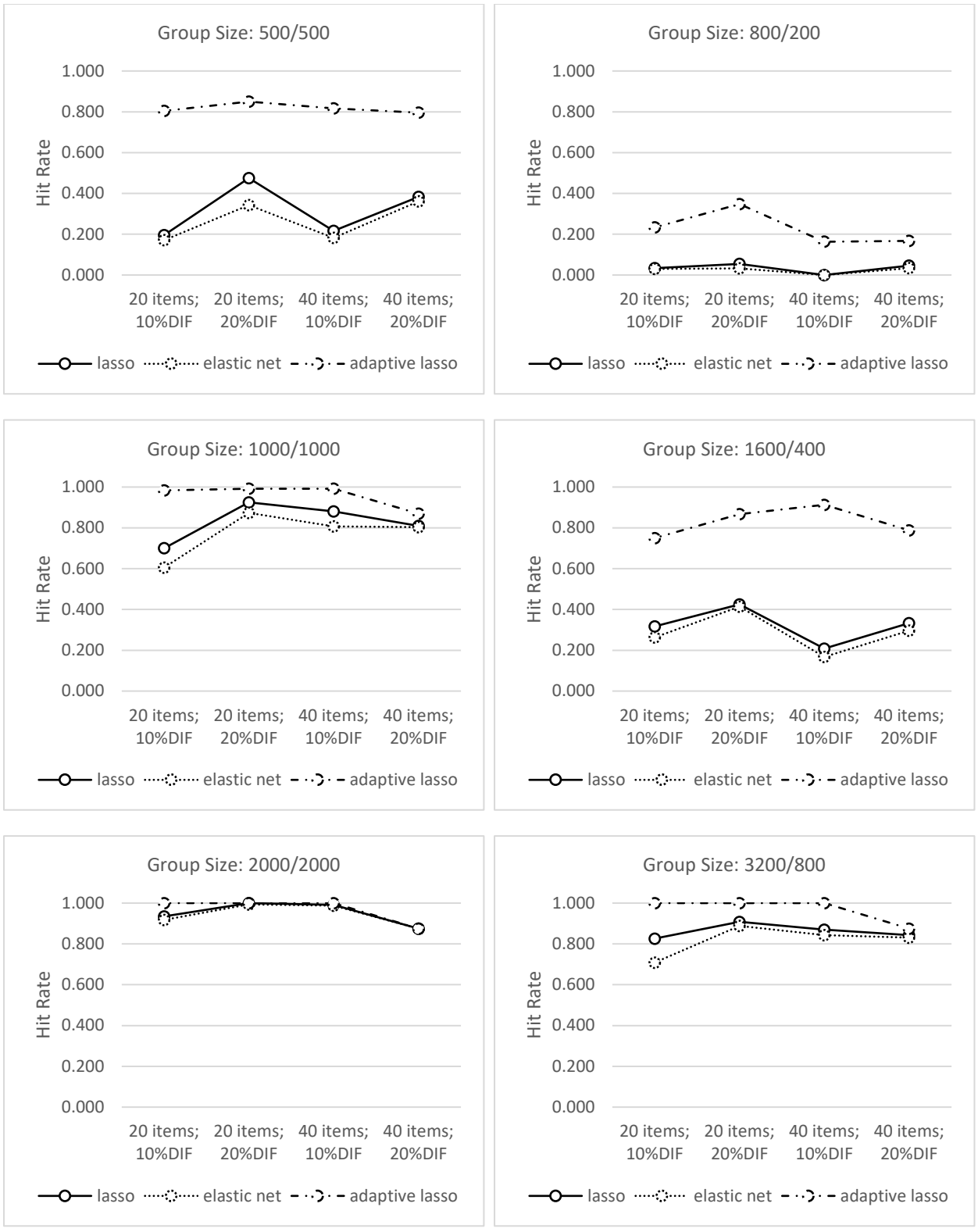

Figure 4.7 Hit Rates for Three Penalty Functions under Nonuniform DIF Conditions with *a*-DIF Magnitude Equal to 1.0

Figure 4.8 False Alarm Rates for Three Penalty Functions under Nonuniform DIF Conditions with *a*-DIF Magnitude Equal to 0.5

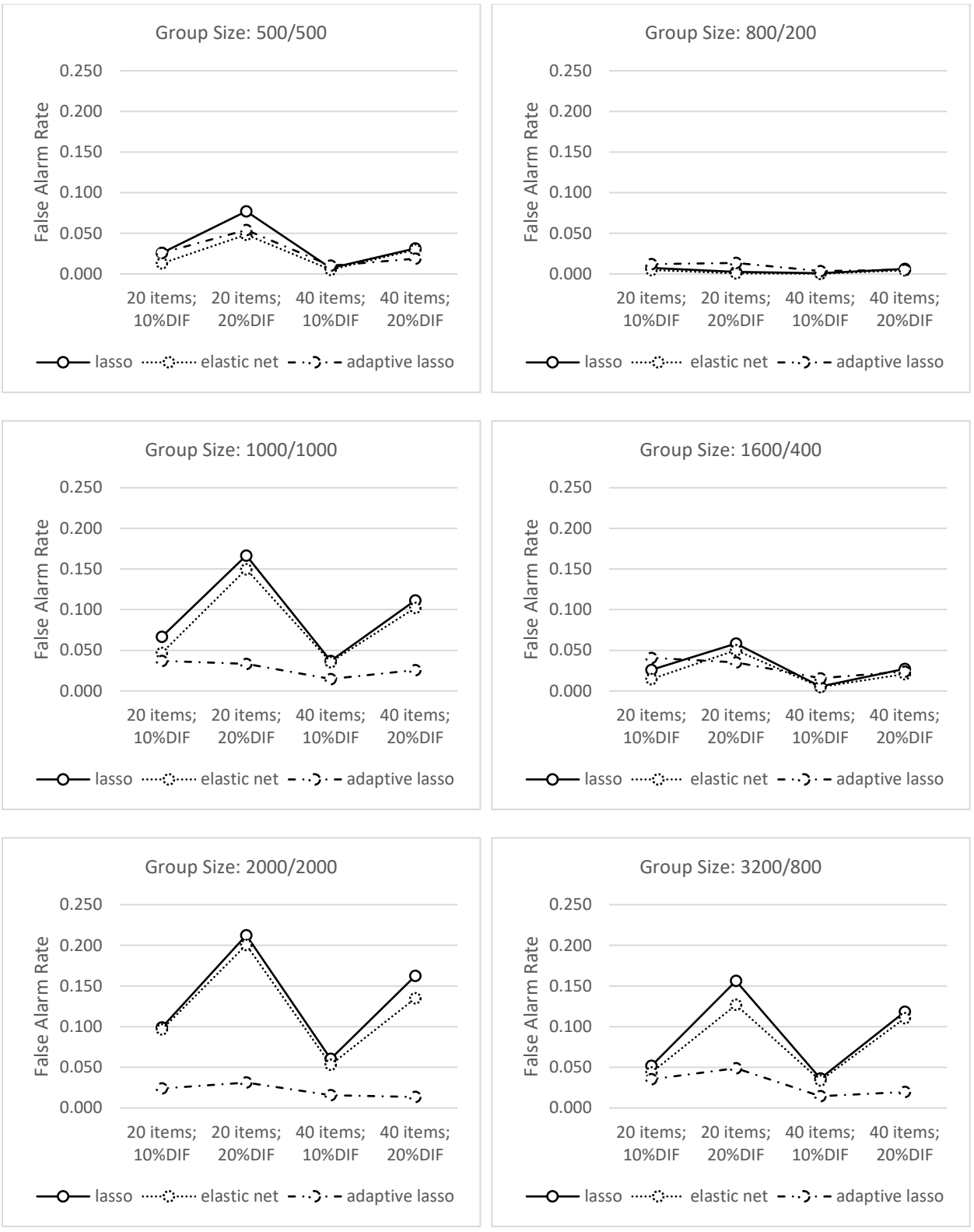

Figure 4.9 False Alarm Rates for Three Penalty Functions under Nonuniform DIF Conditions with *a*-DIF Magnitude Equal to 1.0

Figure 4.10 shows the hit rates and false alarm rates for the 2PL- and 3PL-generated datasets under the same nonuniform DIF conditions with *a*-DIF only. When the ability parameters are estimated using an incompatible IRT model, the average differences in hit rates for the lasso, the elastic net, and the adaptive lasso are decreased by 0.273, 0.253, and 0.413 respectively. Correspondingly, the average false alarm rates are decreased by 0.030, 0.026, and 0.009 for these methods. The results indicate that when inaccurate person's ability estimates are used in the regularized DIF detection model, no matter which penalty function is used in estimation, the model is not able to correctly identify nonuniform DIF items under most conditions.

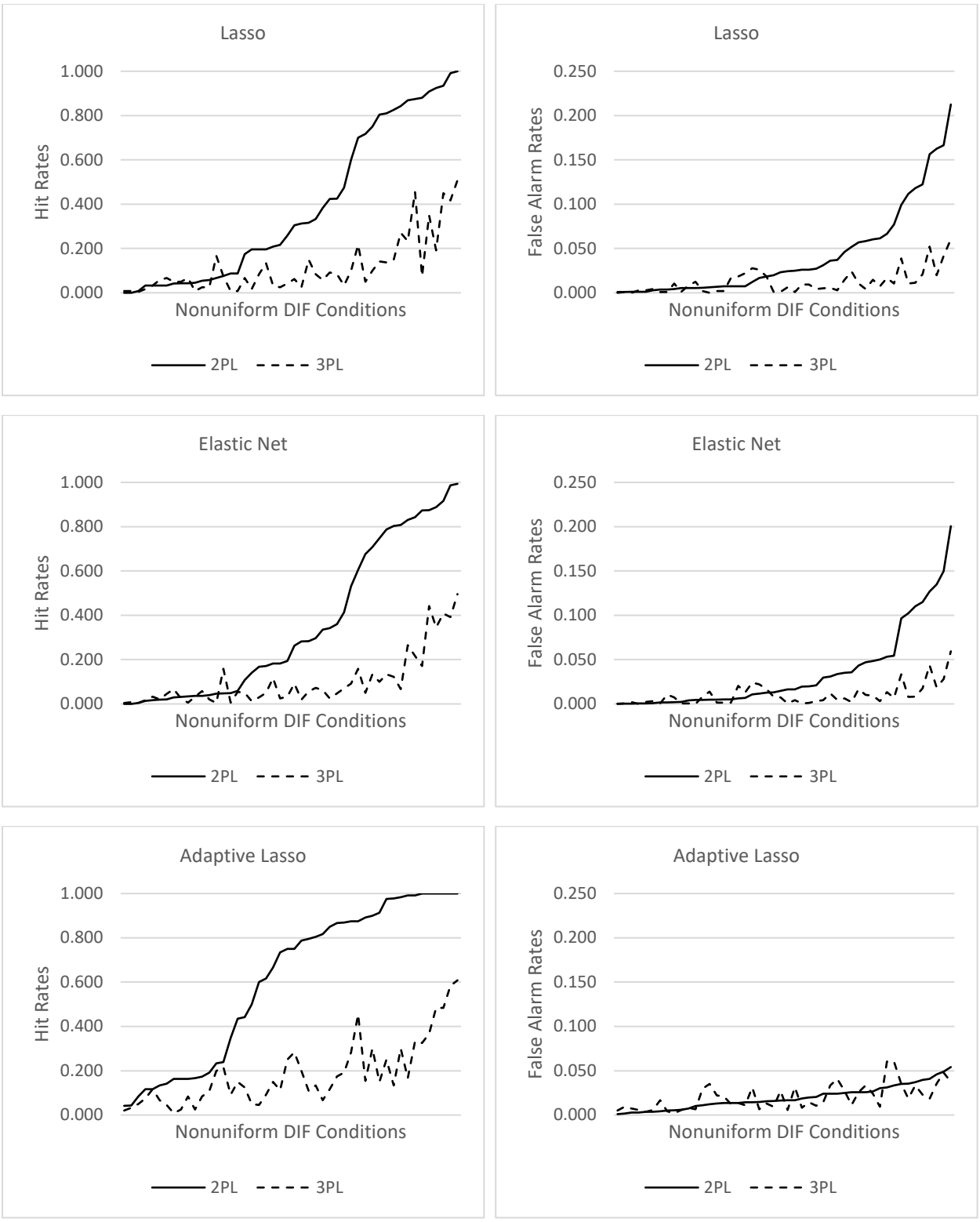

Figure 4.10 Comparison of Hit Rates and False Alarm Rates for 2PL- and 3PL-generated Datasets under 48 Nonuniform DIF Conditions with *a*-DIF only

4.1.2.2 Nonuniform DIF Conditions with Drift on the Discrimination and Difficulty Parameters

Figures 4.11 and 4.13 show the hit rates and false alarm rates using three methodologies with *a*-DIF magnitude equal to 0.5 and *b*-DIF magnitude equal to 0.4, while Figures 4.12 and 4.14 show the hit rates and false alarm rates with *a*-DIF magnitude equal to 1.0 and *b*-DIF magnitude equal to 0.8. Again, the performance of the adaptive lasso method in detecting nonuniform DIF items is much better than the lasso and the elastic net across all conditions. The lasso slightly outperforms the elastic net and the difference in the average hit rates between these two methods is smaller than 0.02.

For most conditions, the hit rates are higher for larger sample sizes, balanced groups, and larger DIF magnitudes. Specifically, the average hit rates are 0.711, 0.413, and 0.168 when there are 4000, 2000, and 1000 examinees correspondingly; the average hit rates are 0.578 and 0.284 for balanced and unbalanced groups; moreover, the average hit rate is 0.467 for large DIF magnitudes according to Figure 4.12 and is 0.395 for small DIF magnitudes according to Figure 4.11. However, it is difficult to tell the relationship between test length and hit rate, and the average hit rates for 20-item and 40-item tests are 0.432 and 0.429 respectively. Also, the impact of percentage of DIF items on hit rates is unclear for nonuniform DIF conditions with both *a*-DIF and *b*-DIF.

According to Figures 4.13 and 4.14, the average false alarm rates using the lasso, the elastic net and the adaptive lasso in regularized DIF detection models are 0.043, 0.036 and 0.024 correspondingly, which are similar to the results in Section 4.1.2.1. We can see that the adaptive lasso has well-controlled false alarm rates across most conditions and the worst false alarm rate is 0.104 (see the left-bottom subplot in Figure 4.14).

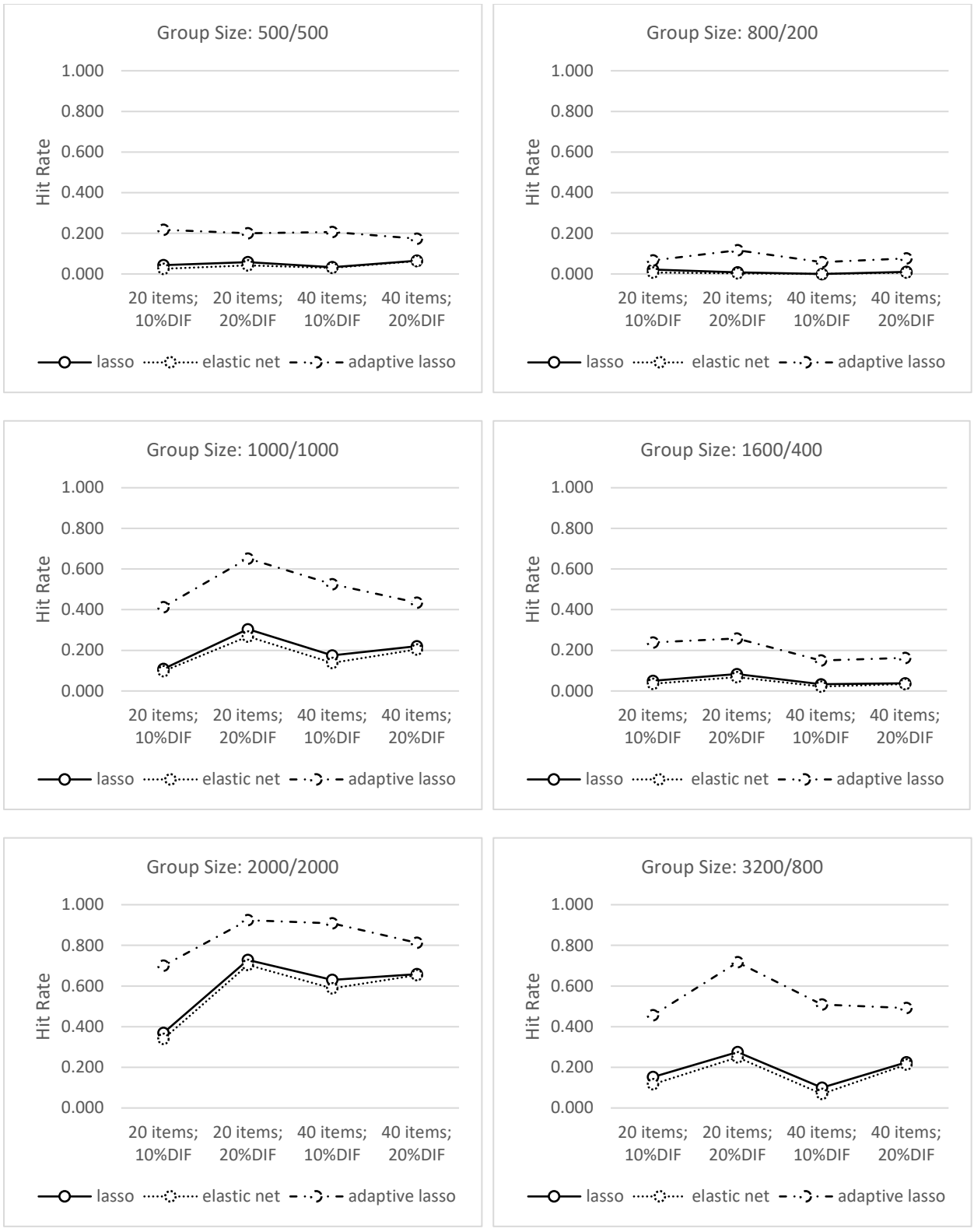

Figure 4.11 Hit Rates for Three Penalty Functions under Nonuniform DIF Conditions with *a*-DIF Magnitude Equal to 0.5 and *b*-DIF Magnitude Equal to 0.4

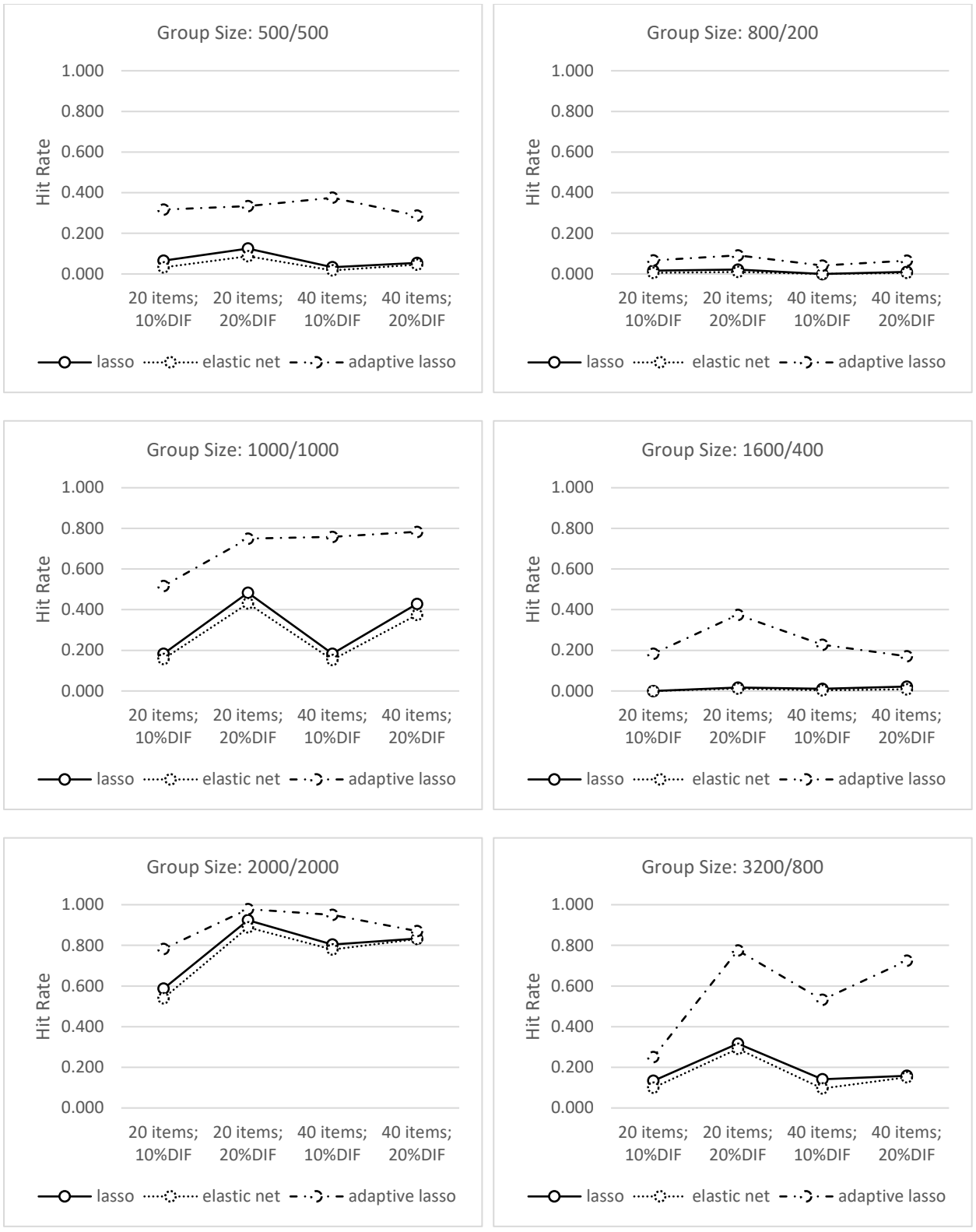

Figure 4.12 Hit Rates for Three Penalty Functions under Nonuniform DIF Conditions with *a*-DIF Magnitude Equal to 1.0 and *b*-DIF Magnitude Equal to 0.8

Figure 4.13 False Alarm Rates for Three Penalty Functions under Nonuniform DIF Conditions with *a*-DIF Magnitude Equal to 0.5 and *b*-DIF Magnitude Equal to 0.4

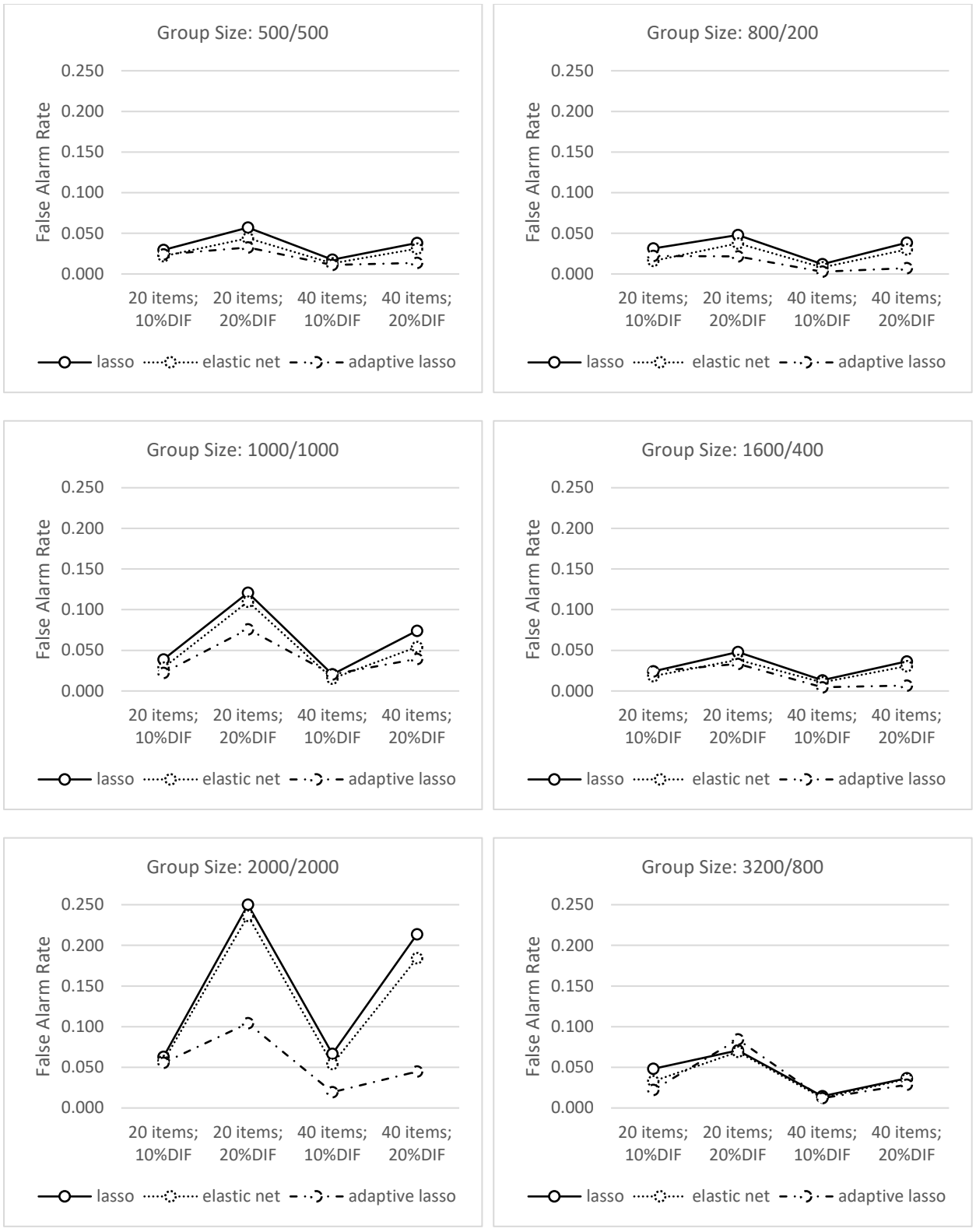

Figure 4.14 False Alarm Rates for Three Penalty Functions under Nonuniform DIF Conditions with *a*-DIF Magnitude Equal to 1.0 and *b*-DIF Magnitude Equal to 0.8

The results demonstrated in Figure 4.15 are similar to Figure 4.10, indicating that the regularized logistic regression model is not robust to biased ability estimates so that it is not able to correctly identify the nonuniform DIF items. Specifically, the average differences in hit rates for the lasso, the elastic net, and the adaptive lasso between 2PL and 3PL scenarios are 0.148, 0.139, and 0.373, while the average differences in false alarm rates are 0.023, 0.019, and 0.016 correspondingly. In addition, we can see that the adaptive lasso penalty is most sensitive to biased ability estimates compared to the other two penalty functions.

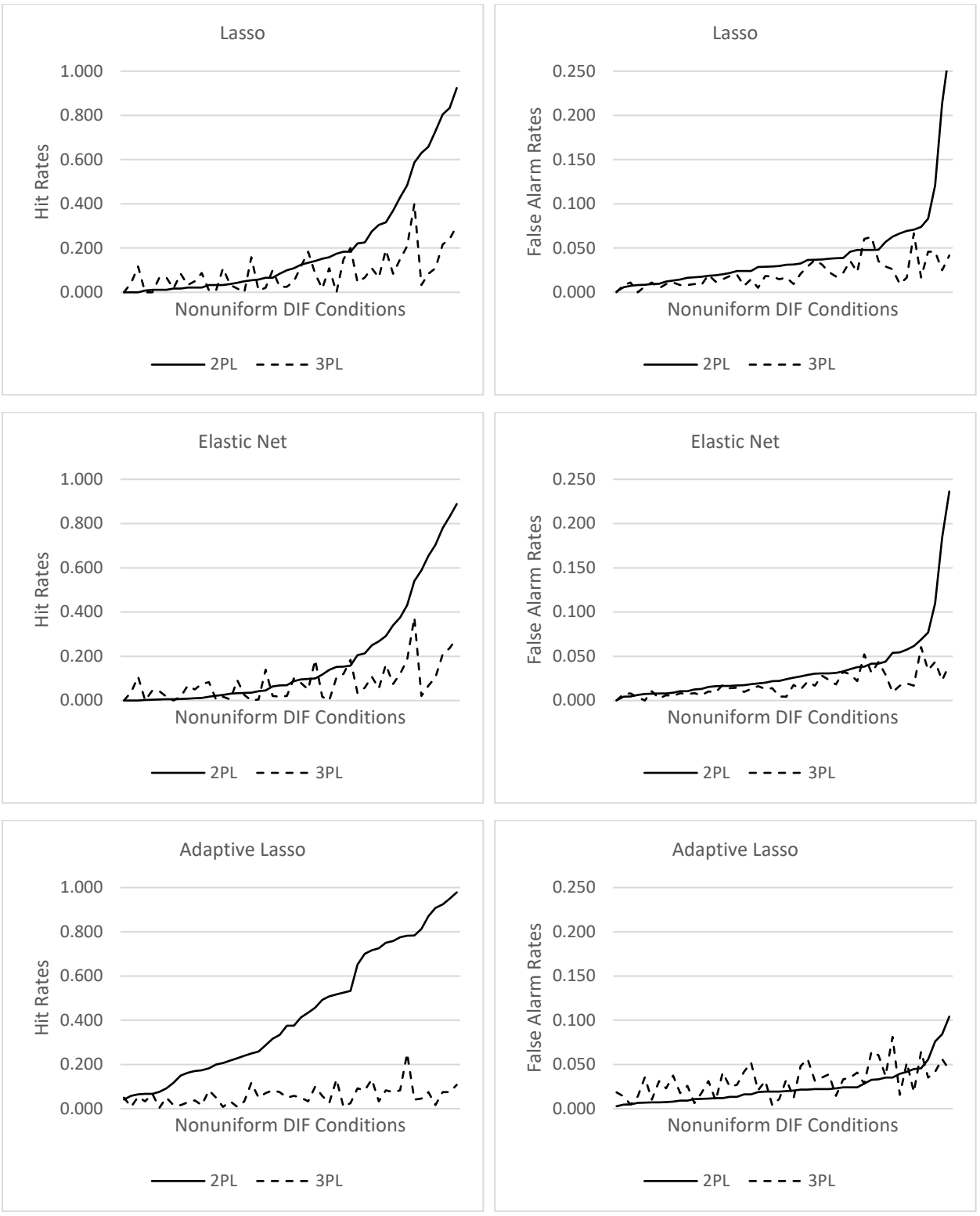

Figure 4.15 Comparison of Hit Rates and False Alarm Rates using 2PL- and 3PL-generated Datasets under 48 Nonuniform DIF Conditions with both *a*-DIF and *b*-DIF

4.2 Simulation Study 2

The second simulation study compares three different DIF detection methods—the regularization method using the adaptive lasso penalty (we called it the adaptive lasso in this section), the logistic regression method, and the likelihood ratio test, aiming to answer the third research question—what are the differences between the proposed method and other existing techniques in terms of their DIF detection performance. As mentioned in Section 3.2.2, the average ROC curves are used to compare these methods. The results are presented according to uniform and nonuniform DIF conditions in the following sections.

4.2.1 Uniform DIF Conditions

Figure 4.16 summarizes the average ROC curves for three DIF detection methods under two different uniform DIF conditions. We can see that controlling for the false alarm rate, the larger the hit rate, the more powerful the method is in identifying true DIF items correctly. In other words, the larger the area under the ROC curve, the better the methodology is in discriminating between DIF and non-DIF items (Magis et al., 2015). Therefore, the adaptive lasso is the most powerful method in detecting uniform DIF items since its ROC curve is always above the other two curves, while the logistic regression method is the least powerful method. Particularly, when the false alarm rate is set to 0.05, the corresponding hit rates of the adaptive lasso are about 0.9 and 0.8 for balanced and unbalanced groups respectively.

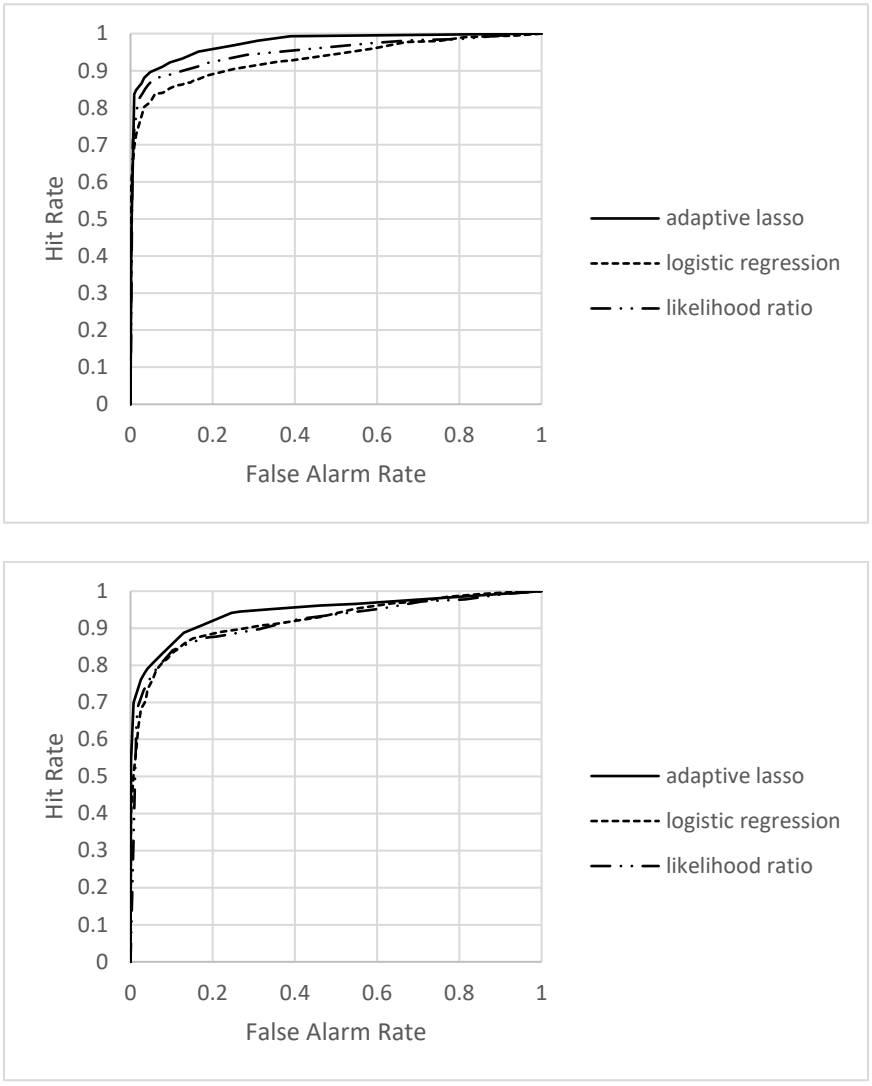

Figure 4.16 Average ROC Curves for Adaptive Lasso, Logistic Regression, and Likelihood Ratio dealing with Tests of 40 Items, 20% Uniform DIF Items with *b*-DIF Magnitude Equal to 0.8, and Sample Size of 2000 (top: group sizes of 1000; bottom: group sizes of 1600 and 400)

4.2.2 Nonuniform DIF Conditions

Figure 4.17 shows the average ROC curves under two nonuniform DIF conditions with *a*-DIF only. According to the areas under the ROC curves, logistic regression and likelihood ratio test have slightly better overall performance in detecting nonuniform DIF items than the adaptive lasso. However, when the false alarm rates are controlled at 0.05,

the top subplot shows that the hit rates of all three methods are greater than 0.9, and the bottom subplot shows that the hit rates of these methods are close to each other and are between 0.85 and 0.9. Therefore, all three DIF methods have similar performance in detecting nonuniform DIF items, particularly when there only exists *a*-DIF.

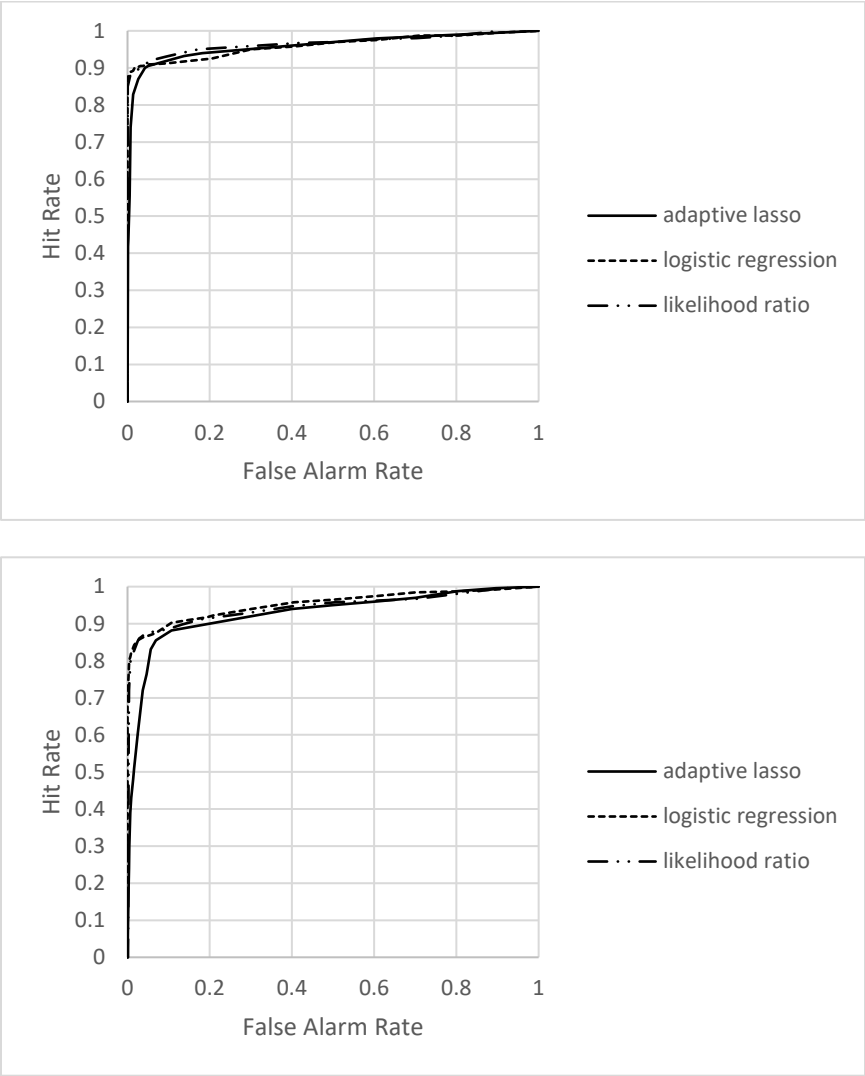

Figure 4.17 Average ROC Curves for Adaptive Lasso, Logistic Regression, and Likelihood Ratio dealing with Tests of 40 Items, 20% Nonuniform DIF Items with *a*-DIF Magnitude Equal to 1.0, and Sample Size of 2000 (top: group sizes of 1000; bottom: group sizes of 1600 and 400)

However, Figure 4.18 shows different findings where logistic regression and likelihood ratio methods have similar performance and are much better than the adaptive lasso in detecting nonuniform DIF items with both *a*-DIF and *b*-DIF. Especially for the adaptive lasso, if the reference and focal group sizes are different, the hit rate is only less than 0.5 when the false alarm rate is set to 0.05. But for equal group sizes, the hit rate of 0.75 is acceptable.

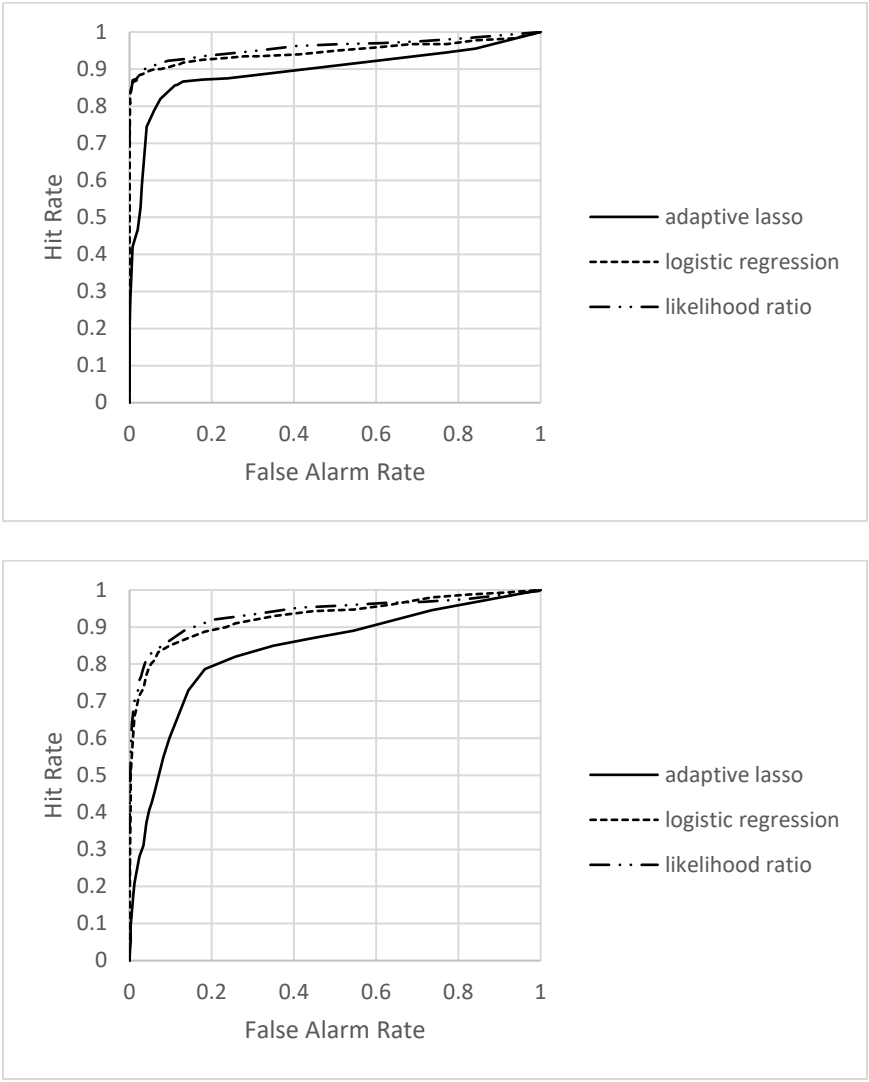

Figure 4.18 Average ROC Curves for Adaptive Lasso, Logistic Regression, and Likelihood Ratio dealing with Tests of 40 items, 20% Nonuniform DIF Items with *a*-DIF Magnitude Equal to 1.0 and *b*-DIF Magnitude Equal to 0.8, and Sample Size of 2000 (top: group sizes of 1000; bottom: group sizes of 1600 and 400)

4.3 Empirical Data Analyses

Three DIF detection methods—the regularization method using the adaptive lasso penalty, the logistic regression method, and the likelihood ratio test, were not only studied in the simulation study but also applied to real datasets, the TIMSS 2015 data.

As stated in Section 3.1.1, based on Equation 3.1, Item *i* shows uniform DIF when $\hat{\beta}_{2i} \neq 0$ and $\hat{\beta}_{3i} = 0$, and item *i* shows nonuniform DIF when $\hat{\beta}_{3i} \neq 0$ regardless of the value of β_{2i}. In terms of the logistic regression and likelihood ratio test, two significance levels of 0.01 and 0.05 were selected. The analysis results on mathematics and science data are given in Section 4.3.1 and Section 4.3.2 correspondingly.

4.3.1 Empirical Dataset 1

Table 4.1 summarizes the DIF analysis results of 18 mathematics items for selected students from the United States. Item 02 and Item 03 are detected as DIF items displaying uniform DIF using adaptive lasso, and the corresponding parameter estimates according to Equation 3.1 are $\hat{\beta}_{22} = 0.217$ and $\hat{\beta}_{23} = 0.132$. When the significance level is set to 0.01, the same items are flagged using both the logistic regression (Item 02: $p = 0.009$; Item 03: $p = 0.004$) and likelihood ratio test (Item 02: $p = 0.010$; Item 03: $p = 0.004$).

However, when 0.05 is chosen as the significance level, more items are found to be problematic using the logistic regression model, and they are Item 08 ($p = 0.045$), Item 10 ($p = 0.029$), and Item 16 ($p = 0.024$). Additionally, Item 03 displays nonuniform DIF and the corresponding p-value is 0.042. For the other two approaches, the DIF detection results remain the same, that is, only Items 02 and 03 are identified as uniform DIF items.

Table 4.1 DIF Detection Results of TIMSS 2015 Mathematics Data

Item	Adaptive Lasso	Logistic Regression		Likelihood Ratio	
		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
01	No DIF	No DIF	No DIF	No DIF	No DIF
02	Uniform DIF	Uniform DIF	Uniform DIF	Uniform DIF	Uniform DIF
03	Uniform DIF	Uniform DIF	Nonuniform DIF	Uniform DIF	Uniform DIF
04	No DIF	No DIF	No DIF	No DIF	No DIF
05	No DIF	No DIF	No DIF	No DIF	No DIF
06	No DIF	No DIF	No DIF	No DIF	No DIF
07	No DIF	No DIF	No DIF	No DIF	No DIF
08	No DIF	No DIF	Uniform DIF	No DIF	No DIF
09	No DIF	No DIF	No DIF	No DIF	No DIF
10	No DIF	No DIF	Uniform DIF	No DIF	No DIF
11	No DIF	No DIF	No DIF	No DIF	No DIF
12	No DIF	No DIF	No DIF	No DIF	No DIF
13	No DIF	No DIF	No DIF	No DIF	No DIF
14	No DIF	No DIF	No DIF	No DIF	No DIF
15	No DIF	No DIF	No DIF	No DIF	No DIF
16	No DIF	No DIF	Uniform DIF	No DIF	No DIF
17	No DIF	No DIF	No DIF	No DIF	No DIF
18	No DIF	No DIF	No DIF	No DIF	No DIF

4.3.2 Empirical Dataset 2

Table 4.2 demonstrates the DIF analysis results of 20 science items for selected students from the United States. We can see that when the statistical significance level is set to 0.01, only Item 03 is flagged as a uniform DIF item by the logistic regression ($p = 0.008$) and likelihood ratio test ($p = 0.005$). Item 03 is also flagged as a uniform DIF items using the adaptive lasso method, where the parameter estimate is $\hat{\beta}_{23} = 0.005$. Additionally, Item 15 is detected as a nonuniform DIF item using the adaptive lasso with $\hat{\beta}_{3,15} = 0.037$. When the significance level is set to 0.05, the results are different between the logistic regression and the likelihood ratio test. For the logistic regression method, in addition to Item 03, Item 17 and Item 20 are flagged as uniform DIF items with p-values

equal to 0.029 and 0.021 correspondingly, while Item 07 ($p = 0.044$), Item 15 ($p = 0.014$) and Item 19 ($p = 0.043$) are flagged as nonuniform DIF items. In terms of the likelihood ratio test, Item 03 still displays uniform DIF, and Item 20 also shows uniform DIF with $p = 0.020$. Apart from these two items, Item 16 and Item 17 display nonuniform DIF with p-values equal to 0.009 and 0.039 respectively.

Table 4.2 DIF Detection Results of TIMSS 2015 Science Data

Item	**Adaptive Lasso**	**Logistic Regression**		**Likelihood Ratio**	
		$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
01	No DIF	No DIF	No DIF	No DIF	No DIF
02	No DIF	No DIF	No DIF	No DIF	No DIF
03	Uniform DIF	Uniform DIF	Uniform DIF	Uniform DIF	Uniform DIF
04	No DIF	No DIF	No DIF	No DIF	No DIF
05	No DIF	No DIF	No DIF	No DIF	No DIF
06	No DIF	No DIF	No DIF	No DIF	No DIF
07	No DIF	No DIF	Nonuniform DIF	No DIF	No DIF
08	No DIF	No DIF	No DIF	No DIF	No DIF
09	No DIF	No DIF	No DIF	No DIF	No DIF
10	No DIF	No DIF	No DIF	No DIF	No DIF
11	No DIF	No DIF	No DIF	No DIF	No DIF
12	No DIF	No DIF	No DIF	No DIF	No DIF
13	No DIF	No DIF	No DIF	No DIF	No DIF
14	No DIF	No DIF	No DIF	No DIF	No DIF
15	Nonuniform DIF	No DIF	Nonuniform DIF	No DIF	No DIF
16	No DIF	No DIF	No DIF	No DIF	Nonuniform DIF
17	No DIF	No DIF	Uniform DIF	No DIF	Nonuniform DIF
18	No DIF	No DIF	No DIF	No DIF	No DIF
19	No DIF	No DIF	Nonuniform DIF	No DIF	No DIF
20	No DIF	No DIF	Uniform DIF	No DIF	Uniform DIF

CHAPTER 5. CONCLUSIONS

5.1 Summary of Findings

In this study, a new DIF detection model based on the regularized logistic regression model was proposed. Unlike many traditional DIF detection approaches, the proposed DIF model allows all test items to be examined for DIF simultaneously, and the DIF analysis is no longer conducted at the item level. As a result, the strict assumption for DIF analysis that all items except the studied item are supposed to be invariant across groups can be avoided, and multiple testing—a major threat to test validity—is no longer a problem.

In order to evaluate the performance of the proposed DIF detection model, comprehensive simulation studies and empirical data analyses were conducted. The first simulation study examined the operating characteristics including hit rates and false alarm rates using three kinds of penalty functions in the regularized logistic regression model under various manipulated conditions. Moreover, the second simulation study compared the performance of the regularized DIF detection model to two commonly used DIF detection methods including the logistic regression method and the likelihood ratio test, and these three methods were also applied to analyzing real datasets.

5.1.1 Research Question One

In order to address the first research question, two preliminary analyses were conducted beforehand to (1) choose a model selection technique that can select the best DIF detection model in terms of their operating characteristics and (2) choose an appropriate ability measure as a proxy for θ_p. Specifically, although under some

simulated conditions AIC and CV are able to identify more true DIF items than BIC, at the same time much more non-DIF items are flagged as DIF items incorrectly. Also, most of the DIF items flagged by AIC and CV only have a minimal impact on parameter estimation and are actually less of concern in practice. Therefore, by taking both hit rates and false alarm rates into consideration, BIC outperforms AIC and CV since it is able to not only flag high-impact DIF items, but also prevent over-identification of DIF items with few false alarms. BIC is considered as a more conservative criterion compared to AIC and CV in selecting the optimal DIF model. It is because typically AIC and CV tend to find a model that gives the best prediction without assuming any of the models are correct. All candidate models are approximations for truth according to AIC or CV, and the truth tends to be high dimensional since a more complex model gives a better fit to the data. On the other hand, BIC tends to find a model that is more likely to be true by assuming one of the candidate models being true. Therefore, AIC and CV always have a chance of selecting a too complex model where only a subset of the selected variables are in the true model, and BIC has a larger chance of choosing a simple model compared to AIC and CV. In DIF detection, selecting the variables that present DIF is much more important than precisely predicting the responses, which explains why BIC outperforms AIC and CV in terms of the operating characteristics.

In the second preliminary study, the DIF detection performance of two types of ability measures mentioned in previous research including the observed total score and the IRT ability estimate were compared. The results indicate that using the IRT ability estimate always has a better DIF detection performance compared to using the observed score. Generally, the inconsistency between the observed total score and the ability

parameter estimates is the result of letting the item discrimination parameters vary among items. Since the definition and classification of DIF were specified in the IRT framework, it is reasonable that using the IRT ability estimate in the proposed model yields better DIF detection results. However, additional analyses also indicate that this conclusion is tenable only when the IRT ability estimates are accurate, which will be discussed in this section later.

The first simulation study was conducted based on the findings of two preliminary analyses, where the IRT ability estimate was used as a proxy for θ_p in the proposed DIF detection model and BIC was used as the model selection criterion. The answer to the first research question is that the adaptive lasso penalty has the best performance with the highest hit rates as well as the lowest and well-controlled false alarm rates among all three penalty functions under most conditions in both uniform and nonuniform DIF detection, while the elastic net penalty obtains the worst DIF detection performance. This indicates that using the adaptive extension of lasso that introduces weights to the penalty on each coefficient in the lasso procedure can improve the operating characteristics significantly, but adding the $\ell 2$ penalty term does not improve the model's performance in detecting either type of DIF. In other words, although the quadratic part of the elastic net penalty encourages grouping effect that allows for grouped selection of correlated variables, it does not improve the model's ability in detecting more true DIF items. Moreover, we may encounter problems when using the traditional lasso such as small nonzero parameters cannot be consistently detected and large non-zero parameters are too small resulting in large bias. Using the adaptive lasso where the $\ell 1$ norms in the penalty are weighted by data-dependent weights allows a relatively higher penalty for small

coefficients and a relatively lower penalty for large coefficient (Huang, Ma, & Zhang, 2008), which helps reduce the estimation bias and improve variable selection accuracy compared to the traditional lasso. Therefore, the adaptive lasso generally yields larger hit rates compared to other approaches. In terms of the false alarm rates of the adaptive lasso approach, they are well-controlled at 0.05 under most conditions. It is expected because not only BIC tends to select a simpler model, but also the adaptive lasso tends to reduce the number of unimportant parameters (e.g., the parameters with small coefficients), making it less probable to incorrectly flag non-DIF items as DIF items.

However, the results also show that the performance of the proposed DIF detection model is dependent on the latent ability estimates—the proposed DIF detection model becomes less powerful when the ability parameters are inaccurately estimated. It is because if the ability estimates are incorrect, there will be additional inaccuracies caused by observation errors in the DIF detection model. According to Equation 3.1, we can see that half of the variables in the model are related to person's abilities. Therefore, using an appropriate ability measure is a crucial prerequisite for a powerful DIF detection model.

5.1.2 Research Question Two

The first simulation also addresses the second research question. Six manipulated factors including DIF type, DIF magnitude, percentage of DIF items, test length, sample size, and sample size ratio were considered in the study. Since we have already known that the adaptive lasso has a better performance than the other two penalties, and also the false alarm rates are too small to compare under many scenarios, the following discussion focuses on how the hit rates are impacted by each manipulated factor using the adaptive lasso approach.

The proposed DIF detection model is more powerful in identifying uniform DIF items than nonuniform DIF items. As expected, the DIF model has a better performance in detecting items with large DIF magnitudes (e.g., b-DIF magnitude equal to 0.8, and a-DIF magnitude equal to 1.0). It is consistent with the findings in previous studies, since larger DIF magnitudes are useful to quantify the effect of DIF magnitude on power and type I error rate, while smaller DIF magnitudes usually represent negligible DIF since they have minimal impact on estimating model parameters (Wells et al., 2002; Magis & De Boeck, 2012). Moreover, we can see that the patterns of hit rates become very different under nonuniform DIF conditions (see Section 4.1.2) compared to uniform DIF conditions (see Section 4.1.1). One potential explanation is that when the nonuniform DIF effects are not strong enough, the proposed DIF model tends to flag those nonuniform DIF items as uniform DIF items. Although not presented, the simulation results show that for some nonuniform DIF items, β_{2i} is nonzero even if β_{3i} equals zero.

Additionally, the hit rates increase as the sample size increases while controlling for other manipulated factors, since typically more precise model parameter estimates require larger sample size. On the other hand, when the sample size increases, each group contains more participants so that the estimates of persons' abilities are likely to be more accurate, thus the proposed DIF model will be more powerful correspondingly.

Besides, the proposed DIF model is less powerful in detecting true DIF items when the reference and focal group sizes are unbalanced. This is because when the reference and focal group size ratio is extremely unbalanced (e.g., group size ratio of 4:1), it becomes much more difficult to precisely estimate the focal group participants' abilities, which affects the model's ability to detect DIF items. As discussed, the

proposed model has a better DIF detection performance with more accurate ability estimates, and this explains why balanced groups are more appreciated.

Furthermore, in uniform DIF detection, the hit rates increase when the total number of items decreases, or when there are fewer DIF items on a test. Specifically, according to Equation 3.2, there are in total $4I$ parameters that need to be estimated, where I represents the total number of test items. When the number of items increases from 20 to 40, the number of estimated parameters increases dramatically from 80 to 160. Adding too many variables to the model may lead to overfitting issues, which easily yields misleading coefficient estimates, thus affects the variable selection results. However, under nonuniform DIF conditions, the patterns become different and it is hard to tell the relationship between test length and hit rates, as well as the relationship between percentages of DIF items and hit rates.

5.1.3 Research Question Three

The third research question is addressed by the second simulation study. According to the ROC curves in Figures 4.16-4.18, the regularized DIF detection model using the adaptive lasso penalty outperforms the traditional logistic regression and likelihood ratio test in uniform DIF detection. And under nonuniform DIF conditions, two traditional approaches demonstrate similar or slightly better DIF detection performance.

These three methods were also applied to two real datasets from TIMSS 2015. Specifically, when the statistical significance level set to 0.01 for the logistic regression and likelihood ratio methods, the DIF detection results of these methods are similar to

each other. However, when the significance level is 0.05, the results are different, especially when evaluating the science items.

Generally, the empirical analyses results are consistent with the simulation study—the adaptive lasso approach is powerful in detecting uniform DIF items, but for those items with small DIF effects, the adaptive lasso tends to treat them as non-DIF items. Since the proposed DIF model is a very new approach, currently it is recommended to use other traditional DIF detection methods at the same time when analyzing empirical datasets and make a final decision by comparing the results of these approaches, until more simulation studies have been done.

5.2 Limitations and Future Research

Since the proposed regularized logistic regression DIF model is the first attempt to detect both uniform and nonuniform DIF using regularization techniques, it still has some features that can be further improved.

First, this study only considered the three most commonly used model selection techniques including AIC, BIC, and CV. Although BIC outperforms AIC and CV by considering both hit rates and false alarm rates, it has a conservative bias tending towards false negative errors, that is, fail to detect true DIF items. Therefore, other model selection criterion, for example, the weighted average information criterion (Wu, Chen, & Yan, 2013; Magis et al., 2015) which is an intermediate criterion between AIC and BIC, can be studied in future research.

Second, as shown in Section 3.1.4.2, when the IRT ability estimates are precise, using these estimates yields much better operating characteristics compared to using the

total scores. However, in certain scenarios, the estimates of ability parameters might be biased (e.g., using the 2PL model in parameter estimation when there exist guessing behaviors in a test). Although the simulations indicate that the DIF detection performance of the proposed model relied on the accuracy of ability measures, additional studies are recommended to examine to what extent the DIF model is able to tolerate biased ability estimates, and to investigate if other variables can be used as a proxy for person's latent ability in addition to the IRT ability estimate and the observed total score.

Third, this study first applied the elastic net and the adaptive lasso penalty to DIF detection. The simulation results show that the adaptive lasso penalty has a great performance in detecting DIF items especially uniform DIF items. However, other penalty functions can be studied, for example, the group lasso penalty (Meier, Van De Geer, & Bühlmann, 2008) which is another extension of the lasso, handling variable selection on groups of variables in generalized linear models.

Moreover, the design for the simulation study and the empirical data analyses can be further improved. Specifically, for the simulation study, additional manipulated factors (e.g., unmatched ability distributions for the reference and focal groups) or more levels of each factor (e.g., small, medium and large DIF magnitudes) can be studied. It is also worth investigating other reasonable ranges for sample size ratio as well as percentage of DIF items on a test. For the empirical data analyses, instead of the achievement data, other types of test data such as personality data can be studied.

Finally, more variables can be added to the current DIF model in order to better quantify the nonuniform DIF effects. The DIF model can also be modified and extended to detect DIF among multiple groups, as well as assess DIF effects in polytomous items.

REFERENCES

Acar, T. (2012). Determination of a differential item functioning procedure using the hierarchical generalized linear model: a comparison study with logistic regression and likelihood ratio procedure. *SAGE Open, 2*(1), 1-8.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, *95*(3), 631-636.

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

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 32*(2), 283-301.

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26*(1), 31-44.

Angoff, W. H., & Ford, S. F. (1973). Item‐race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95-105.

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*(3), 870-897.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic model. *ETS Research Report Series, 1981*(1), 1-8.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 57*(1), 289-300.

Berk, R. A. (1982). *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: Johns Hopkins University Press.

Binici, S. (2007). *Random effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: a comparison of estimation methods* (Doctoral dissertation). Retrieved from http://purl.flvc.org/fsu/fd/FSU_migr_etd-3755

Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision-making problems. Series Report No. 58-16. Project No. 7755-23. USAF School of Aviation Medicine, Randolph Air Force Base, TX.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-479). Charlotte, NC: Information Age Publishing.

References

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for dichotomously scored items. *Psychometrika, 35*(2), 179-197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement*, *37*(4), 307-327.

Bonferroni, C. E. (1936). Statistical theory of classification and calculation of the probability. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, *8*, 3-62.

Cardall, C., & Coffman, W. E. (1964). *A method for comparing the performance of different groups on the items in a test.* Research Bulletin RB-64-61. Princeton, NJ: Educational Testing Service.

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

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

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*(4), 269-279.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15-26.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Orlando, FL: Holt, Rinehart and Winston.

Cubedo, M., & Oller, J. M. (2002). Hypothesis testing: a model selection approach. *Journal of Statistical Planning and Inference*, *108*(2002), 3-21.

De Mol, C., De Vito, E., & Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity*, *25*(2), 201-230.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39*(1), 1-38.

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

References

Dorans, N. J., & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*(1), 77-90.

Du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22.

González de San Román, A., & de la Rica, S. (2016). Gender gaps in PISA test scores: the impact of social norms and the mother's transmission of role attitudes. *Estudios de Economía Aplicada, 34*(1), 79-108.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Han, K. T. (2013). Item response models used within WinGen. Retrieved from https://www.umass.edu/remp/software/simcata/wingen/modelsF.html

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics, 43*(2), 417-423.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Houts, C. R., & Cai, L. (2016). *flexMIRT User's Manual Version 3.5: Flexible Multilevel Multidimensional Item Analysis and Test Scoring.* Chapel Hill, NC: Vector Psychometric Group.

Huang, J., Ma, S., & Zhang, C. H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica, 18*(4), 1603-1618.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York, NY: Springer.

References

Jia, J., & Yu, B. (2010). On model selection consistency of the elastic net when p>>n. *Statistica Sinica, 20*(2), 595-611.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling, 16*(4), 642-657.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement (Fourth Edition)* (pp. 17–64). Lanham, MD: Rowman & Littlefield Publishers.

Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics, 28*(5), 1356-1378.

Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*(4), 291-312.

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*(3), 217-228.

Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*(3), 458-470.

Koehler, A. B., & Murphree, E. S. (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Applied Statistics, 37*(2), 187-195.

Kopf, J., Zeileis, A., & Strobl, C. (2013). Anchor methods for DIF detection: a comparison of the iterative forward, backward, constant and all-other anchor class (Technical Report 141). Munich, Germany: Department of Statistics, LMU Munich.

Lautenschlager, G. J., Flaherty, V. L., & Park, D. G. (1994). IRT differential item functioning: an examination of ability scale purifications. *Educational and Psychological Measurement, 54*(1), 21-31.

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72*(5), 847-861.

Li, Z. (2014). Power and sample size calculations for logistic regression tests for differential item functioning. *Journal of Educational Measurement*, *51*(4), 441-462.

References

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*(4), 989-1020.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale NJ: Lawrence Erlbaum Associates.

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

Magis, D., Beland, S., & Raiche, G. (2013). difR: Collection of methods to detect dichotomous differential item functioning in psychometrics. Retrieved from http://www2.uaem.mx/r-mirror/web/packages/difR/difR.pdf

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111-135.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement*, *18*(4), 229-248.

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE-Life Sciences Education*, *16*(2), ar19-rm2.

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

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 53-71.

Meinshausen, N., & Bühlmann, P. (2006). Consistent neighborhood selection for sparse high-dimensional graphs with the lasso. *The Annals of Statistics, 34*(3), 1436-1462.

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

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

Mosier, C. I. (1940). Psychophysics and mental test theory: fundamental postulates and elementary theorems. *Psychological Review, 47*(4), 355-366.

Mosier, C. I. (1941). Psychophysics and mental test theory II: the constant process. *Psychological Review, 48*(3), 235-249.

References

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS Advanced 2015 International Results in Advanced Mathematics and Physics.* Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/advanced/

Muraki, E., & Bock, R. D. (1997). *PARSCALE 3: IRT Based Test Scoring and Item Analysis for Graded Items and Rating Scales*. Lincolnwood, IL: Scientific Software International.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

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

Osterlind, S. J. (1983). *Test Item Bias.* Beverly Hills, CA: Sage Group.

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

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value in Health, 18*(1), 25-34.

Raju, N. S. (1998). 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.

R Development Core Team. (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago, IL: University of Chicago Press.

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.

Şahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice, 17*(1n), 321–335.

San Martín, E., González, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*(2), 450-467.

References

Schauberger, G. (2015). *Regularization methods for item response and paired comparison models* (Doctoral dissertation). Retrieved from https://edoc.ub.uni-muenchen.de/19007/1/Schauberger_Gunther.pdf

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*(1), 561-584.

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

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*(1), 117-126.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1), 1–16.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 44-47.

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

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*(1), 53-75.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*(4), 397-412.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21-43.

References

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443-464.

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

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

Wang, W. C., Shih, C. L., & Sun, G. W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement, 72*(4), 687-708.

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

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*(1), 77-87.

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

Wu, M., Adams, R., Wilson, M. R., & Haldane, S. (2007). *ConQuest (Version 2.0) [Computer Software].* Camberwell, England: ACER.

Wu, T. J., Chen, P., & Yan, Y. (2013). The weighted average information criterion for multivariate regression model selection. *Signal Processing*, *93*(1), 49-55.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika, 92*(4), 937-950.

Yuan, M., & Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(2), 143-161.

Zimowski, M. F., Muraki, E., Mislevy, R., & Bock, R. D. (2003). *BILOG-MG [Computer Software].* Mooresville, IN: Scientific Software International.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.