ROBUST SCALE TRNASFORMATION METHODS IN IRT TRUE SCORE EQUATING UNDER COMMON-ITEM NONEQUIVALENT GROUPS DESIGN

A Dissertation Presented to the Faculty of the

Department of Educational, School and Counseling Psychology

University of Missouri

In Partial Fulfillment

of the Requirements for the Doctoral Degree

---

by

YONG HE

Dr. Steven Osterlind, Dissertation Supervisor

MAY, 2013

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

ROBUST SCALE TRNASFORMATION METHODS IN IRT TRUE SCORE EQUATING UNDER COMMON-ITEM NONEQUIVALENT GROUPS DESIGN

presented by **Yong He**,

a candidate for the degree of **Doctor of philosophy**,

and hereby certify that in their opinion it is worthy of acceptance.

_____

Dr. Steven Osterlind

_____

Dr. Ze Wang

_____

Dr. Alex Waigandt

_____

Dr. Roberta Scholes

_____

Dr. Christopher Wikle

ACKNOWLEDGEMENTS

I would like to thank my dissertation supervisor Dr. Steven Osterlind for giving me the opportunity to pursue my interest, and I am grateful for his guidance, patience and understanding during the study and research at the University of Missouri-Columbia.

I am appreciative of my committee members, Drs. Ze Wang, Christopher Wikle, Alex Waigandt, and Roberta Scholes for their assistance and advisement throughout this degree. Particularly, I would like to thank Dr. Wikle for his insightful suggestions by using the least absolute deviation method in my proposed methods.

Special thanks go to Dr. Paul Speckman for his instruction and help with the robust Deming method used in the dissertation.

I also thank the ACT Inc. for providing me with summer internship in 2011. The dissertation is inspired by the research project during the internship. I am grateful for Drs. Zhongmin Cui, Yu Fang and Hanwei Chen at ACT Inc. for mentoring during my internship. Particularly, I wish to express my appreciation to Dr. Zhongmin Cui for his invaluable advice on my dissertation.

I thank my fellow classmates, Ihui Su, Chia-Lin Tsai, Ping Yang, Enkhee Chong, Guohui Wu, and Haiying Wang for their assistance during my study.

Finally, I am truly thankful to my wife, Dijie Liu, and my children, Jennifer (Jinghan) and Zechariah (Zhiyuan), for their love and inspiration.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii
LIST OF TABLES vi
LIST OF FIGURES viii
ABSTRACT x
CHAPTER 1 INTRODUCTION 1
1.1 Introduction 1
1.2 Common-Item Nonequivalent Groups Design 3
1.3 IRT True Score Equating 3
1.4 Scale Transformation 4
1.5 Statement of Problem 5
1.6 Purpose of Study 7
1.7 Significance of the Study 9
1.8 Limitations of the Study 12
1.9 Overview of the Subsequent Chapters 12
CHAPTER 2 LITERATURE REVIEW 14
2.1 Overview of Test Equating 14
2.1.1 Equating Design 15
2.1.1.1 Single Group Design 16
2.1.1.2 Random Groups Design 17
2.1.1.3 Common-Item Nonequivalent Groups Design 17
2.1.2 Equating Methods 18
2.1.2.1 Equating Methods for the SG and RG designs 19
2.1.2.2 CTT-Based Equating Methods for the CINEG Design 20
2.1.2.3 IRT-Based Equating Methods for the CINEG Design 23
2.2 Scale Transformation in IRT Equating 26

2.2.2 The Haebara Method 29
2.2.3 The Stocking-Lord Method 31
2.2.3 Comparison among Scale Transformation Methods 32
2.3 Issues in Common Item 32
2.4 Robustness 37
CHAPTER 3 METHODOLOGY 41
3.1 Overview of Research Questions 41
3.2 Scale transformation using Characteristic Curves 42
3.3 The Proposed Robust Methods 43
3.3.1 Robust Haebara Method 43
3.3.2 The Least Absolute Values 46
3.3.3 Robust Deming Approach 46
3.4 Simulated Data 51
3.4.1 Traditional Scale Transformation Methods with Outliers 51
Table 1. Summary of Conditions of Outlier Simulation (Sections 3.4.1 & 3.4.2) 53
Table 2. Summary of Conditions of Outlier Simulation (Section 3.4.3) 58
3.5 Empirical Data 60
CHAPTER 4 RESULTS 61
4.1 Simulation Studies 61
4.1.1 Traditional Scale Transformation Methods with a Single Outlier 61
4.1.2 Performance of the Proposed Robust Methods of Scale Transformation 77
4.1.3 Comparisons between Robust Methods of Scale Transformation and Outlier Removal 107
4.2 Numeric Illustration with Empirical Data 137

CHAPTER 5 DISCUSSION AND CONCLUSION 145
5.1 Overview of the Study 145
5.2 Summary of Findings 146
5.2.1 Traditional Scale Transformation Methods with a Single Outlier 147
5.2.2 Performance of the Proposed Robust Methods of Scale Transformation 148
5.2.3 Comparisons between Robust Methods of Scale Transformation and Outlier Removal 150
5.2.4 Numeric Illustration using Empirical Data 151
5.2.5 Conclusions 152
5.3 Discussions and Future Directions 152
REFERENCES 156
VITA 165

LIST OF TABLES

Table 1. Summary of Conditions of Outlier Simulation (Sections 3.4.1 & 3.4.2) 53

Table 2. Summary of Conditions of Outlier Simulation (Section 3.4.3) 58

Table 3. Weighted RMSE Statistics for IRT True Score Equating with Traditional Scale Transformation 62

Table 4. Weighted Bias Statistics for IRT True Score Equating with Traditional Scale Transformation 64

Table 5. Weighted Standard Error Statistics for IRT True Score Equating with Traditional Scale Transformation 65

Table 6. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0,1)$ 78

Table 7. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0.25,1.1^2)$ 79

Table 8. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0.5,1.2^2)$ 80

Table 9. Mean (M) and Standard Deviation (SD) of Scale Transformation Coefficients, 100 replications, $\theta \sim N(-0.25,1.1^2)$ 81

Table 10. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(-0.5,1.2^2)$ 82

Table 11a. Weighted RMSE Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 84

Table 11b. Weighted RMSE Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 85

Table 12a. Weighted Bias Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 86

Table 12b. Weighted Bias Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 87

Table 13a. Weighted Standard Error Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 88

Table 13b. Weighted Standard Error Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 89

Table 14. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0,1)$ 109

Table 15. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0.25,1.1^2)$ 110

Table 16. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0.5,1.2^2)$ 111

Table 17. Mean (M) and Standard Deviation (SD) of Scale Transformation Coefficients, 100 replications, $\theta \sim N(-0.25,1.1^2)$ 112

Table 18. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(-0.5,1.2^2)$ 113

Table 19a. Weighted RMSE Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 115

Table 19b. Weighted RMSE Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 116

Table 20a. Weighted Bias Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 117

Table 20b. Weighted Bias Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 118

Table 21a. Weighted Standard Error Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 119

Table 21b. Weighted Standard Error Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods 120

Table 22. Scale Transformation Coefficients for *CBASE* English and Mathematics . 137

LIST OF FIGURES

Figure 1. The RMSE statistics of IRT equating (θ ~ N(0,1)) 66

Figure 2. The RMSE statistics of IRT equating (θ ~ N(0.25,1.1^2)) 67

Figure 3. The RMSE statistics of IRT equating (θ ~ N(0.5,1.2^2)) 68

Figure 4. The Bias statistics of IRT equating (θ ~ N(0,1)) 71

Figure 5. The Bias statistics of IRT equating (θ ~ N(0.25,1.1^2)) 72

Figure 6. The Bias statistics of IRT equating (θ ~ N(0.5,1.2^2)) 73

Figure 7. The Standard Error statistics of IRT equating (θ ~ N(0,1)) 74

Figure 8. The Standard Error statistics of IRT equating (θ ~ N(0.25,1.1^2)) 75

Figure 9. The Standard Error statistics of IRT equating (θ ~ N(0.5,1.2^2)) 76

Figure 10. The RMSE statistics of IRT equating (θ ~ N(0,1)) 90

Figure 11. The RMSE statistics of IRT equating (θ ~ N(0.25,1.1^2)) 91

Figure 12. The RMSE statistics of IRT equating (θ ~ N(0.5,1.2^2)) 92

Figure 13. The RMSE statistics of IRT equating (θ ~ N(-0.25,1.1^2)) 93

Figure 14. The RMSE statistics of IRT equating (θ ~ N(-0.5,1.2^2)) 94

Figure 15. The Bias statistics of IRT equating (θ ~ N(0,1)) 96

Figure 16. The Bias statistics of IRT equating (θ ~ N(0.25,1.1^2)) 97

Figure 17. The Bias statistics of IRT equating (θ ~ N(0.5,1.2^2)) 98

Figure 18. The Bias statistics of IRT equating (θ ~ N(-0.25,1.1^2)) 99

Figure 19. The Bias statistics of IRT equating (θ ~ N(-0.5,1.2^2)) 100

Figure 20. The Standard Error statistics of IRT equating (θ ~ N(0,1)) 103

Figure 21. The Standard Error statistics of IRT equating (θ ~ N(0.25,1.1^2)) 104

Figure 22. The Standard Error statistics of IRT equating (θ ~ N(0.5,1.2^2)) 105

Figure 23. The Standard Error statistics of IRT equating (θ ~ N(-0.25,1.1^2)) 106

Figure 24. The Standard Error statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$) 107

Figure 25. The RMSE statistics of IRT equating ($\theta \sim N(0,1)$) 121

Figure 26. The RMSE statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$) 122

Figure 27. The RMSE statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$) 123

Figure 28. The RMSE statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$) 124

Figure 29. The RMSE statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$) 125

Figure 30. The Bias statistics of IRT equating ($\theta \sim N(0,1)$) 127

Figure 31. The Bias statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$) 128

Figure 32. The Bias statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$) 129

Figure 33. The Bias statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$) 130

Figure 34. The Bias statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$) 131

Figure 35. The Standard Error statistics of IRT equating ($\theta \sim N(0,1)$) 132

Figure 36. The Standard Error statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$) 133

Figure 37. The Standard Error statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$) 134

Figure 38. The Standard Error statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$) 135

Figure 39. The Standard Error statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$) 136

Figure 40. The transformations of item difficulties (*b*-parameters) for English 138

Figure 41. The transformations of item discrimination (*a*-parameters) for English... 139

Figure 42. Estimated *to* scale true score equivalents of *from* scale true scores using IRT true score equating (English) 140

Figure 43. The transformations of item difficulties (*b*-parameters) for Mathematics 141

Figure 44. The transformations of item discrimination (*a*-parameters) for Mathematics 142

Figure 45. Estimated *to* scale true score equivalents of *from* scale true scores using IRT true score equating (Mathematics) 143

ROBUST SCALE TRNASFORMATION METHODS IN IRT TRUE SCORE EQUATING UNDER COMMON-ITEM NONEQUIVALENT GROUPS DESIGN

Yong He

Dr. Steven Osterlind, Dissertation Supervisor

ABSTRACT

Common test items play an important role in equating multiple test forms under the common-item nonequivalent groups design. Inconsistent item parameter estimates among common items can lead to large bias in equated scores for IRT true score equating. Current methods extensively focus on detection and elimination of outlying common items, which usually leads to enlarged random equating error and inadequate content representation of common items.

New robust scale transformation methods based on robust regression, the robust Deming regression method, the robust Haebara method, and the least absolute values (LAV) method, were proposed. In simulation studies, performances of the proposed methods were compared to the Stocking-Lord method which yields the least equating errors among the traditional method and to outlier removal methods. The results indicate: 1) the robust Haebara method and the LAV method usually outperform the robust Deming regression method, 2) the robust Haebara method and the LAV method perform as well as the Stocking Lord method under the condition of *No outlier*, 3) the robust Haebara method and the LAV method perform better than the Stocking-Lord method when a single outlying common item is simulated, 4) the LAV method and the robust Haebara method are better than, or at least comparable to, the existing outlier removal

methods in the presence of a single outlying common item, and 5) the LAV method and the robust Haebara method have smaller equated scores than the Stocking-Lord method using the *CBASE* data of English and Mathematics.

Keywords: item response theory, test equating, outlier, robust regression, Deming regression, common item

CHAPTER 1

INTRODUCTION

1.1 Introduction

Most large-scale testing programs typically have multiple forms administered on different dates and/or at different locations. It prevents spread of test items which threatens test security (Kolen & Whitney, 1982). In addition, new versions of test forms, which must be equivalent (parallel), are continuously produced with the efforts of numerous test developers. Various types of parallel test forms include randomly parallel tests, τ-equivalent (tau) measures, essentially τ-equivalent measures, and congeneric tests. This study deals only with tests that are congeneric, which means "the scores between tests are characterized by a uniform linear relationship", and "the error variances of each test form … must be random and symmetric for the population" (Osterlind, 2009, p. 70).

If at all possible, these alternate test forms should be identical in both content specifications and statistical properties. However, differences, especially in statistical specifications, unavoidably exist in these alternate test forms despite having a common blueprint. As a consequence, scores of examinees who take different test forms are not always directly comparable because the scores are typically not equivalent. To make matters worse, high-stakes decisions such as college admissions are often on the basis of direct comparisons of test scores from different forms, which have great impact on examinees. To fairly compare examinees in terms of their test scores and make sound decisions, we need to put the scores on the same scale; namely, make test scores from

different forms interchangeable (Petersen, Kolen, & Hoover, 1989). The statistical process to "express the scores of one test on the scale of another test with maximum precision" (Osterlind, 2009, p. 442) or to "adjust scores on test forms so that scores on the forms can be used interchangeably" (Kolen & Brennan, 2004, p. 2) is referred to as test equating. It is worth noting that test equating is not a technique for comparing contents specificities of different test forms.

Since its debut decades ago, a variety of equating approaches have been developed. These approaches integrate three components for test equating: equating designs, expected correspondence among scores, and equating method. However, the second component is not always required by test equating procedures (Cook & Eignor, 1991). Therefore, procedures for test equating in the literature predominately include only two parts: equating design and equating method. Equating designs consider how to collect equating data. For each design, several statistical methods could be implemented to equate different test forms. Depending on different theoretical frameworks, score correspondence and equating methods can be classified into two categories: one is based on the classical test theory (CTT), and the other rests on the item response theory (IRT).

A detailed literature review of equating design and their related equating methods is presented in Chapter 2. For this work, and as is widespread in large-scale testing programs, I employ the common-item nonequivalent groups (CINEG) design. While no single process is the most popular out of many possibilities, in practice, IRT true score equating is broadly used to equate the number-correct observed scores from different test forms (Lord, 1982). This study addresses concerns with the IRT true score equating

process, and provides a brief introduction of the CINEG design and IRT true score equating method.

1.2 Common-Item Nonequivalent Groups Design

In the CINEG design, the groups of examinees taking different test forms are not statistically equivalent. Differences of score distribution characteristics on multiple test forms stem from two sources: form difficulty and the examinee groups. A main goal of the CINEG design is to separate the differences. To do so, two groups of examinees are administered two separate test forms, each embedded with a set of common items. In order to represent the group differences, the common item set needs to be a "mini version" of the tests to be equated (Angoff, 1984). In other words, the common-item set should "behave similarly in the old and new forms" and "be proportionally representative of the total test forms in content and statistical characteristics" (Kolen & Brennan, 2004, p. 19). Research has shown that the content-representative common items have less equating bias than the unrepresentative common items (Cook & Petersen, 1987; Klein & Jarjoura, 1985; Marco, Petersen, & Stewart, 1983).

1.3 IRT True Score Equating

The IRT equating methods are built on the basis of item response theory, which relates the responses of examinees to both the item parameters and the proficiency – an unobservable (latent) trait. The IRT true score equating (Lord, 1982) engages three steps (Cook & Eignor, 1991): choosing an appropriate equating design, establishing a common scale for estimated ability and item parameters (i.e., scale transformation), and equating

test scores. Under the CINEG design, if the estimates of ability and item parameters are separately calibrated, they are on different scales due to the nature of non-equivalent groups. After scale transformation, item parameters and ability of the two forms are placed onto the same scale. In practice, equating results are usually reported as number-correct true scores. Therefore, the equating establishes a relationship between the number-correct true scores on the two forms through the estimated abilities on the common scale.

1.4 Scale Transformation

The process of scale transformation, which is critical in test equating, converts estimates of item parameters and proficiency parameters from separate calibrations into a common scale. Theoretically, due to the indeterminacy of scale location and spread in the IRT models, two sets of separately calibrated parameters only differ by a single linear transformation (e.g., Baker & Kim, 2004; De Ayala, 2009). As a result, the probability values for any linearly related scales are identical.

In practice, four methods are extensively implemented to determine scale transformation coefficients. Two of them, the mean/mean method (Loyd & Hoover, 1980) and the mean/sigma method (Marco, 1977), are based on moments of *a*- and *b*-parameters. The haebara method (Haebara, 1980) and the stocking-lord method (Stocking & Lord, 1983), on the other hand, are based on the characteristic curves. In Chapter 2, I provide more details of scale transformations.

The moment methods are widely used due to their simplicity of computation. However, the methods could be affected by outlying common items, especially considering the larger difference between the separate *b*-parameter estimates (Baker & Al-Karni, 1991; Hu et al., 2008). To decrease outlier effects, methods using the characteristic curve were proposed. The Haebara approach estimates the transformation coefficients, A and B, by minimizing the summed squares of differences between the item characteristic curves over examinees. The Stocking-Lord method, however, estimates A and B by minimizing the summed squared differences between the test characteristic curves over examinees. A detailed description of the Haebara method and the Stocking-Lord method is in Chapter 3.

1.5 Statement of Problem

To ensure a successful scale transformation, item parameter estimates for common items should be consistent between new and old test forms. However, in practice, a distorted scale transformation often occurs due to inconsistency on item parameter estimates of common items from two test forms. The phenomenon of inconsistent item parameters has been extensively addressed in two fields: IRT-based Differential Item Functioning (e.g. Lord, 1980; review in Osterlind & Everson, 2009; Thissen, Steinberg, & Wainer, 1993) and Item Parameter Drift (e.g. Bock, Muraki, & Pfeiffenberger, 1988; Cook, Eignor, & Taft, 1988; Wells, Subkoviak, & Serlin, 2002). However, the results in terms of scale transformation using traditional methods, i.e., Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord, are not consistent (e.g., Hu et al., 2008; Ogasawara, 2000).

To address the distortion of scale transformation, current methods extensively focus on detection and elimination of outlying common items. This does improve the stability of scale transformation that follows IRT equating to some extent. Based on the IRT framework, Kolen and Brennan (2004) further recommended checking for outliers using scatter plots of common item parameter estimates. Other works that have attempted to improve this overt sensitivity includes the works of Murphy, Little, Fan, Lin, and Kirkpatrick (2010), who used a displacement method with a predetermined cut-off value, and He, Cui, Fang, and Chen (in press), who employed a residual analysis based on the ordinary least square linear regression to detect outliers.

Generally, after outliers are identified, they are simply eliminated from the scale transformation. However, elimination of outlying common items can cause problems in IRT equating, and may lead to further problems such as inadequate content representation of common items (Cook & Petersen, 1987). To maintain content representation, additional items, along with the outlying common items, should be eliminated from the common-item set (Kolen & Brennan, 2004). This, however, has the effect of dramatically decreasing the number of items for the scale transformation, which, in turn, may increase random equating error due to the resultant smaller number of common items using in the scale transformation (Petersen, Cook, & Stocking, 1983).

Simply eliminating outlying common items jeopardizes content representativeness. To solve the problem, robust methods have been recommended in the moment approaches. For instance, Cook, Eignor, and Hutten (1979) proposed a method of restricting a range for the item difficulties to remove the effect of the outlying common

items. Linn, Levine, Hastings, and Wardrop (1980) used weighted moments by defining weights as inversely proportional to the estimated standard error of the estimated item difficulties. Bejar and Wingersky (1981) used a robust method that gives smaller weights to outlying items used to estimate the moments. Stocking and Lord (1983) proposed an iterative weighted weighted Mean/Sigma method, which integrated procedures of Linn et al. (1980) and Bejar and Wingersky (1981). However, the existing robust procedures solely considered the estimated item difficulties. As a consequence, scale transformation methods considering more item information are needed.

As a result, the characteristic curve methods were developed to simultaneously consider *a*- and *b*- parameters. Although the characteristic curves methods are more "robust" than the moment approaches (Osagawara, 2001a), a simulation study (He et al., in press) found distorted scale transformation and increased equating errors in the presence of mild outliers (*b*-parameter decreased by 0.5). In addition, Haebara (1980; p. 149) acknowledged that "a possible modification of the method to make it more robust to the existence of outliers may be to remove those items … from the equating process." Apparently, elimination of outlying common items is considered as an alternative method to solve the problem even for the characteristic-curve method. Therefore, characteristic-curve-based scale transformation methods that are more robust to outlying common items are also on demand.

1.6 Purpose of Study

The above mentioned dilemma between the existence of outlying common items and the destructive effect of eliminating the outlying common items leads me to propose

several new methods of deriving scale transformation coefficients to reduce the influence of outlying common items instead of eliminating them. The main purpose of the study is to develop robust methods that not only solve problems caused by outlying common items, but also produce relatively consistent scale transformation coefficients in the absence of outlying common items. In addition, this study also explores the performances of current methods of scale transformations in the presence of various outlying common items are also explored in this study.

To illustrate the robust approaches of the proposed model, analyses were carried out on data from the following: 1) simulated item responses based on Kolen and Brennan's item parameters in their classic work Test Equating, Scaling, and Linking (2004), and 2) *College Basic Academic Subjects Examination* (*CBASE*). *CBASE* is currently required by State of Missouri for teacher certification. It is used nationally by over 130 colleges or universities. In addition to a diagnostic assessment of applicants' knowledge and skills in Mathematics, English, Science, and Social Studies, this exam provides information about the reasoning competencies of the student. The complete test battery comprises 180 multiple-choice test items for all four subjects. The English test includes 41 items, the mathematics test includes 56 items, the science test includes 41 items, and the social science test includes 42 items.

Specifically, the study intends to address the following questions:

1. What are the performances of the traditional scale transformation methods, i.e., Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord, when a variety of conditions of a single outlier are simulated (both a- and b- parameters are allowed

to vary)? Which one performs better in the absence and/or in the presence of a single outlier?

2. How do the proposed robust methods perform in the presence of a simulated outlying common item as compared to the current characteristic-curve methods?
3. How do the proposed robust methods perform in the presence of a simulated outlying common item as compared to the current characteristic-curve methods after outlier detection and removal?
4. Do the proposed robust methods perform as well as the traditional scale transformation methods when there is no outlying common item with various simulated situations including test length, number of common items, distribution of ability, and the spread of common items?
5. How well do the proposed robust methods perform with empirical data from *CBASE*? A detailed description of the exam is provided in Chapter 3.

1.7 Significance of the Study

Test equating is a critical statistical procedure to adjust scores from multiple test forms in order to make the scores comparable. A meaningful and fair interpretation of test scores from multiple forms prevents inaccurate equating that could potentially result in wrong decisions for important events such as college admission. The rise in varying item parameters due to exposure, guided practices, or changed curriculum may deteriorate the accuracy of scale transformation. Although there are methods to deal with the problem, they are limited to detecting and eliminating outlying items, which results in weakening the content representation and increasing random errors in equating.

In this study, I first compared the performances of traditional scale transformation methods under a variety of situations of simulated outliers to systematically study the performance of the traditional scale transformation methods. In addition, it is well known that not a single method can perfectly identify outliers. Keeping outliers in the common item set when conducting scale transformation could produce large equating errors. Therefore, it is important to identify methods that are more robust to treat outlying common items.

In this study, I propose a solution for the above-mentioned problem by incorporating concepts of robust regression approaches into the scale transformation. Researchers have been working hard to solve the problem in the last several decades. After the applications of moment approaches in scale transformation, researchers noticed that outlying item parameters had a serious effect on the computation of the means or standard deviation. Particularly, the mean/sigma method is less stable than the mean/mean method in scale transformation (Ogasawara, 2000). There are two approaches to solve this problem. The first approach focused on reducing the impact of outlying items by using weighting schemes (Bejar & Wingersky, 1981; Cook, Eignor, & Hutten, 1979; Linn, Levine, Hastings, & Wardrop, 1980; Stocking & Lord, 1983). However, the robust approaches are rarely used in practice. The other approach is minimizing a loss function defined in the characteristic curves (Haebara, 1980; Stocking & Lord, 1983). Although the methods of characteristic curves are more "robust" than the moment approaches, outlier elimination is still practiced. To avoid the consequences caused by eliminating items in the methods of characteristic curves, methods incorporating robust regression and characteristic curves are proposed in this study.

In addition, more realistic treatments of outlying common items are used in simulation study. In the studies of detection and elimination of outlying common items, only one parameter, particularly the b-parameter, or one direction of change, is usually used (e.g., He et al., 2011; Hu et al., 2008; Murphy et al., 2010), specifically when the b-parameter decreases. However, the change of item parameters is complex. Therefore, in this study, both a- and b- parameters are set to either increase or decrease by a random value. Under such conditions, the proposed approaches are compared to the current scale transformation methods in terms of computing scale transformation coefficients and accuracy of equating. These detailed and realistic conditions are helpful to generalize the comparison results from the current and proposed methods for practical application in the future.

Provided that the proposed approaches outperform the current approaches in the presence of outlying common items, one may question whether the proposed approaches are reliable in the absence of outlying common items. To answer this question, I compared the two sets of approaches in various conditions such as test length, number of common items, and ability distribution. In this study, no outlying common item is simulated. The simulation studies have advantages as it is not always feasible to collect particular types of data such as the outlying common items. However, the simulation studies also have disadvantages such as the distribution of examinees is defined by the researchers. In this study, I implement empirical data from a real examination, which should make the study more reliable.

Overall, if the proposed approaches perform as well as, if not better than the current approaches, it is evident that the proposed approaches would be recommended for the scale transformation.

1.8 Limitations of the Study

The limitations of the current study are as follows:

1. This study is limited to the unidimensional IRT true score equating. In the future, multidimensional IRT true score equating could be studied.
2. Empirical data from *CBASE* were used for this study. However, the existence of outlying common items for the specific data was not clear. To generalize the conclusion, data from other testing programs should be used in the future
3. The simulation study generated a single outlying common item for the investigation. In addition, the study with an outlying common item did not consider situations with varying test length, number of common items. In the future, a study could be done to investigate these situations.
4. The equating procedure is limited to two test forms. In the future, multiple test forms could be tested.

1.9 Overview of the Subsequent Chapters

The subsequent chapters are organized as follows: Chapter 2 provides a review of equating design, equating methods, issues in scale transformation using the common item nonequivalent groups design, and introduction of concepts in relation to robust regression; Chapter 3 describes the current scale transformation methods based on the characteristic

curves and the proposed approaches, as well as the data used in the procedures; Chapter 4 presents the results and discussions based on the four research questions; and Chapter 5 summarizes the findings and provides considerations of limitations of the study and directions for future research.

CHAPTER 2

LITERATURE REVIEW

Chapter 2 consists of four main sections. The first section provides an overview of a general framework of test equating, including the commonly used data collection designs and equating methods. The second section surveys the current methods of scale transformation under the CINEG design. The third section describs the issues related to the common item set. The fourth section reviews the concepts of robust regression, which are related to the topics in this dissertation.

2.1 Overview of Test Equating

Test scores often represent the performance of students on tests, particularly in test programs that assist in high-stakes decision-making, such as college admissions, scholarships, or professional licenses. Theoretically, all examinees need to take the same test form, no matter when or where they take the test. However, it is not practical, because item exposure threatens test security (Kolen & Whitney, 1982). For instance, some test items are administered six times a year. Examinees who have taken a test may re-take the same test at a later date to obtain a higher score, or share test items with others who are planning to take the test. This undermines many aspects of a test fairness, security, and test validity. To avoid this problem, more than one form and date of tests are implemented in practice.

The alternate forms of a test include test items constructed on the same content and statistical specifications. However, differences amongst these tests, especially in difficulty, make the scores from alternate forms incomparable. To fairly evaluate

examinees taking different test forms in terms of their performances, we need to employ a statistical process to establish "a relationship between raw scores on two test forms that can then be used to express the scores on one form in terms of the scores on the other form" (Petersen et al., 1989, p. 242). In other words, a statistical process is needed to make the scores from alternate test forms interchangeable. This statistical process is referred to as test equating, which serves to "express the scores of one test on the scale of another test with maximum precision" (Osterlind, 2009, p. 442) or to "adjust scores on test forms so that scores on the forms can be used interchangeably" (Kolen & Brennan, 2004, p. 2). However, Test equating could not be used to examine whether the test contents of different forms are the same. With this limitation, the test forms should test the same construct with similar psychometric properties in terms of difficulty, symmetry, equity, and population invariance (Angoff, 1984; Lord, 1980).

In general, a complete equating procedure comprises three important components: sampling design, expected correspondence among scores, and specific statistical procedures used to estimate score correspondence (Kolen, 1988; Cook & Eignor, 1991). The second component, however, is not required by all test equating procedures (Cook & Eignor, 1991). Usually, the last two components are combined as one component in textbooks or literatures. Consequently, the procedure of test equating is commonly described in two parts: equating design and equating method.

2.1.1 Equating Design

An equating design is "mainly a matter of logistical aspects of test administration and data collection" (Osterlind, 2009; p. 447). In this sense, the equating design establishes a

structure of data collection including design of test forms, administration of tests, and groups of examinees. Like any other research designs, the equating design aims to establish a connection between two test administrations in a way that the confounding variables can be removed.

In practice, three major equating designs are used: single group (SG) design, random groups (RG) design, and common item non-equivalent groups (CINEG) design (Kolen & Brennen, 2004). However, von Davier et al. (2004) use the equivalent groups (EG) design as an alternateive to the random groups (RG) design, and the nonequivalent groups with anchor test (NEAT) design as the common item non-equivalent groups (CINEG) design. In this study, the notations from Kolen and Brennan (2004) are used to diminish confusions.

2.1.1.1 Single Group Design

In the SG design, all examinees in a group receive two forms of a test, Form X and Form Y, in a single administration. In this particular design, the potential confounding effect of ability differences in the groups is under control so that the differences between the test scores are only attributed to the test forms. However, the order for taking the test has significant effect on scores, such as learning or fatigue, which introduces errors to this design. To control the order effect of administration of the test, an improved version, a single group with a counterbalancing design, is often adopted. In this design, the examinees in a group are divided into two subgroups, but the test still includes Form X and Form Y. Examinees in one subgroup take Form X first and Form Y second, and examinees in the other subgroup take Form Y first and Form X second. The procedure is

referred to as a spiraling process. The main advantage of using this design is that it requires a smaller sample size and decreases sampling errors, while eliminating the obvious disadvantage of prolonging test time.

2.1.1.2 Random Groups Design

In the RG design, examinees are randomly assigned to a test administration of two equivalent test forms, typically by using spiraling process. Here, the examinees who took different forms are considered to be independent random samples from the same population of examinees. In this way, the differences in scores between the two groups are attributed to the differences in difficulty between the two forms because the examinees in two groups are expected to be randomly equivalent. This design is often preferred over the SG design because it shortens testing time and eliminates the order effect on test scores. From a practical consideration, it is also easier to implement than the SG design. However, in this design, a larger number of examinees are needed than in the SG design because (1) the groups should be representative of the same population (Osterlind, 2009), and (2) the ability distributions of two groups may not be the same when sample size is small (Kolen & Brennan, 2004).

2.1.1.3 Common-Item Nonequivalent Groups Design

The CINEG design is very useful when only one form of the test could be administered per test date due to practical issues such as test security. In this design, the groups of examinees taking different test forms are not assumed to be equivalent. Therefore, the differences in test scores are attributed to both differences in test forms

and the ability levels of examinees. For this reason, a set of common items are embedded in both forms to adjust for population differences, either internally or externally depending on whether their scores contribute to the total score on the test. The advantage of this design is its flexibility in administration because only one test form is needed at a given test date. Its disadvantage is that it requires strong statistical assumptions to separate the differences in test forms from differences in group ability levels. In addition, the group differences should not be very large. Otherwise, there is no statistical procedure that could adjust for these differences (Kolen, 1988).

Two important issues are related to the CINEG design. First, the common items must be representative of all contents and items on the test forms. In other words, the common-item sets should be a "mini" version of the overall test (Angoff, 1984), such that it "should be proportionally representative of the total test forms in content and statistical characteristics" (Kolen & Brennan, 2004, p. 19). Second, the number of common items should be large enough to maintain good reliability.

2.1.2 Equating Methods

For any specific equating design, several equating methods are available. Generally, one group of equating methods is based upon the classical test theory (CTT), and the other group is based on the item response theory (IRT). In this section, I will provide an overview of the CTT-based equating methods according to the equating designs. Then, I will focus on the IRT-based equating, especially the true score equating which is used in this dissertation. The fundamental distinction between the CTT equating and the IRT

equating is that the CTT methods focus on test level data, but the IRT methods focus on item level data.

2.1.2.1 Equating Methods for the SG and RG designs

Both the CTT equating methods and the IRT equating methods can be used for these two equating designs. For convenience, let *X* stand for the new form and *Y* for the old form.

The CTT based methods focus on matching scores correspondence by simple transformation of the observed test score distribution from alternate forms. Usually the means, standard deviations, or percentile ranks, are set to equal for a specific group of examinees (Kolen, 1988). There are three equating methods based on CTT that are used for the SG and RG designs: mean equating, linear equating, and equipercentile equating.

In mean equating, the means of the two forms are set equal. Since the difference between two forms to be equated is a constant along the score scale, the score correspondence of *x* (a score on Form X) on the Form Y is

$$y = x + [\mu(Y) - \mu(X)],$$

where $\mu(X)$ and $\mu(Y)$ are means of scores on Form X and Form Y, respectively.

In linear equating, the means and standard deviations on equated forms are set equal. In other words, the z-scores on the equated forms are set equal:

$$\frac{y - \mu(Y)}{\sigma(Y)} = \frac{x - \mu(X)}{\sigma(X)},$$

where $\sigma(X)$ and $\sigma(Y)$ are standard deviations of observed scores on Form X and Form Y, respectively. Therefore, the score y on Form Y corresponding to score x on Form X is

$$y = \frac{\sigma(Y)}{\sigma(X)} x + [\mu(Y) - \frac{\sigma(Y)}{\sigma(X)} \mu(X)].$$

In equipercentile equating, the scores on the alternate forms are set equal if the percentile ranks are the same. Thus the score y on From Y corresponding to score x on Form X is

$$y = Q^{-1}P(x),$$

where Q^{-1} is the inverse of percentile rank function of Form Y, and $P(x)$ is percentile rank function of Form X.

In the SG and RG designs, the parameter estimates from either separate calibration or simultaneous calibration are on the same scale since the groups responding to the items are identical in ability. It should be noted that in the SG design the parameters for the alternate forms are estimated on the same examinees, and in the RG design, the groups of examinees are assumed to be identical. Since the item parameter estimates are on the same scale, it is not necessary to have further scale transformation. In other words, the correspondence scores are not required in the equating process here.

2.1.2.2 CTT-Based Equating Methods for the CINEG Design

In the CINEG design, the common items are used to adjust for ability differences in the two groups for the CTT based equating methods. The scores to be equated are

different. The basic model of CTT is that observed raw scores comprise two components: the true score which is "the unbiased estimation of a construct" (Osterlind, 2009, p. 56), and the expected scores for the observed scores and the measurement error. Accordingly, one type of the equating methods focuses on the observed scores and the other on the true scores.Under the CTT framework, equating with the CINEG design has two main types: linear and equipercentile, as described in the previous section. However, the equating procedure is much more complicated than those in the SG and the RG designs. As mentioned earlier, the common items are used to adjust for the differences between the groups taking alternate forms by generalizing a statistical relationship between the total score and the common item scores to a target population.

In linear equating with the common items, the z-scores are set to equal for the alternate forms via a synthetic population (Braun & Holland, 1982) in which the two equated populations are weighted with w_1 and w_2 where $w_1 + w_2 = 1$. Therefore, the equation for equating scores on Form X to the scale of Form Y is

$$y^* = \frac{\sigma_S(Y^*)}{\sigma_S(X^*)}[x^* - \mu_S(X^*)] + \mu_S(Y^*),$$

where the subscript *S* indicated the synthetic population, and the asterisk indicates that the scores could be either observed scores or true scores. To carry out the equating, additional assumptions related to the common items are required. In the Tucker method for the observed scores (Gulliksen, 1950; pp. 299-301), the regression of the observed scores on Form X and the observed scores on the common items is assumed to be the same for both alternate populations, and the conditional variance of scores on Form X

given the scores on the common items is assumed to be the same for both alternate populations. In the Levine methods for observed scores or true scores (Levine, 1955), the correlation between the true scores either on Form X or on Form Y and true scores on the common items is assumed to be 1. In addition to the direct matching between the scores, a chained process is also suggested (Angoff, 1971; Livingston, 1996), where the scores on the new form are equated to the scores on the common items, which are then scored on the common items that are equated to the reference form.

In the equipercentile equating with common items, the frequency estimation (FE) method (Angoff, 1971; Braun & Holland, 1982) is used. By using this method, the conditional distributions of total scores given a score on common items for the alternate forms are assumed to be the same in both populations. A target (synthetic) population is obtained by weighted combination of the distributions for the alternate populations. Then, the equated scores are obtained according to the equal percentile ranks:

$$y = Q_S^{-1}[P_S(x)],$$

where the subscript *S* indicates the synthetic population. In addition, Braun and Holland (1982) generalized the Tuker method by considering the regression of total test on the common items as nonlinear, whereas Angoff, 1971, Dorans, 1990; Livingston, Dorans, & Wright, 1990 suggested a chained process for equipercentile equating (chained equipercentile or CE method).

2.1.2.3 IRT-Based Equating Methods for the CINEG Design

The IRT equating methods are built on the basis of the item response theory (e.g., Baker & Kim, 2004; de Ayala, 2009; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Osterlind, 2009; van der Linden & Hambleton, 1997). In IRT, the responses from examinees and the item parameters are considered. Depending on the number of constructs measured and considered in the test, the IRT models could be classified into two large categories: unidimensional IRT models and multidimensional IRT models. A unidimensional IRT model with dichotomous responses (1 for correct and 0 for incorrect) is used in this study, thus; its properties are reviewed here. A unidimensional IRT model measures only one construct, such as students' proficiency in mathematics. In addition, the model assumes that local independence, which means that an examinee's performance on one item is independent of the performance on other items. Item characteristic curve defines the characteristics of a test. In most cases, three item parameters are estimated: including discrimination (*a*-parameter), difficulty (*b*-parameter), and pseudo-chance or pseudo-guessing (*c*-parameter). The model accordingly is named three-parameter logistic (3PL) IRT model, which is used in this study.

Under a 3PL IRT model, the probability for an examinee *i* with ability θ_i to correctly answer item *j* is defined as

$$p_{ij} = p_{ij}(\theta_i; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}},$$

where a_j, b_j, and c_j are the item parameters indicating discrimination, difficulty, and pseudo-guessing for item j, respectively. D is a constant which equals 1.7 to make the frequency function to closely approximate a normal distribution. To obtain a 2PL model, let $c_j = 0$, and to obtain a 1PL model, let $c_j = 0$ and $a_j = 1$.

Besides the feature of invariant measurement of IRT models, the statistics in item parameters of the IRT models render more appealing in test equating. Under the CINEG design, there are two ways to place the item parameter estimates on a common scale. In one method, the parameters for both forms are estimated together, which is referred to as concurrent calibration (Wingersky & Lord, 1984). In this condition, all item parameters are automatically placed on the same scale. The other method has the parameters for both forms estimated separately, where the scales are not the same. As the common items are placed on the same scale of item parameter estimates, this procedure is referred to as scale transformation. The scale transformation will be discussed in detail in the next section because it is the core concept of this dissertation.

Comparison between concurrent and separate calibrations has been done, but the results are not conclusive. Some studies indicated that concurrent estimations performed slightly better than separate estimations (Petersen, Cook, & Stocking, 1983; Wingersky, Cook, & Eignor; 1987) using ~~the~~ joint maximum likelihood estimation to obtain item parameters. However, separate calibration is recommended when the number of common items is small using ~~the~~ marginal maximum likelihood estimation to obtain item parameters (Kim & Cohen, 1998). Hanson and Béguin (2002) found that the small resultant errors from concurrent calibration might be due to large sample size of the

method. In addition, they concluded that "there are not sufficient to recommend completely avoiding separate estimation in favor of concurrent estimation." (p. 19). Further, they recommended that separate estimation could be used to identify potential problems, such as violation of assumptions or outlying common items. Following this logic, "In practice, separate estimation using the test characteristic curve methods seems to be safest." (Kolen & Brennan, 2004; p. 174) Moreover, the item parameters of the old test form (Form Y) are estimated before the administration of new test form (Form X). In this circumstance, only item parameters on the new form need to be estimated when conducting an equating process from the new form to the old form. Consequently, it is more reasonable to use separate calibration than to use concurrent calibration.

Similar to the CTT equating methods, the IRT equating methods also include true score equating and observed equating. The IRT observed score equating (Lord, 1982) uses the IRT model to obtain distribution of the observed scores on alternate forms. Then, the distributions are used for an equipercentile equating. Although its results are similar to the true score equating (Lord & Wingersky, 1984), it is computational expensive. In this study, I will focus on the IRT true score equating, instead.

IRT true score equating (Lord, 1982) uses the number-correct scores to set up the relationship between two test forms. The number-correct true scores, also known as test characteristic curves, are used because the measurement errors are high for both high and low ability examinees (Lord, 1980, p. 183). After scale transformation (which will be discussed later), the true scores are used to relate the alternate forms according to their corresponding abilities.

The number-correct true score on Form X for a given ability θ_i is defined as

$$\tau_X(\theta_i) = \sum_{j:X} p_{ij}(\theta_i; a_j, b_j, c_j),$$

and the number-correct true score on Form Y for a given ability θ_i is defined as

$$\tau_Y(\theta_i) = \sum_{j:Y} p_{ij}(\theta_i; a_j, b_j, c_j).$$

However, in the 3PL model, the true scores under c_j are not achievable because the lower asymptote tends to be c_j rather than 0. Therefore, the true scores for the alternate forms fall in the range between c_j and the numbers of items on the forms.

The true score equating first finds true scores on Form X (τ_X), followed by finding the corresponding θ_i by using the Newton-Raphson method and last obtains the true scores on Form Y (τ_Y). For a given ability θ_i, the true score on Form Y corresponding to the Form X true score is expressed as

$$irt_Y(\tau_X) = \tau_Y(\tau_X^{-1}), \qquad \sum_{j:X} c_j < \tau_X < K_X,$$

where τ_X^{-1} is the θ_i corresponding to true score τ_X.

2.2 Scale Transformation in IRT Equating

When the parameters for both forms are estimated separately under the CINEG design, the scales are not on the same scale. Consequently, the common items are needed to place item parameter estimates on the same scale, which is scale transformation. Under

the IRT framework, item parameters from two separately calibrated forms of varying ability levels would differ by a linear transformation due to the invariant measurements of IRT models. Suppose two test forms with either the *to* scale (T) [cf. focal or new scale] or the *from* scale (F) [cf. reference or old scale] are used for equating, the equations to relate the two scales with the coefficients *A* and *B* are:

$$\theta_{Ti} = A\theta_{Fi} + B,$$

$$a_{Tj} = \frac{a_{Fi}}{A},$$

$$b_{Ti} = Ab_{Fi} + B,$$

$$c_{Ti} = c_{Fi}.$$

where *a*, *b*, and *c*'s denote the item parameters. It is obvious that the relationships between the corresponding *a*- and *b*- parameters are linear, and the c-parameters are free of transformation. The goal of scale transformation is to obtain the transformation coefficients, A and B, which also represent the slope and intercept of the linear transformation.

Practically, four methods are predominantly applied in determination of the scale transformation coefficients. The mean/mean method (Loyd & Hoover, 1980) and the mean/sigma method (Marco, 1977) are based on moments of *a*- and *b*- parameters, focusing on computations of means and standard deviations for the parameters on the alternate forms. The Haebara method (Haebara, 1980) and the Stocking-Lord method

(Stocking & Lord, 1983) are based on characteristic curves, where the probability of correct answers considering both the item parameter estimates and examinees' abilities is used. Besides the widely used ones, other methods have also been proposed, but are not widely adopted in practice. For instance, Divgi (1985) proposed a minimum chi-squares method to minimize an alternative criterion as an alternative to that used in the Stocking-Lord method. Osagawara (2001) investigated the least square methods in obtaining scale transformation coefficients, including unweighted least squares, generalized least squares, and weighted least squares. He found that the weighted least squares method produces very similar results to the characteristic curve methods with comparable standard errors.

In this section, four popular methods are reviewed with particular focus on the characteristic curve methods.

2.2.1 The Moment Methods of Scale Transformation

The mean/mean approach estimates the transformation coefficients, *A* and *B*, by:

$$A = \frac{\mu(a_{Fi})}{\mu(a_{Ti})},$$

$$B = \mu(b_{Ti}) - A\mu(b_{Fi}).$$

and the mean/sigma method estimates the coefficients by:

$$A = \frac{\sigma(b_{Ti})}{\sigma(b_{Fi})},$$

$$B = \mu(b_{Ti}) - A\mu(b_{Fi}).$$

The mean/mean method and the mean/sigma method are easy to calculate because either means or standard deviations from two sets of *a*- or *b*- parameter estimates are involved in the procedure. Generally, the mean/mean method is more stable than the mean/sigma method because the means are typically more consistent than standard deviations (Baker & Al-Karni, 1991). In addition, the methods could be affected by outlying common items, especially larger difference between the separate *b*-parameter estimates (Baker & Al-Karni, 1991; Hu et al., 2008).

2.2.2 The Haebara Method

The Haebara method considers minimizing a defined loss function (Haebara, 1980). In this method, a relative frequency distribution of ability is arbitrarily partitioned into a number of small equal-distance intervals. Then, the probability of correct responses to each item is computed for a given ability interval in both alternate scales. A weighted least squares method is used to obtain the slope and intercept of the linear transformation which could minimize the loss function over the relative frequency distribution of ability.

In detail, the method includes four steps. First, the difference between the equated scales in terms of the probability of correct answers for item *j* and examinee *i* is expressed as:

$$e_{ij} = p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Fi}; a_{Fj}, b_{Fj}, c_{Fj}) = p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj}).$$

Second, a loss function L which evaluates the resultant loss by e_{ij} is defined as

$$L(e_{ij}) = e_{ij}^2.$$

Third, an equating error for item *j* is given by

$$Q_j = \sum_i L(e_{ij}) = \sum_i e_{ij}^2 = \sum_i [\, p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})]^2.$$

Last, a total equating error is defined as

$$Q = \sum_{j:V} \sum_i L(e_{ij}) = \sum_{j:V} \sum_i e_{ij}^2 = \sum_{j:V} \sum_i [\, p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})]^2,$$

where *V* is the set of common items. To obtain the transformation coefficients *A* and *B*, one needs to minimize the criterion function *Q*. In other words, the *A* and *B* should be the optimum number of the loss function *L*. In practice, a relative frequency distribution $h_1(\theta_i)$ is determined by dividing the range of θ_i into small intervals old form quadrature points. The *Q* is minimized by approximately minimizing Q_1 (or Q_1+Q_2 if one also considers transforming scores from *old* form to *new* form):

$$Q_1 = \sum_j \sum_q e_{ij}^2 h_1(\theta_i) = \sum_j \sum_q [\, p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})]^2 h_1(\theta_{i1}),$$

$$Q_2 = \sum_j \sum_q e_{ij}^2 h_1(\theta_i) = \sum_j \sum_q [\, p_{ij}(\theta_{Fi}; Aa_{Tj}, \frac{b_{Tj} - B}{A}, c_{Tj}) - p_{ij}(\theta_{Fi}; a_{Fj}, b_{Fj}, c_{Fj})]^2 h_2(\theta_{i2}).$$

To simplify the expression, Kolen and Brennan (2004) used

$$Hcrit = \sum_i \sum_{j:V} [p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})]^2.$$

Briefly, the Haebara method is to minimize the sum of the squares of the difference between the item characteristic curves for all θ_i.

2.2.3 The Stocking-Lord Method

The Stocking-Lord method is established on the concept of true score (Stocking & Lord, 1983). In an optimal condition, the estimated true score of the new form after being transformed to the scale of the old form should be identical to the one of the old form. However, difference always exists due to sampling errors or lack of fit of the IRT models. As a consequence, the sum of squared differences between the two estimated true scores of the equated forms needs to be minimized.

The estimated true scores for the old form and the transformed new form are defined as the sum of the probability of correctly answering the items under a given ability θ_i:

$$p(\theta_{Ti}) = \sum_{j:V} p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}),$$

$$p(\theta^*_{Fi}) = \sum_{i} p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj}).$$

The function to be minimized is

$$F = \sum_{i} [\, p(\theta_{Ti}) - p(\theta^*_{Fi})]^2.$$

The function is expressed by Kolen and Brennan (2004) as follows:

$$SLcrit = \sum_i [\sum_{j:V} p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - \sum_{j:V} p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})]^2.$$

To obtain the coefficients, one needs to minimize the sum of the squared difference between the test characteristic curves for all θ_i.

2.2.3 Comparison among Scale Transformation Methods

Comparison among the most widely used four scale transformation methods seems to be inconclusive, especially within the same category of methods. However, there is a consensus that characteristic curve methods usually outperform moment methods. For instance, Baker and Al-Karni (1991), Ogasawar (2001b), Kaskowitz and De Ayala (2001), and Hanson and Béguin (2002), found that IRT scale transformation methods using characteristic curve methods give more stable results than moment methods. Kim and Lee (2006) also compared these methods and drew a similar conclusion in the mixed-format test. In addition, their results indicated that the Heabara method had the least error among them. However, Way and Wang (1991) found that the results from both characteristic curve methods are very similar.

2.3 Issues in Common Item

To ensure a successful scale transformation, item parameter estimates for common items should be consistent between new and old test forms. However, inconsistency in item parameter estimates of common items from two test forms is often found in practice. According to Cook and Eignor (1991), the difficulty parameter (*b*-parameter) is more stable than the other parameter estimates including discrimination parameter (*a*-

parameter) and pseudo-guessing parameter (*c*-parameter). Nevertheless, in practice, a large difference in *b*-parameter estimates frequently occurs due to various reasons. For example, the *b*-parameter estimates for certain items on the new form may be lower than the same items on the old form due to item overexposure. Changes in curriculum (e.g., more emphasis on a particular content area) may also lead to lower estimates of *b*-parameters of relevant items. Changes in parameter estimates lead to inaccurate equating results (Hu, Rogers, & Vukmirovic, 2008; Huang and Shyu, 2003), partially owing to the distorted scale transformation. The phenomenon of inconsistent item parameters has been extensively addressed in two fields based on their research objectives: IRT-based Differential Item Functioning (e.g. Lord, 1980; Osterlind & Everson, 2009; Raju, 1988; Rudner, 1977; Thissen, Steinberg, & Wainer, 1993) and Item Parameter Drift (e.g. Bock, Muraki, & Pfeiffenberger, 1988; Cook & Eignor, 1991; Cook, Eignor, & Taft, 1988; Hu, Rogers, & Vukmirovic, 2008, Huang & Shyu, 2003; Wells, Subkoviak, & Serlin, 2002).

Current methods extensively focus on detection and elimination of outlying common items, which, to some extent, improve the stability of scale transformation as well its subsequent IRT equating. These methods are based upon either classical test theory (CTT) or IRT. For instance, the delta-plot or transformed item difficulties procedure (Angoff, 1972) is the most used CTT based method to detect aberrant items by comparing the transformed item *p* values. Using the method, items that deviate significantly from the principal axis line between the two sets of deltas, which are retained by item difficulties in terms of p values, would be considered as outliers. However, the method might give misleading results unless all the items have the same discrimination power (Angoff, 1982). Another popular method is the Mantel-Haenszel procedure (Holland & Thayer,

1988; Mantel & Haenszel, 1959; Osterlind & Everson, 2009) based on a Chi-square distribution to detect the aberrant items.

Under the IRT framework, Raju (1988) suggested using an area by either a signed area or an unsigned area as a criterion to detect aberrant items. The closed-interval signed area (CSA) for the 3PL IRT model is defined as

$$CSA = S_R(\theta_1, \theta_2) - S_F(\theta_1, \theta_2)$$
$$= (c_R - c_F)(\theta_2 - \theta_1) + \ln\{\frac{\{1+\exp[Da_R(\theta_2 - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_1 - b_F)]\}^{(1-c_F)/Da_F}}{\{1+\exp[Da_R(\theta_1 - b_R)]\}^{(1-c_R)/Da_R}\{1+\exp[Da_F(\theta_2 - b_F)]\}^{(1-c_F)/Da_F}}\},$$

where θ_1 and θ_2 are two ends of an infinite interval of ability, subscript R denotes the reference group, subscript F denotes the focal group, and *a*, *b*, and *c* are the item parameters in the 3PL IRT model. The computation of unsigned area relies on different situations of *a*, *b*, and *c*'s. Therefore, the procedure is somewhat computationally expensive.

Based on the Riemann sum approximations, Rudner (1977) suggested an approximated way to acquire the area by using the sum of the rectangular areas for 200 small intervals of θ between -5 and 5:

$$S_j = \sum_q | p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Fi}; a_{Fj}, b_{Fj}, c_{Fj}) | * \Delta\theta,$$

where $\Delta\theta$ is the difference between two quadrature point on an ability scale, *q* is the set of quadrature points, and the *a*, *b*, and *c*'s are the item parameters in the 3PL IRT model. Compared with the Raju area, this method is conceptually simpler and easier to compute.

In addition, Kolen and Brennan (2004) recommended that the scatter plots of common item parameter estimates should be checked for outliers, but the procedure is a subjective process so that different people tend to determine outliers differently. The displacement method with a predetermined cutoff value is another commonly used method in practice (Murphy, Little, Fan, Lin, & Kirkpatrick, 2010). This method suggests that an item is flagged as an outlier if its absolute difference in the parameter estimates (after being placed on the same scale) of any specific common item exceeds a predetermined cutoff value. The cutoff value, however, is typically arbitrarily determined. Both inspecting scatter plots of common item parameter estimates and examining the differences of common item parameter estimates (i.e., the displacement method) are subject to the investigator's subjectivity. He et al. (in press) proposed a more objective method based on residual analysis of ordinary least square linear regression to detect outliers, based upon the fact that the two sets of item parameter estimates for the common items on both new and old forms have a linear relationship. Unlike previous methods which require scale transformation before detecting outliers and re-run scale transformation after outlier removal, the linear regression method detects and eliminates outliers before scale transformation. In addition, after removing flagged outliers, the accuracy of IRT equating is dramatically improved as compared to the method with outliers in the scale transformation.

However, elimination of outlying common items might cause problems in IRT equating, due to inadequate content representation of common items, especially when the groups of examinees differ noticeably (Cook & Petersen, 1987; Klein & Jarjoura, 1985). In addition, it is evident that eliminating common items might increase random equating

error as a result of small number of common items used in the scale transformation (Petersen, Cook, & Stocking, 1983). Instead of dropping items from the common-item set, Harris (1991) suggested using a weighting method on common items to balance the content. In her study, more weights are given to score point differences occurring frequently.

Along with detection and elimination, robust methods are also considered, but these are mostly related to the moment methods of scale transformation. For example, one group of studies focused on reducing the impact of outlying items by using weighting schemes (Bejar & Wingersky, 1981; Cook, Eignor, & Hutten, 1979; Linn, Levine, Hastings, & Wardrop, 1981; Stocking & Lord, 1983). Cook, Eignor, and Hutten (1979) proposed a method of restricting the range for item difficulty to remove the effect of the outlying common items. Linn, Levine, Hastings, and Wardrop (1980) used weighted moments by defining weights as inversely proportional to the estimated standard error of the estimated item difficulties. To obtain weights, indices including the area enclosed by two ICCs within θ = -3 and θ = 3 are used, after the estimated item parameters are transformed to the same scale. Bejar and Wingersky (1981) utilized a robust method to estimate the moments assigning smaller weights to outlying items. Stocking and Lord (1983) proposed an iterative weighted Mean/Sigma method, which integrats procedures of Linn et al. (1980) and Bejar and Wingersky (1981). The procedure first uses the Mean/Sigma scale transformation with initial weights from estimated standard errors of the estimated b-parameter of common items. Then, a new set of weights is obtained by using the Tukey bi-square weights for the common items according to the absolute perpendicular distance of points to the transformed line. The procedure is repeated until

the change in the absolute perpendicular distance is less than a cutoff value. However, the existing robust procedures solely consider estimated item difficulties. As a consequence, a new scale transformation method considering more item information is needed to improve the approaches, especially in the presence of outlying common items, to reduce the effect of outliers on scale transformation coefficient, and to avoid deleting outliers from the scale transformation.

As a matter of fact, the characteristic curve methods were developed to solve the problem by minimizing a loss function defined in the characteristic curves (Haebara, 1980; Stocking & Lord, 1983). Although the characteristic curves methods are more "robust" than the moment approaches (Osagawara, 2001a), a simulated study (He et al., in press) found distorted scale transformation and increased equating errors in the presence of mild outliers (Δb = -0.5). Therefore, there is also a need for robust methods based on characteristic curve methods. In addition, elimination of outlying common items is still deemed as an alternative way to solve the problem (Haebara, 1980).

2.4 Robustness

To remedy the problem without eliminating outlying common items from scale transformation, robust regression that is not easily affected by the outlying common items has been considered. In robust regression, most of the data is considered (Rousseeuw & Leroy, 1987), but the outlying observations are either dropped or down-weighted. The robust regression considers not only decreasing the influence of outlying observations, but also maintains the sample variance by using certain functions (Andersen, 1998; p. 3-4).

The M-estimators method was proposed by Huber (1964). To obtain the estimators, one needs to minimize the loss function of the residuals, namely,

$$\min \sum_i \rho(r_i),$$

where r_i is the residual of the ith observation, and ρ is the loss function of the residuals with a unique minimum at zero. It is equivalent to minimizing the iterated reweighted least-squares:

$$\min \sum_i w_i * r_i^2$$

where the weights are recomputed after each iteration.

Under the framework to estimate M-estimators, there are several robust loss functions, such as least squares, least absolute values, Huber function, Tukey's biweight function, Andrews function, Hampel function, Blake-Zisserman function, among others (Hartley & Zisserman, 2003; Maronna, Martin, & Yohai, 2006; Rousseeuw & Leroy, 1987). Although the M-estimators work differently, they all assign less weight to an outlying observation. The most widely used weight functions include Huber weighting (Huber, 1977), Tukey bi-square weighting (or bi-weight), and least absolute values.

The Huber loss function is defined as

$$\rho(e_i) = \begin{cases} \frac{1}{2} e_i^2 & |e_i| \le k \\ k|e_i| - \frac{1}{2}k^2 & |e_i| > k \end{cases},$$

where e_i indicates a residual for observation *i*, and *k* is a tuning constant establishing the extent of down-weight. The Huber weight function is defined as

$$w_i(e_i) = \begin{cases} 1 & |e_i| \le k \\ k/|e_i| & |e_i| > k \end{cases}.$$

Smaller values of *k* produce more resistant to outliers, but decrease efficiency under the normal distribution. Through the weight functions, outlying observations are given relatively smaller weights so that their influence on the overall regression estimator is smaller (Draper and Smith 1998). The Huber weighting function has an attractive property, that is, it hardly tends to be zero unless the number of e is enormously large. In this study, $k = 1.345$ is chosen to obtain about 90% efficiency when data are normally distributed.

It should be pointed out that both the estimator of least squares and the estimator of least absolute deviations are special cases Huber M-estimator. When *k* approaches infinity or a number larger than all the residuals, the Huber M-estimator changes to the least square estimator, where the loss function is defined as the sum of squares of the deviation scores. In other words, the weights for the individual observations are all equal to 1 as defined. On the other hand, when *k* approaches zero, the Huber M-estimator approaches to the estimator of least absolute deviations.

The Tukey's bi-weight function is defined as

$$\rho(e_i) = \begin{cases} (1-[1-(e_i/k)^2]^3)k^2/6 & |e_i| \le k \\ k^2/6 & |e_i| > k \end{cases},$$

where e_i indicates a residual for observation i, and k is a tuning constant. The weight function is given by

$$w_i(e_i) = \begin{cases} [1-(e_i/k)^2]^2 & |e_i| \le k \\ 0 & |e_i| > k \end{cases}.$$

Unlike the Huber weight function, the Tukey's bi-square weight function decreases when $|e_i|$ deviates from zero. Additional, it has a faster decrease of weight when $|e_i|$ is large, which is a desirable property due to the objectives of this study. However, two unfavorable features of the bi-square weight function are: (1) its designation to zero when the $|e_i|$ is larger than the predefined tuning constant; and (2) the weight starts decreasing when the residual largely deviates from zero. As a consequence, the Tukey's bi-weight function is not pursued in this study.

CHAPTER 3

METHODOLOGY

This chapter is organized into five main sections. In the first section, the research questions are restated. In the second section, the equations of scale transformation equations based on the characteristic curves i.e., the Haebara method and the Stocking-Lord method are reviewed. In the third section, the procedures of the proposed robust approaches are presented. In the fourth section, studies with simulated data are presented. In the last section, a study with simulated data is presented.

3.1 Overview of Research Questions

The major purpose of the study is to decrease influence of outlying common items on scale transformation by using proposed robust methods. The specific research questions related to the approaches being proposed are as follows:

1. What are the performances of the traditional scale transformation methods, i.e., Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord, when a variety of conditions of outlier are simulated (both *a*- and *b*- parameters are allowed to vary)?
2. How do the proposed robust methods perform in the presence of a simulated outlying common item as compared to the current scale transformation methods?
3. How do the proposed robust methods perform in the presence of a simulated outlying common item as compared to the current scale transformation methods after outlier detection and removal?

4. Do the proposed robust methods perform as well as the traditional scale transformation methods when there is no outlying common item under various simulated situations?
5. How well do the proposed robust methods perform with empirical data from the *CBASE*?

3.2 Scale transformation using Characteristic Curves

Under a three-parameter logistic (3PL) IRT model, the probability for an examinee *i* with ability θ_i to correctly answer item *j* is given by

$$p_{ij} = p_{ij}(\theta_i; a_j, b_j, c_j) = c_j + (1-c_j)\frac{e^{Da_j(\theta_i - b_j)}}{1+e^{Da_j(\theta_i - b_j)}}$$

where a_j, b_j, and c_j are the item parameters indicating discrimination, difficulty, and pseudo-guessing for item *j*, respectively. *D* is a constant which equals 1.7 to make the frequency function to closely approximate a normal distribution.

Under the CINEG design, suppose two test forms with either the *to* scale (*T*, old form) or the *from* scale (*F*, new form) are used for equating. The two sets of item parameters and abilities differ by a linear transformation with the coefficient *A* and *B*:

$$\theta_{Ti} = A\theta_{Fi} + B,\ a_{Tj} = \frac{a_{Fi}}{A},\ b_{Ti} = Ab_{Fi} + B,\ c_{Ti} = c_{Fi}.$$

The Haebara method minimizes the sum of the squares of the difference between the item characteristic curves for all θ_i to obtain the scale transformation coefficients:

$$Hcrit = \sum_i \sum_{j:V} [p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})]^2,$$

and the Stocking-Lord method minimizes the sum of the squared difference between the test characteristic curves for all θ_i:

$$SLcrit = \sum_i [\sum_{j:V} p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - \sum_{j:V} p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})]^2.$$

3.3 The Proposed Robust Methods

3.3.1 Robust Haebara Method

Due to the fact that outlying common items distort scale transformation and eliminating of them leads to unbalanced content representation, a method is proposed to assign weights to common items, so that outlying items carry smaller weights than others. This will allow one to minimize the impact of outliers without removing them so that the content balance is preserved. In robust regression, a weight is assigned to an observation based on the distance between a residual from the OLS regression and the median residual. If the distance is large, the weight is small. Instead of focusing on the estimators in robust regression, I am interested in acquiring weights of the common items. The idea of obtaining weights is the same as the robust regression: a large discrepancy on item parameter estimates between the old and the new test forms is associated with a small weight for the item in computing the coefficients. The procedure does not only down-weight the influence of obvious outlying items, but also considers the influence of other common items with smaller discrepancy.

To implement the weights in the scale transformation, a reiterative weighting method was used. Two elements are essential in the proposed approaches: 1) define a discrepancy between the two test forms for each item, and 2) select a weight function.

The areas enclosed by the two corresponding ICCs after transforming the new test form onto the old scale, as suggested by Linn et al. (1981), serve as the index of the discrepancy between the two test forms. There are two leading estimates of area between two characteristic curves. Raju (1988) suggested two ways to obtain the area by either a signed area or an unsigned area, as reviews in Chapter 2. However, the method is rather computationally expensive. Rudner (1977) suggested an approximated way to acquire the area by using the sum of the rectangular areas for 200 small intervals of θ between -5 and 5. It does not require an equal c-parameter, which fits my requirement well. However, the θ is typically set to within -4 and 4 in the most popular software programs such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002) or ST (Hanson & Zeng, revised by Cui, 2004a) with 31 quadrature points. Therefore, Rudner's approximation is used, except that $\theta \in [-4, 4]$ with 31 quadrature points:

$$Area_j = \sum_q | p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Fi}; a_{Fj}, b_{Fj}, c_{Fj}) | * \Delta\theta,$$

where q indexes the number of quadrature points.

Under the framework of Huber function, the formula weight function is defined as

$$w_i(e_i) = \begin{cases} 1 & | e_i | \le k \\ k / | e_i | & | e_i | > k \end{cases},$$

where

$$e_i = \frac{Area_j}{\sigma},$$

and

$$\sigma = Median(Area_j) / 0.6745.$$

The tuning constant *k* establishes the extent of down-weight. A tuning constant is set to 1.345.

The robust Haebara function, therefore, is defined as the sum of squares of weighted difference of item characteristic curves between the old form and the transformed new form:

$$HC\_Hcrit = \sum_i [\sum_{j:V} \{w_j [p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})]\}^2],$$

To obtain the scale transformation coefficients, *A* and *B*, one needs to minimize the functions above. To serve as initial points for the reiterative estimation procedure, scale transformation coefficients obtained by traditional Haebara method or Stocking-Lord method can be used.

3.3.2 The Least Absolute Values

Instead of minimizing the sum of squares of the difference of item characteristic curves, the approach based on the least absolute deviations which minimizes the absolute difference of two ICCs between the equated test forms:

$$LAV\_Hcrit = \sum_{i}\sum_{j:V} |p_{ij}(\theta_{Ti}; a_{Tj}, b_{Tj}, c_{Tj}) - p_{ij}(\theta_{Ti}; \frac{a_{Fj}}{A}, Ab_{Fj} + B, c_{Fj})| .$$

3.3.3 Robust Deming Approach

The robust mean and sigma method introduced by Stocking & Lord (1983) is a method that implements a regression of estimated item difficulty of the *to* scale on that of the *from* scale. One important property of the test equating is symmetry. That is, if a score x on Form X equates to a score y on Form Y by using a function, then the score y on Form Y will equate to the score x on Form X by using the inverse of the original function. To satisfy this requirement, both directions of scale transformation need to be considered. In statistics, a special type of linear model could be used to obtain the best fit with errors in both variables, and it is called error-in-variables model. In this study, a robust method of scale transformation is proposed under a framework of Deming regression (Deming, 1943).

In linear Deming regression with observations subject to errors on both *x* and *y*, a model is defined as

$$x_j = \xi_j + e_{xj},$$

$$y_j = \eta_j + e_{yj} = \beta_0 + \beta_1 \xi_j + e_{yj},$$

where x_j and y_j are the observed values for individual *j*, ξ_j and η_j are true values for the independent and dependent variables, e_{xj} and e_{yj} are the error terms for the independent and dependent variables, and the β_0 and β_1 are the estimated intercept and slope of the best fit line. The likelihood function of the Deming regression is

$$f(x_1,...,x_n,y_1,...,y_n) = \prod_{j=1}^{n}\left\{\frac{1}{(2\pi\sigma_x)^2}\exp[-\frac{(x_j-\xi_j)^2}{2\sigma_x^2}]\frac{1}{(2\pi\sigma_x)^2}\exp[\frac{(y_j-\beta_0-\beta_1 x_j)^2}{2\sigma_y^2}]\right\}.$$

A critical feature of the new method is that an M-estimates introduced by Huber (1977) is applied to the likelihood function for joint distribution of both *a*- and *b*-parameters under the framework of Deming regression.

For the *to* scale, let

$$a_j = \alpha_j + \varepsilon_{\alpha j},$$

where a_j is the estimated *a*-parameter, α_j is the "true" *a*-parameter, and $\varepsilon_{\alpha j}$ is the error term for the *a*-parameter of item *j*. Assume $\varepsilon_{\alpha j} \sim N(0,\sigma_\alpha^2)$.

For the *from* scale, let

$$\tilde{a}_j = \tilde{\alpha}_j + \tilde{\varepsilon}_{\alpha j},$$

where $\tilde{a}_j$ is the estimated *a*-parameter, $\tilde{\alpha}_j$ is the "true" *a*-parameter, and $\tilde{\varepsilon}_{\alpha j}$ is the error term for the *a*-parameter of item j. Assume $\tilde{\varepsilon}_{\alpha j} \sim N(0, \lambda\sigma_\alpha^2)$. In addition, assume $\lambda = 1$, so that $\tilde{\varepsilon}_{\alpha j} \sim N(0, \sigma_\alpha^2)$.

According to the linear scale transformation, the true *a*-parameters from both test from have the relationship, such that,

$$\alpha_j = \tfrac{1}{A}\tilde{a}_j,$$

so that

$$a_j = \tfrac{1}{A}\tilde{a}_j + \varepsilon_{\alpha j}.$$

Similarly, for the *to* scale, let

$$b_j = \beta_j + \varepsilon_{\beta j},$$

where b_j is the estimated *b*-parameter, β_j is the "true" *b*-parameter, and $\varepsilon_{\beta j}$ is the error term for the *b*-parameter of item *j*. Assume $\varepsilon_{\beta j} \sim N(0, \sigma_\beta^2)$.

For the *from* scale, let

$$\tilde{b}_j = \tilde{\beta}_j + \tilde{\varepsilon}_{\beta j},$$

where $\tilde{b}_j$ is the estimated *b*-parameter, $\tilde{\beta}_j$ is the "true" *b*-parameter, and $\tilde{\varepsilon}_{\beta j}$ is the error term for the *b*-parameter of item *j*. Assume $\tilde{\varepsilon}_{\beta j} \sim N(0, \sigma_\beta^2)$.

According to the linear scale transformation, the true *b*-parameters from both test from have the relationship, such that,

$$\beta_j = A\tilde{\beta}_j + B,$$

so that

$$b_j = A\tilde{\beta}_j + B + \varepsilon_{\beta j}.$$

Based on the likelihood function of Deming regression, the likelihood function for joint distribution of both *a*- and *b*- parameters could be written as

$$f(A, B, \tilde{\alpha}_1, ..., \tilde{\alpha}_n, \tilde{\beta}_1, ..., \tilde{\beta}_n, \sigma_\alpha^2, \sigma_\beta^2)$$
$$= \prod_{j=1}^{n} \left\{ \frac{1}{(2\pi\sigma_\alpha\sigma_\beta)^2} \exp\left[-\frac{(\tilde{a}_j - \tilde{\alpha}_j)^2 + (a_j - \frac{1}{A}\tilde{\alpha}_j)^2}{2\sigma_\alpha^2} - \frac{(\tilde{b}_j - \tilde{\beta}_j)^2 + (b_j - A\tilde{\beta}_j - B)^2}{2\sigma_\beta^2} \right] \right\}.$$

Consequently, the log likelihood function is

$$-2n(\ln\sigma_\alpha + \ln\sigma_\beta) - \frac{1}{2\sigma_\alpha^2}\sum_{j=1}^{n}[(\tilde{a}_j - \tilde{\alpha}_j)^2 + (a_j - \tfrac{1}{A}\tilde{\alpha}_j)^2] - \frac{1}{2\sigma_\beta^2}\sum_{j=1}^{n}[(\tilde{b}_j - \tilde{\beta}_j)^2 + (b_j - A\tilde{\beta}_j - B)^2].$$

To obtain $\tilde{\alpha}_j$, let

$$\frac{\partial \ln L}{\partial \tilde{\alpha}_j} = \frac{\partial}{\partial \tilde{\alpha}_j}[\frac{(\tilde{a}_j - \tilde{\alpha}_j)^2 + (a_j - \frac{1}{A}\tilde{\alpha}_j)^2}{2\sigma_\alpha^2}] = \frac{2(\tilde{a}_j - \tilde{\alpha}_j) + \frac{2}{A}(a_j - \frac{1}{A}\tilde{\alpha}_j)}{2\sigma_\alpha^2}.$$

Set the derivative equal to zero, yields

$$\tilde{\alpha}_j = \frac{A^2\tilde{a}_j + Aa_j}{1+A^2}$$

To obtain $\tilde{\beta}_j$, let

$$\frac{\partial \ln L}{\partial \tilde{\beta}_j} = \frac{\partial}{\partial \tilde{\beta}_j}[\frac{(\tilde{b}_j - \tilde{\beta}_j)^2 + (b_j - A\tilde{\beta}_j - B)^2}{2\sigma_\beta^2}] = \frac{2(\tilde{b}_j - \tilde{\beta}_j) + 2A(b_j - A\tilde{\beta}_j - B)}{2\sigma_\beta^2}.$$

Set the derivative equal to zero, yields

$$\tilde{\beta}_j = \frac{\tilde{b}_j + Ab_j - AB}{1+A^2}.$$

To obtain the transformation coefficients, *A* and *B*, based on the least squares criterion, we need to maximize the log-likelihood function.

According to Huber's M-estimation, the negative log-likelihood function, which is to be minimized to obtain the transformation coefficients, of the joint distribution of both *a*- and *b*- parameters could be written as

$$-\ln L = n\ln(\sigma_\alpha\sigma_\beta\sigma_{\alpha 1}\sigma_{\beta 1}) + \frac{1}{2\sigma_\alpha^2}\sum_{j=1}^{n}(a_j - \tfrac{1}{A}\tilde{\alpha}_j)^2 + \rho(\frac{\tilde{a}_j - \tilde{\alpha}_j}{\sigma_{\alpha 1}}) + \frac{1}{2\sigma_\beta^2}\sum_{j=1}^{n}(b_j - A\tilde{\beta}_j - B)^2 + \rho(\frac{\tilde{b}_j - \tilde{\beta}_j}{\sigma_{\beta 1}}),$$

where ρ is a Huber function, and $\sigma_{\alpha 1}$ and $\sigma_{\beta 1}$ are scale parameters for *a*- and *b*- parameters. The Huber function is defined as

$$\rho(e_j) = \begin{cases} \frac{1}{2}e_j^2 & |e_j| \le k \\ k|e_j| - \frac{1}{2}k^2 & |e_j| > k \end{cases},$$

where e_j indicates a residual for item j, and k is a tuning constant. Smaller values of k produce more resistance to outliers, but decrease efficiency under the normal distribution. In this study, k = 1.345 is chosen to obtain about 90% efficiency when data are normally distributed.

3.4 Simulated Data

3.4.1 Traditional Scale Transformation Methods with Outliers

The purpose of this study was to investigate the performances of the traditional scale transformation methods, i.e., Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord, when a variety of conditions of outlier are simulated. The method with the least equating errors in the presence of outliers was used for further studies.

Data

This simulation studies used item parameters presented by Kolen and Brennan in Table 6.4 of their classic work *Test Equating, Scaling, and Linking* (2004, p. 186). The data include 36 items with 12 common items. Based on the item parameters, the true item parameter values from which item responses were generated. Score conversions from IRT true score equating using these item parameter estimates were assumed to be the population equating relationship.

Assuming ability distribution for the group taking the old test form was $\theta \sim N(0, 1)$, three ability distributions, $\theta \sim N(0, 1)$, $\theta \sim N(0.25, 1.1^2)$, and $\theta \sim N(0.5, 1.2^2)$, were used for the new form to represent magnitudes of ability differences between two groups of examinees who took the new form and the old form. These ability distributions are commonly encountered in practice. Given ability distributions and item parameter values, dichotomous item responses are simulated by assuming a 3PL IRT model with a sample size of 1000 examinees by using R. The simulated item responses were served as the data for this study.

Outlier Manipulation

In most simulated studies in relation to outliers in common items, only the *b*-parameter was manipulated. Wells, Subkoviak & Serlin (2002) used increased *a*- and *b*-parameter in their study focusing on impact of item parameter drift on ability estimate. In this study, both *a*- and *b*- parameters were allowed to vary in both directions to simulate more complicated situations. However, the change of *a*-parameter was relatively small, similar to the commonly encountered conditions in practice. As showed in Table 1, to represent the complexity of the item parameter discrepancy, nine conditions are summarized. In the study, a single common item was randomly selected to be an outlier except for the condition of *no outlier*. To simulate an outlier, the *a*- and *b*- parameters of a randomly selected item were adjusted according to randomly assigned numbers in the outlying range. For example, the original *a*- and *b*- parameter are 1.1445 and -0.1301, respectively. Random generated numbers for a- and b- parameters are 0.2 and -0.7, respectively. As a result, *a*- and *b*- parameters of the outlying item are adjusted to be

1.3445 (= 1.1445 + 0.2) and -0.8301 (= -0.1301 - 0.7), respectively. Last, item responses were simulated according to the adjusted item parameters for 1000 examinees. The entire procedure of simulation is conducted in R environment.

Table 1. Summary of Conditions of Outlier Simulation (Sections 3.4.1 & 3.4.2)

Conditions	Change of *b*- parameter	Change of *a*- parameter
1) No outlier	$\Delta b = 0$	$\Delta a = 0$
2) IMoBIA	$0.5 < \Delta b < 1$	$0.1 < \Delta a < 0.5$
3) IMiBIA	$0.1 < \Delta b < 0.5$	$0.1 < \Delta a < 0.5$
4) DMoBIA	$-1 < \Delta b < -0.5$	$0.1 < \Delta a < 0.5$
5) DMiBIA	$-0.5 < \Delta b < -0.1$	$0.1 < \Delta a < 0.5$
6) IMoBDA	$0.5 < \Delta b < 1$	$-0.5 < \Delta a < -0.1$
7) IMiBDA	$0.1 < \Delta b < 0.5$	$-0.5 < \Delta a < -0.1$
8) DMoBDA	$-1 < \Delta b < -0.5$	$-0.5 < \Delta a < -0.1$
9) DMiBDA	$-0.5 < \Delta b < -0.1$	$-0.5 < \Delta a < -0.1$

Separate estimates of item parameters for the new form need to be put on the same scale of the old form using a scale transformation method. Four traditional scale transformation methods (i.e., Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord) are compared in this study under a variety of conditions of outlying common item.

Totally, 108 conditions are included in this study. The manipulated variations included

- three ability distributions: $\theta \sim N(0, 1)$, $\theta \sim N(0.25, 1.1^2)$, and $\theta \sim N(0.5, 1.2^2)$;
- four scale transformation methods;
- Nine outlier conditions.

The whole procedure was replicated one hundred times in order to control random sampling error. In the case when the calibration didn't converge, a new random sample was drawn as a replacement. Statistics described in the Evaluation Criteria subsection are computed to evaluate the performance of each outlier detection method in different conditions.

IRT True Score Equating

In IRT, the θ_i -equivalent number-correct true scores on test form X is

$$\tau_X(\theta_i) = \sum_{j:X} p_{ij}(\theta_i; a_j, b_j, c_j).$$

In the IRT true score equating, the number-correct true scores on separately calibrated forms are set to be equivalent at a given θ_i (Lord, 1982). The *to* scale true score equivalent of a given true score on the *from* scale is calculated as

$$irt(\tau_F) = \tau_T(\tau_F^{-1}), \quad if \quad \sum_{j:F} c_j < \tau_F < K_F,$$

where τ_F^{-1} is the θ_i corresponding to the true score τ_F , and K_F is the number of items on the *from* form. For the scores outside the range of the possible true scores (τ_F^*), Kolen's ad hoc procedure (1981) was used:

$$irt(\tau_F^*) = \frac{\sum_{j:T} c_j}{\sum_{j:F} c_j} \tau_F^*, \quad if \quad 0 \le \tau_F^* \le \sum_{j:F} c_j,$$

$$irt(\tau_F^*) = K_T, \quad if \quad \tau_F^* = K_F,$$

where K_F and K_T are the numbers of items on the two equated forms, *to* and *from* scales.

Equating Errors

At each score point, the standard error of equating equivalents can be compared among outlier detection methods to see the variation of each method. Denote $\hat{e}_Y(x_i)$ as an estimate of *to* scale (Y, old form) equivalent to a score x_i on *from* scale (X, new form), the corresponding standard error is calculated by

$$S\hat{E}[\hat{e}_Y(x_i)] = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\{\hat{e}_{Y(r)}(x_i) - \hat{\bar{e}}_Y(x_i)\}^2},$$

where

$$\hat{\bar{e}}_Y(x_i) = \frac{1}{R}\sum_{r=1}^{R}\hat{e}_{Y(r)}(x_i),$$

R denotes the number of replications, and *r* indexes each replication. Considering all score points together, the *Weighted Standard Error* is defined as

$$WSE[\hat{e}_Y(x)] = \sum_{i=0}^{K} P_{x_i} \cdot S\hat{E}[\hat{e}_Y(x_i)],$$

where P_{x_i} is the proportion of examinees at score point x_i and *K* is the number of items on the test.

To take into account both systematic error and random error, a *Root Mean Squared Error* (*RMSE*) statistic can also be calculated at each score point as

$$RMSE[\hat{e}_Y(x_i)] = \sqrt{\{S\hat{E}[\hat{e}_Y(x_i)]\}^2 + \{\hat{B}ias[\hat{e}_Y(x_i)]\}^2},$$

where

$$\hat{B}ias[\hat{e}_Y(x_i)] = \hat{\bar{e}}_Y(x_i) - e_Y(x_i).$$

Considering all score points together, the *Weighted Root Mean Squared Error* (*WRMSE*) is defined as

$$WRMSE\left[\hat{e}_Y(x)\right] = \sum_{i=0}^{K} P_{x_i} \cdot RMSE\left[\hat{e}_Y(x_i)\right].$$

The Weighted Absolute Bias (WAB) statistic indexes the overall bias when all score points are considered and is defined as

$$WAB[\hat{e}_Y(x)] = \sum_{i=0}^{K} P_{x_i} \cdot |\hat{B}ias[\hat{e}_Y(x_i)]|.$$

3.4.2 Comparing Robust Methods of Scale Transformation to the Traditional Method

The purpose of this study is to investigate the performances of the proposed robust methods in scale transformation as compared to the traditional scale transformation method with least equating errors when a single outlier is simulated.

The conditions of simulation are the same as those described in Section 3.4.1. Separate estimates of item parameters for the new form need to be put on the same scale of the old form using a scale transformation method. In this particular study, the proposed robust scale transformation methods are compared to the traditional method. Totally, 180 conditions are included in this study. The manipulated variations included

- five ability distributions: $\theta \sim N(0, 1)$, $\theta \sim N(0.25, 1.1^2)$, $\theta \sim N(0.5, 1.2^2)$, $\theta \sim N(-0.25, 1.1^2)$, and $\theta \sim N(-0.5, 1.2^2)$;
- four scale transformation methods (Stocking-Lord, Least Absolute Values, Robust Deming, and Robust Haebara);
- Nine outlier conditions.

As before, the whole procedure was replicated one hundred times in order to control the random sampling error. In the case when the calibration cannot converge, a new random sample was drawn as a replacement. Descriptive statistics of scale transformation coefficients were obtained. To evaluate the accuracy of the equating, the standard error, bias, and root mean square error (RMSE) of equating equivalents are compared. Considering all score points together, the weighted standard error, weighted absolute bias, and weighted RMSE will be compared. Detailed description of the indexes is referred to the Section 3.4.1.

3.4.3 Comparing Robust Scale Transformation to Outlier Removal Methods

The purpose of this study is to investigate the performances of selected robust methods in scale transformation on the basis of Study 2 as compared to the outlier removal methods.

Because differences amongst tested methods were relatively small for a single outlier with mild *b*-parameter changes, a single severe outlier was introduced in this study to replace the mild outlying condition. The nine conditions of a single outlier are summarized in Table 2.

Table 2. Summary of Conditions of Outlier Simulation (Section 3.4.3)

Conditions	Change of *b*- parameter	Change of *a*- parameter
1) No outlier	$\Delta b = 0$	$\Delta a = 0$
2) IMoBIA	$0.5 < \Delta b < 1$	$0.1 < \Delta a < 0.5$
3) ISBIA	$1 < \Delta b < 1.5$	$0.1 < \Delta a < 0.5$
4) DMoBIA	$-1 < \Delta b < -0.5$	$0.1 < \Delta a < 0.5$
5) DSBIA	$-1.5 < \Delta b < -1$	$0.1 < \Delta a < 0.5$
6) IMoBDA	$0.5 < \Delta b < 1$	$-0.5 < \Delta a < -0.1$
7) ISBDA	$1 < \Delta b < 1.5$	$-0.5 < \Delta a < -0.1$
8) DMoBDA	$-1 < \Delta b < -0.5$	$-0.5 < \Delta a < -0.1$
9) DSBDA	$-1.5 < \Delta b < -1$	$-0.5 < \Delta a < -0.1$

Separate estimates of item parameters for the new form need to be put on the same scale of the old form using a scale transformation method. Five scenarios handling the outlier and conducting scale transformations are considered in the study: 1) Baseline (entire common item set, Stocking-Lord method), 2) Robust (entire common item set, Robust Haebara method), 3) LAV (entire common item set, least absolute values method), 4) Displacement (identify and eliminate outlier(s) by displacement method (Murphy et al., 2010) due to its simplicity of application, Stocking-Lord method), and 5) Exclusion (directly eliminate the simulated outlying common item, Stocking-Lord method). By using the displacement method, differences of the estimated a- and b-parameters of the common item set between the *from* scale and the *to* scale are calculated. An item is considered as an outlier if its absolute difference of either a- or b- parameter between two test forms is larger than 0.5.

The procedure used in this particular study is as follows:

1. obtain item parameters (*a*, *b*, and *c*) for both new and old forms;

2. randomly choose one common items as an outlier for further analysis, except for the condition of no outlier;
3. simulate ability parameters (θ) for the examinee groups taking the new form;
4. generate dichotomous item responses using the new form item parameters and simulated examinee abilities θ;
5. calibrate the item responses with the 3PL IRT model using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002) with a sample control card shown in the Appendix;
6. conduct scale transformation using either the traditional method or the robust methods;
7. identify and remove outliers from the common items using the displacement method, and re-conduct scale transformation using the traditional method;
8. conduct IRT true score equating (Kolen & Brennan, 2004).

The whole procedure is replicated one hundred times in order to control random sampling error. In the case when the calibration didn't converge, a new random sample is drawn as a replacement. To evaluate the accuracy of the equating, the standard error, bias, and root mean square error (RMSE) of equating equivalents are compared. Considering all score points together, the weighted standard error, weighted absolute bias, and weighted RMSE will be compared. Detailed description of the indexes is referred to the Study 1.

3.5 Empirical Data

The purpose of this study was to investigate whether the proposed methods are comparable to the current Haebara and Stocking-Lord methods when implemented to empirical data such as the *College Basic Academic Subject Examination* (*CBASE*).

The *CBASE* is currently required by the State of Missouri to assess the qualifications of individuals who want to enter teacher education programs. Additionally, as a nationwide examination, it is used by over 130 colleges or universities. The examination provides not only information about the diagnostic assessment of a college student's knowledge and skills in mathematics, English, Science, and Social Studies, but also information about the reasoning competencies of the student. The complete test battery includes 180 multiple-choice test items for all four subjects. The English test includes 41 items, the mathematics test includes 56 items, the science test includes 41 items, and the social science test includes 42 items. The examinees' responses are dichotomized: 1 for correct answers and 0 for incorrect answers. There are several forms of the examination. Two forms, LF and LM are chosen, and they are defined as the old and the new test forms in this study. Two subjects, mathematics and English are studied. The Mathematics test includes 14 common items, and the English test includes 17 common items. The examinees' responses are dichotomized, given 1 for correct answers and 0 for incorrect answers.

The scale transformation coefficients and transformed item parameters are compared between the proposed methods and the traditional Stocking-Lord methods. In addition, equated scores and reported scores based on the equating results are compared.

CHAPTER 4

RESULTS

In this chapter, results based on the data and procedures described in Chapter 3 are summarized in two main sections: studies with simulated data and a study of empirical *CBASE* data. The research questions are discussed:

1. How do the traditional scale transformation methods, i.e., Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord, perform when a variety of conditions of outlier are simulated?
2. Do the proposed robust methods perform better than the traditional scale transformation methods in the presence of simulated outlying common? Do the robust methods perform as well as the traditional scale transformation methods when there is no outlying common item?
3. Do the new robust methods perform better than or at least close to the current outlier manipulation process (identification and elimination) in the presence of simulated outlying common item?
4. Can the proposed robust methods be used for empirical data such as the *CBASE*?

4.1 Simulation Studies

4.1.1 Traditional Scale Transformation Methods with a Single Outlier

The purpose of this study was to investigate the accuracy of IRT true score equating when different traditional scale transformation methods were used. Equating accuracy was determined by the error indices including root mean square error (RMSE) of

equating, which is a combination of bias (systematic error) and standard error (random error).

Weighted Indices of Equating Errors

Table 3. Weighted RMSE Statistics for IRT True Score Equating with Traditional Scale Transformation

Theta distribution	Outlier	Mean/Mean	Stocking-Lord	Haebara	Mean/Sigma
N(0,1)	*1) No outlier*	0.300	0.171	0.172	0.326
	2) IMoBIA	0.544	0.511	0.416	0.478
	3) IMiBIA	0.379	0.280	0.244	0.360
	4) DMoBIA	0.516	0.490	0.399	0.457
	5) DMiBIA	0.362	0.278	0.251	0.360
	6) IMoBDA	0.428	0.239	0.359	0.506
	7) IMiBDA	0.335	0.206	0.268	0.419
	8) DMoBDA	0.460	0.267	0.380	0.554
	9) DMiBDA	0.342	0.215	0.248	0.445
$N(0.25,1.1^2)$	*1) No outlier*	0.272	0.167	0.169	0.319
	2) IMoBIA	0.500	0.451	0.368	0.436
	3) IMiBIA	0.364	0.251	0.236	0.322
	4) DMoBIA	0.503	0.454	0.387	0.486
	5) DMiBIA	0.317	0.260	0.233	0.334
	6) IMoBDA	0.415	0.241	0.379	0.530
	7) IMiBDA	0.350	0.219	0.268	0.410
	8) DMoBDA	0.448	0.239	0.375	0.537
	9) DMiBDA	0.315	0.209	0.258	0.375
$N(0.5,1.2^2)$	*1) No outlier*	0.253	0.158	0.164	0.280
	2) IMoBIA	0.526	0.448	0.386	0.447
	3) IMiBIA	0.307	0.247	0.246	0.278
	4) DMoBIA	0.544	0.450	0.396	0.482
	5) DMiBIA	0.359	0.240	0.232	0.329
	6) IMoBDA	0.334	0.223	0.385	0.479
	7) IMiBDA	0.264	0.186	0.269	0.371
	8) DMoBDA	0.361	0.209	0.350	0.469
	9) DMiBDA	0.308	0.204	0.283	0.395

The weighted indices of equating errors are based on all score points. To show the overall impact of outlier on equating, the weighted root mean square error, weighted

absolute bias, and weighted standard error of equating indices are presented in Tables 3 through 5.

As can be seen from Table 3, the Stocking-Lord method and the Haebara method of scale transformation had the least weighted equating errors when no outlier was simulated. The moment methods, Mean/Mean and Mean/Sigma, generally had larger equating errors than the characteristic curve methods. The moment methods had also doubled weighted RMSE than the characteristic-curve based methods. Between the two moment methods, Mean/Sigma had relatively larger equating errors when no outlier was simulated. When a single outlying common item was simulated, the equating errors increased according to the magnitude of the outlying common item as expected. When a-parameter was increased, the Haebara method performed the best and the Mean/Mean the worst in equating among the traditional methods. When a-parameter was reduced, the Stocking-Lord method performed the best and the Mean/Sigma the worst in equating among the traditional methods. Again, the characteristic curve methods outperformed the moment methods. These findings are consistent among the ability distributions. It also indicates that equating errors slightly decreased when the difference between two groups was large.

Table 4 shows that, when a single outlying common item was simulated, the equating bias increased. The characteristic curves methods usually had smaller weighted absolute bias than the moment methods, particularly when the group differences were large. The Mean/Sigma method had the least weighted absolute bias especially under the conditions of mildly increased *b*-parameter ($0.1 < \Delta b < 0.5$) with increased *a*-parameter ($0.1 < \Delta a <$

0.5). Similar to the pattern found in the weighted RMSE, the Mean/Mean had the largest bias when a-parameter increased, and the Mean/Sigma had the largest bias when a-parameter reduced.

Table 4. Weighted Bias Statistics for IRT True Score Equating with Traditional Scale Transformation

Theta distribution	Outlier	Mean/Mean	Stocking-Lord	Haebara	Mean/Sigma
N(0,1)	*1) No outlier*	0.037	0.031	0.026	0.050
	2) IMoBIA	0.354	0.290	0.234	0.237
	3) IMiBIA	0.149	0.115	0.096	0.085
	4) DMoBIA	0.309	0.262	0.210	0.195
	5) DMiBIA	0.132	0.114	0.095	0.057
	6) IMoBDA	0.222	0.075	0.159	0.228
	7) IMiBDA	0.127	0.076	0.074	0.175
	8) DMoBDA	0.268	0.111	0.168	0.300
	9) DMiBDA	0.132	0.070	0.062	0.173
$N(0.25,1.1^2)$	*1) No outlier*	0.049	0.040	0.044	0.060
	2) IMoBIA	0.329	0.237	0.186	0.215
	3) IMiBIA	0.146	0.099	0.080	0.066
	4) DMoBIA	0.327	0.248	0.209	0.213
	5) DMiBIA	0.142	0.116	0.099	0.082
	6) IMoBDA	0.211	0.080	0.139	0.239
	7) IMiBDA	0.137	0.068	0.032	0.172
	8) DMoBDA	0.244	0.075	0.145	0.253
	9) DMiBDA	0.103	0.060	0.043	0.141
$N(0.5,1.2^2)$	*1) No outlier*	0.033	0.028	0.029	0.044
	2) IMoBIA	0.361	0.233	0.191	0.234
	3) IMiBIA	0.147	0.112	0.102	0.088
	4) DMoBIA	0.385	0.248	0.212	0.259
	5) DMiBIA	0.170	0.112	0.104	0.115
	6) IMoBDA	0.147	0.060	0.157	0.232
	7) IMiBDA	0.071	0.037	0.020	0.164
	8) DMoBDA	0.167	0.046	0.143	0.193
	9) DMiBDA	0.068	0.028	0.034	0.134

As showed in Table 5, weighted standard errors of equating had a very similar pattern to the weighted RMSE, except that the Mean/Sigma method consistently had the largest weighted standard errors in all conditions in this study. The Haebara method had the least

weighted standard error when a-parameter was increased, while Stocking-Lord method had the least weighted standard error when a-parameter was reduced. The finds are consistent among the ability distributions.

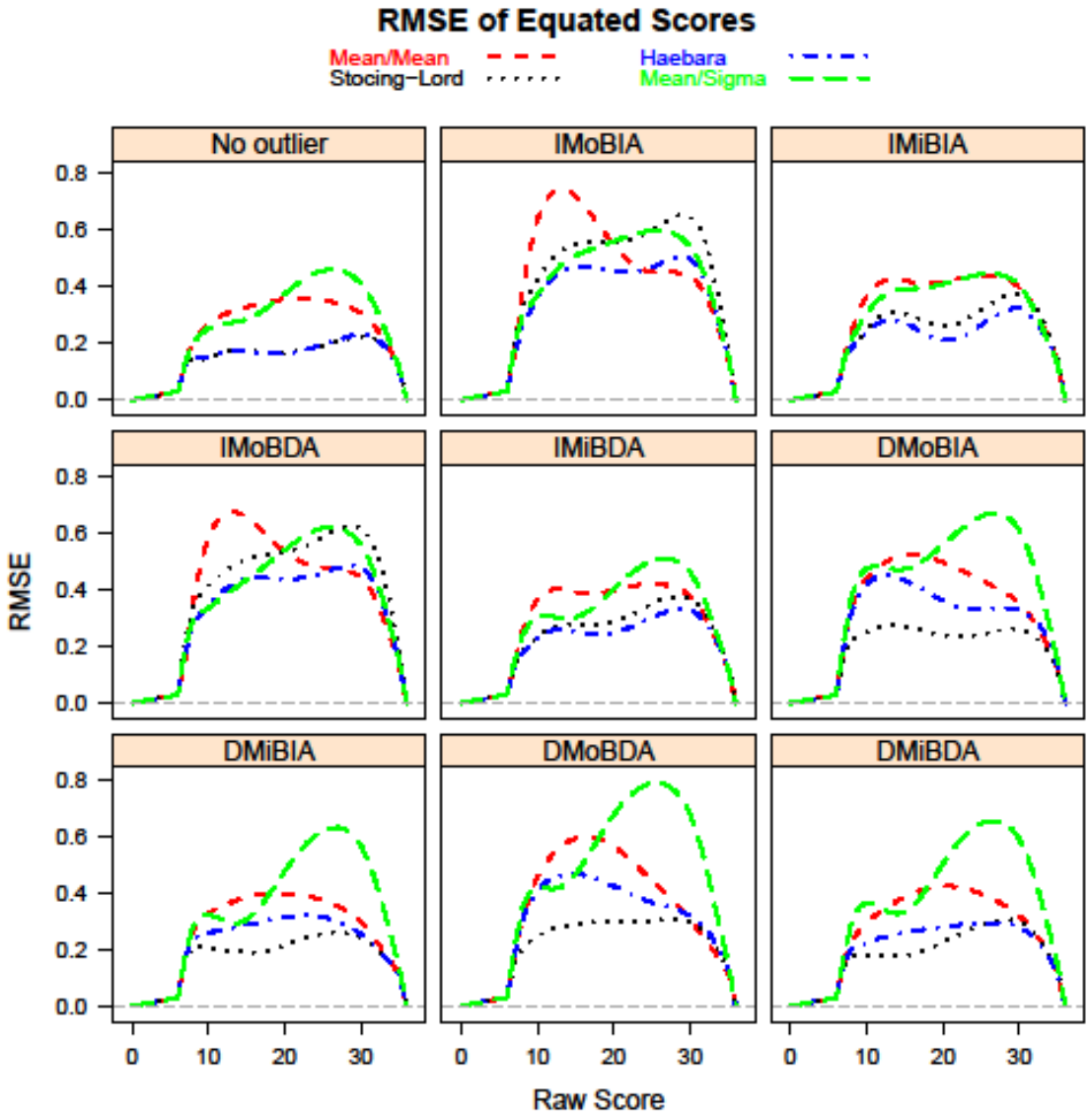
Table 5. Weighted Standard Error Statistics for IRT True Score Equating with Traditional Scale Transformation

Theta distribution	Outlier	Mean/Mean	Stocking-Lord	Haebara	Mean/Sigma
N(0,1)	*1) No outlier*	0.297	0.168	0.169	0.321
	2) IMoBIA	0.344	0.390	0.310	0.394
	3) IMiBIA	0.338	0.244	0.214	0.347
	4) DMoBIA	0.360	0.389	0.313	0.401
	5) DMiBIA	0.328	0.245	0.224	0.353
	6) IMoBDA	0.338	0.223	0.304	0.426
	7) IMiBDA	0.306	0.190	0.256	0.362
	8) DMoBDA	0.348	0.238	0.321	0.419
	9) DMiBDA	0.310	0.200	0.239	0.393
$N(0.25,1.1^2)$	*1) No outlier*	0.266	0.161	0.161	0.311
	2) IMoBIA	0.311	0.358	0.289	0.361
	3) IMiBIA	0.320	0.222	0.217	0.312
	4) DMoBIA	0.323	0.350	0.296	0.420
	5) DMiBIA	0.269	0.223	0.201	0.320
	6) IMoBDA	0.331	0.225	0.338	0.447
	7) IMiBDA	0.317	0.204	0.266	0.354
	8) DMoBDA	0.341	0.223	0.332	0.443
	9) DMiBDA	0.289	0.199	0.252	0.335
$N(0.5,1.2^2)$	*1) No outlier*	0.249	0.155	0.161	0.275
	2) IMoBIA	0.305	0.360	0.308	0.360
	3) IMiBIA	0.251	0.213	0.213	0.258
	4) DMoBIA	0.302	0.349	0.303	0.379
	5) DMiBIA	0.296	0.201	0.195	0.302
	6) IMoBDA	0.282	0.214	0.334	0.396
	7) IMiBDA	0.249	0.180	0.267	0.321
	8) DMoBDA	0.299	0.204	0.304	0.408
	9) DMiBDA	0.296	0.201	0.279	0.364

Indices of Equating Errors

Plots of RMSE at each score point for IRT true score equating for both test forms are shown in Figures 1 to 3, plots of bias at each score point for IRT true score equating for both test forms are shown in Figures 4 to 6, and plots of standard error at each score point for IRT true score equating for both test forms are shown in Figures 7 to 9.

Figure 1. The RMSE statistics of IRT equating ($\theta \sim N(0,1)$)



Note that equating errors of raw score between 0 and 4 were consistently low in all conditions. In this range of raw score, the effect of c-parameter on total scores is below chance, which leads the scores to fall outside the range of possible true scores. As a consequence, the Kolen's ad hoc procedure (1981) was used to manipulate this part of scores.

Figure 2. The RMSE statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

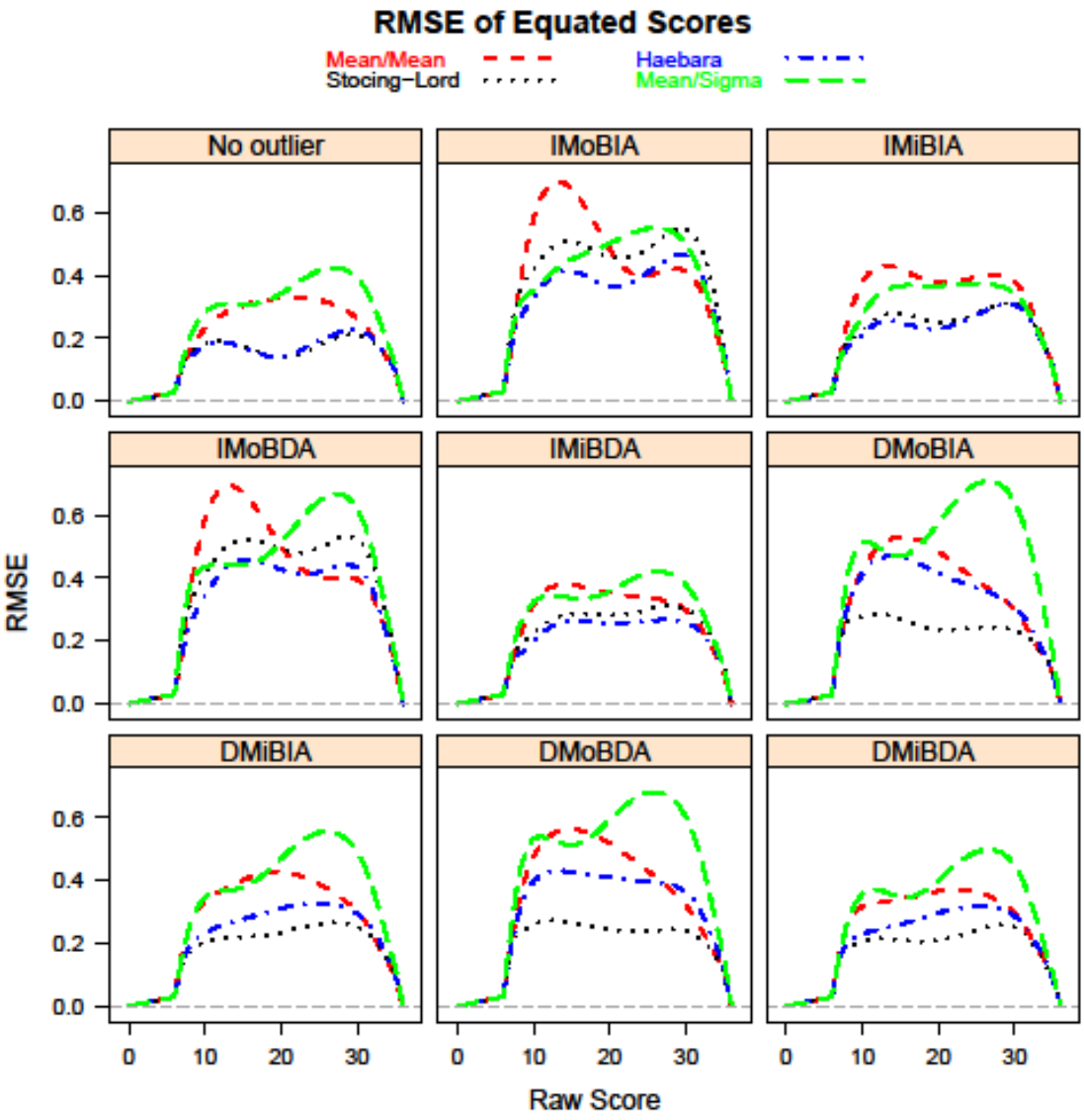
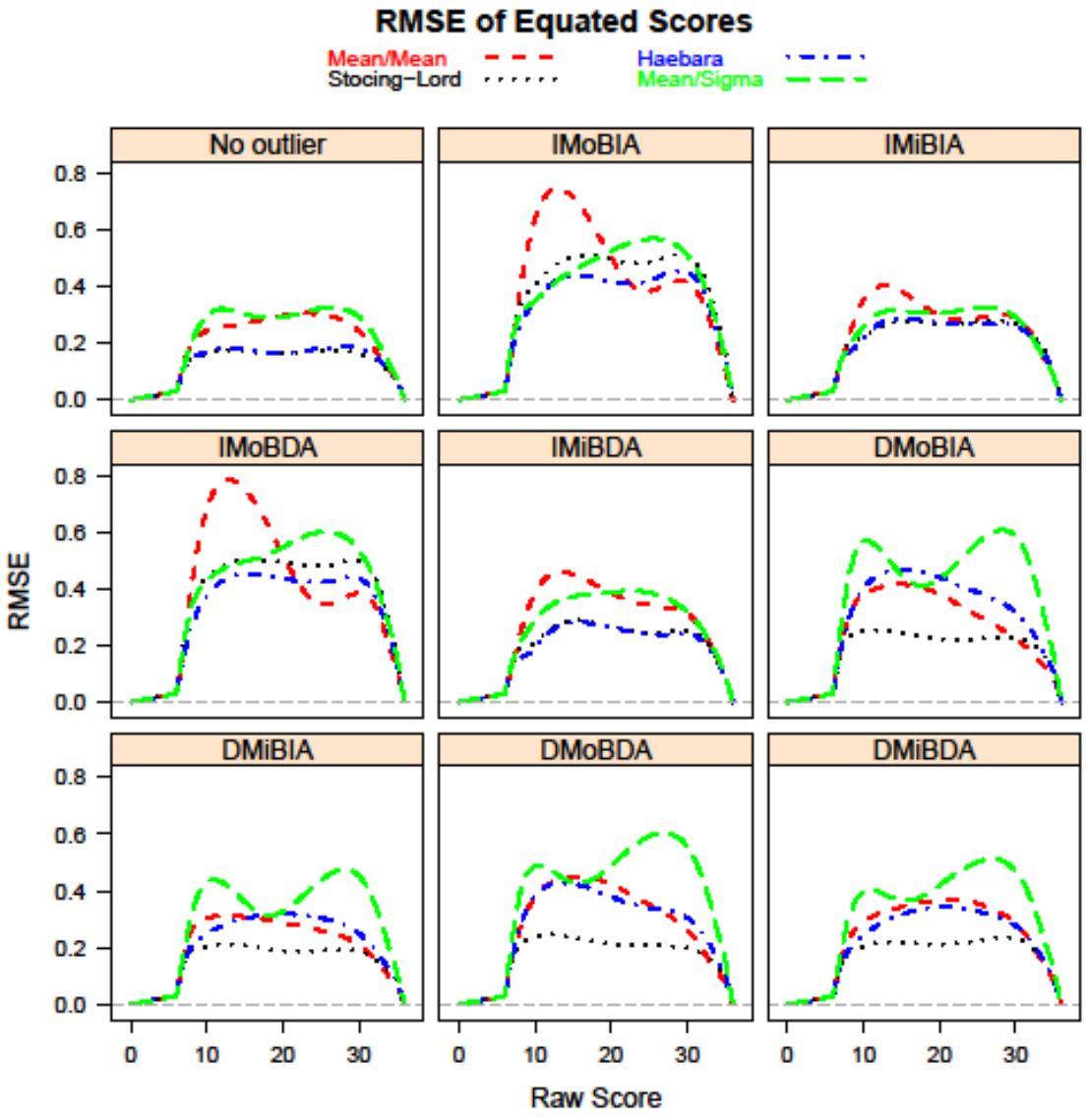


Figure 3. The RMSE statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)



The RMSE represents the overall error in the test equating. Figure 1 presents the RMSE results for the equated scores under different outlier situations and methods of scale transformation when the two groups of examinees are very similar. The horizontal axis is the raw score of the equated scores, and the vertical axis is the RMSE. The figure indicates that when no outlier was simulated, the characteristic curve method performed

better than the moment methods throughout the entire scale, and the Haebara method was nearly the same as the Stocking-Lord method. The Mean/Sigma method had very large RMSE when raw scores are larger than 20.

When a single outlying item parameter existed in the common item set, the equating errors dramatically increased. The RMSE of outlier removal and robust approach are smaller than the baseline when outlier is presented in the common item set. In general, the characteristic curve methods had smaller equating errors as expected. However, the characteristic curve methods differed when the a-parameter changed in different directions. When a-parameter was reduced, the Stocking-Lord method performed much better than the Haebara method. Although RMSE curves of the characteristic methods were close to each other throughout the entire scale when a-parameter was increased, the Haebara method produced slightly smaller RMSE than the Stocking-Lord method. When an outlying common item was included in scale transformation, the Mean/Mean method had enlarged RMSE for the low scores, and the Mean/Sigma method had enlarged RMSE for the high scores, especially when moderate b-parameter change was introduced in the common item set. When the ability difference increases (Figures 2 and 3), the pattern is very similar to what was observed in Figure 1.

The equating bias indicates the difference between the mean of repeated application of a particular equated score and the corresponding true score. Figure 4 shows the bias plot for the equated scores under different outlier situations and methods of the traditional scale transformation when the two groups of examinees are similar. In general, four scale transformation methods produced approximately identical absolute biases when no

outlier is simulated, and the four bias curves were very close to each other. When a single outlier was simulated in the common item set, it impaired the equating accuracy by introducing more systematic errors to the equating procedure after transformation. The characteristic methods had smaller absolute values of bias than the moment methods, and the two absolute bias curves of the characteristic methods were close to each other throughout the entire scale when a-parameter was increased. However, the Stocking-Lord method produced substantially smaller absolute bias than the Haebara method. The direction of varied absolute bias usually was opposite to the direction of b-parameter changes, except that the moment methods had reverse directions at given score ranges.

When the two groups were more dissimilar (Figures 5 and 6), the pattern was very similar to what was observed in Figure 4.

Figure 4. The Bias statistics of IRT equating ($\theta \sim N(0,1)$)

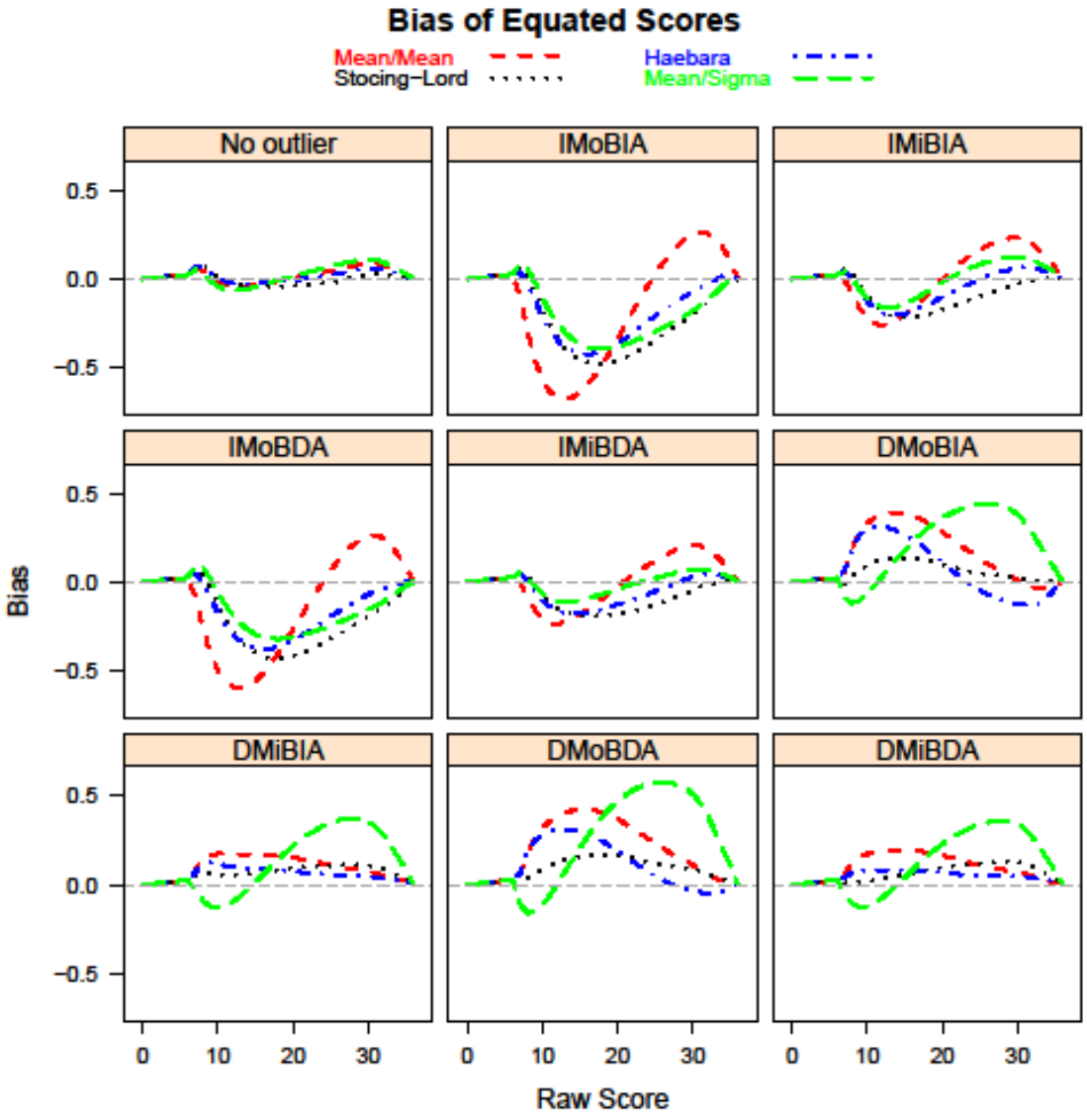


Figure 5. The Bias statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

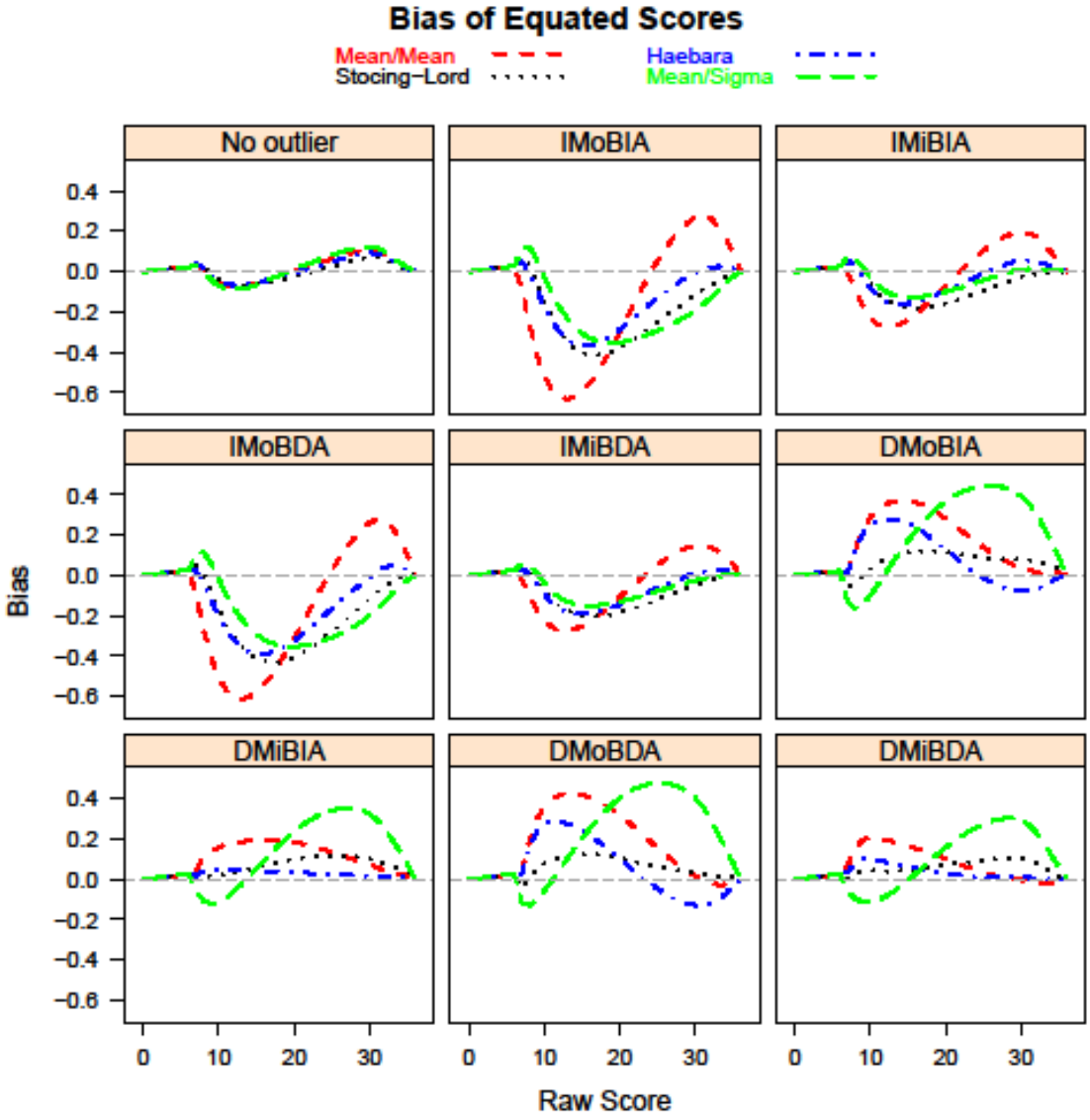
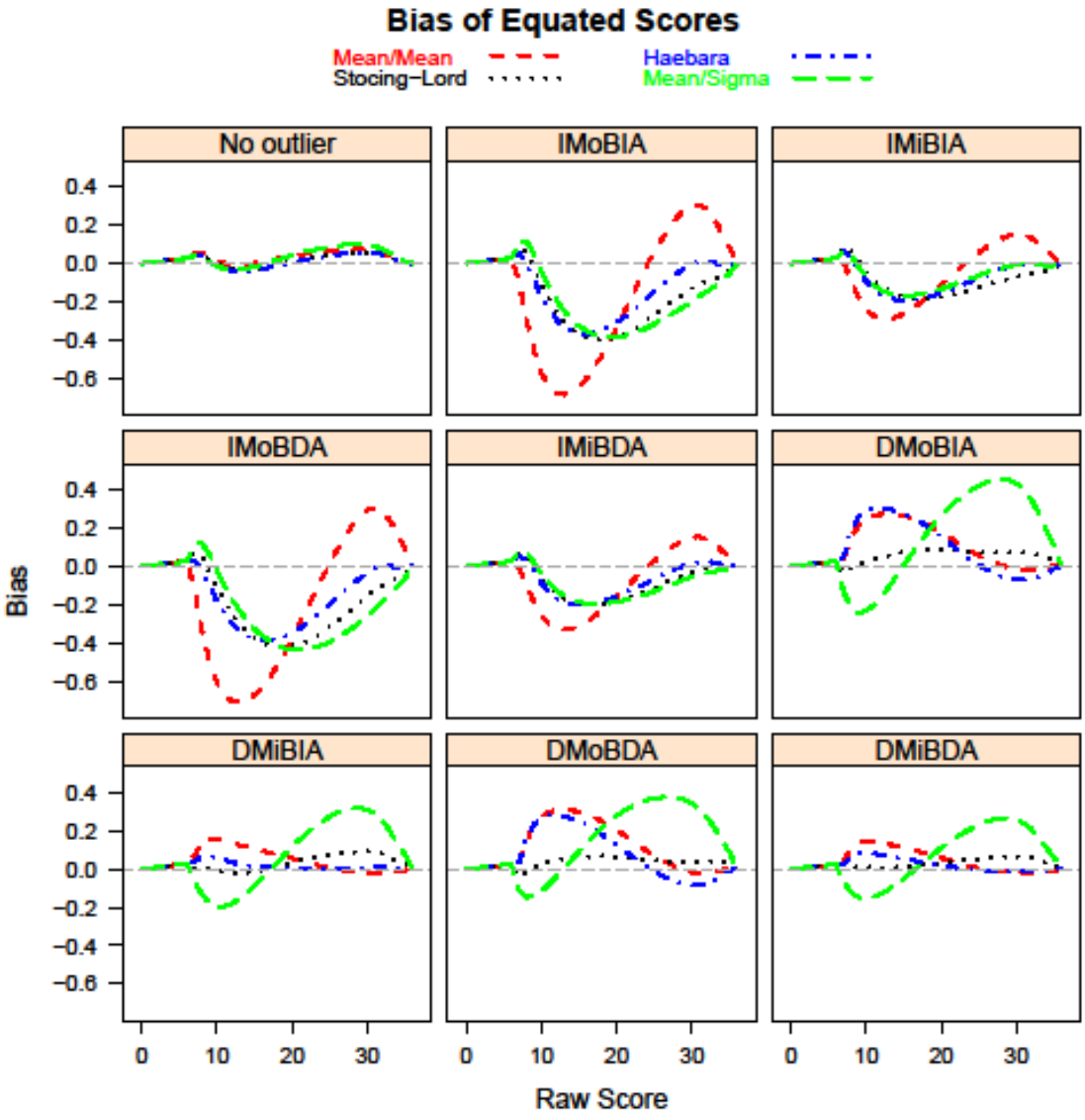


Figure 6. The Bias statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)



Standard error of the equating procedure represents the random error. As showed in Figures 7 to 9, the inclusion of a single outlying common item also increased standard errors at most of the score points, regardless of the ability distributions of the examinees and conditions of outlying common items. The characteristic curve methods produced smaller random errors in equating.

Figure 7. The Standard Error statistics of IRT equating ($\theta \sim N(0,1)$)

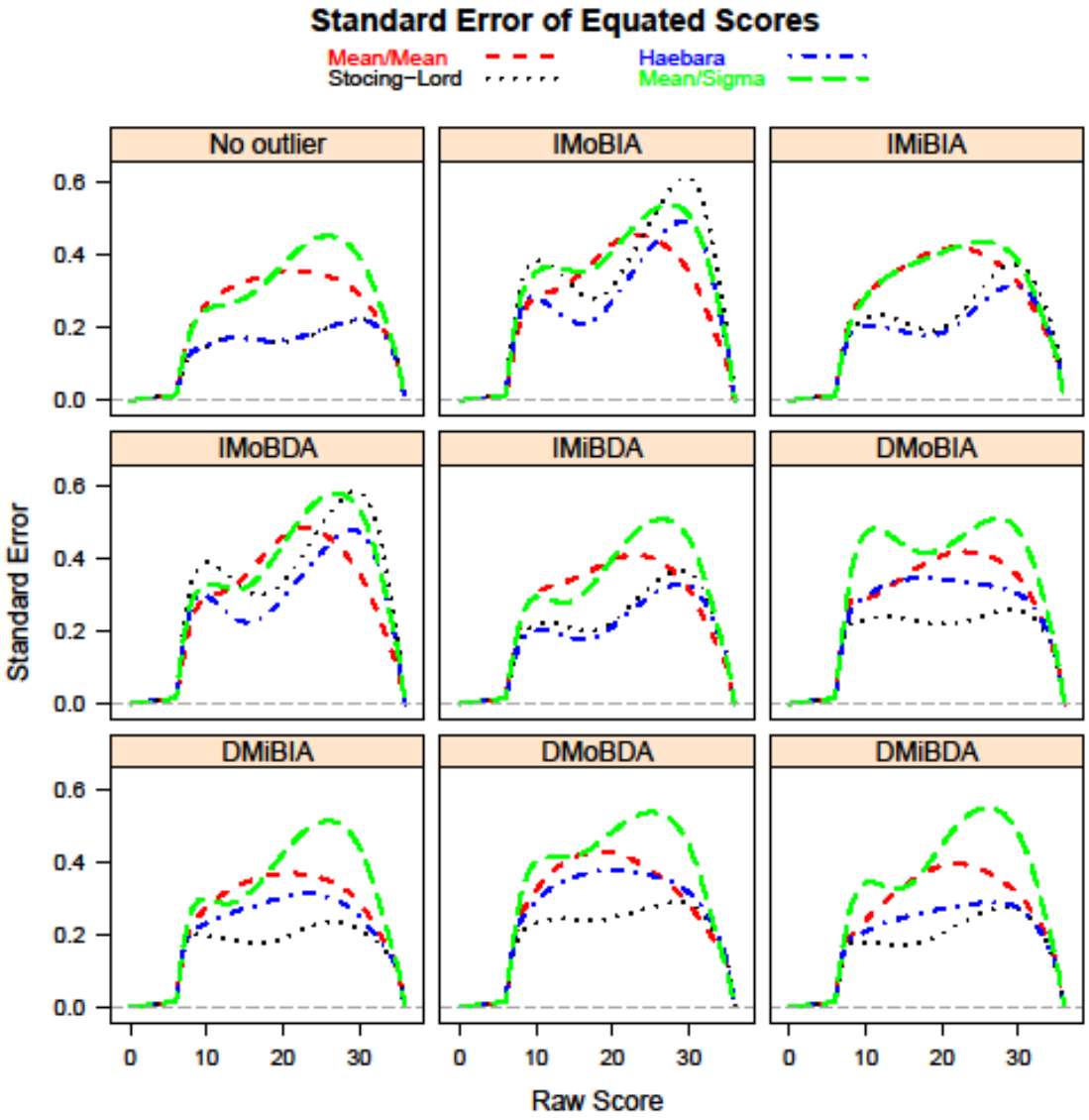


Figure 8. The Standard Error statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

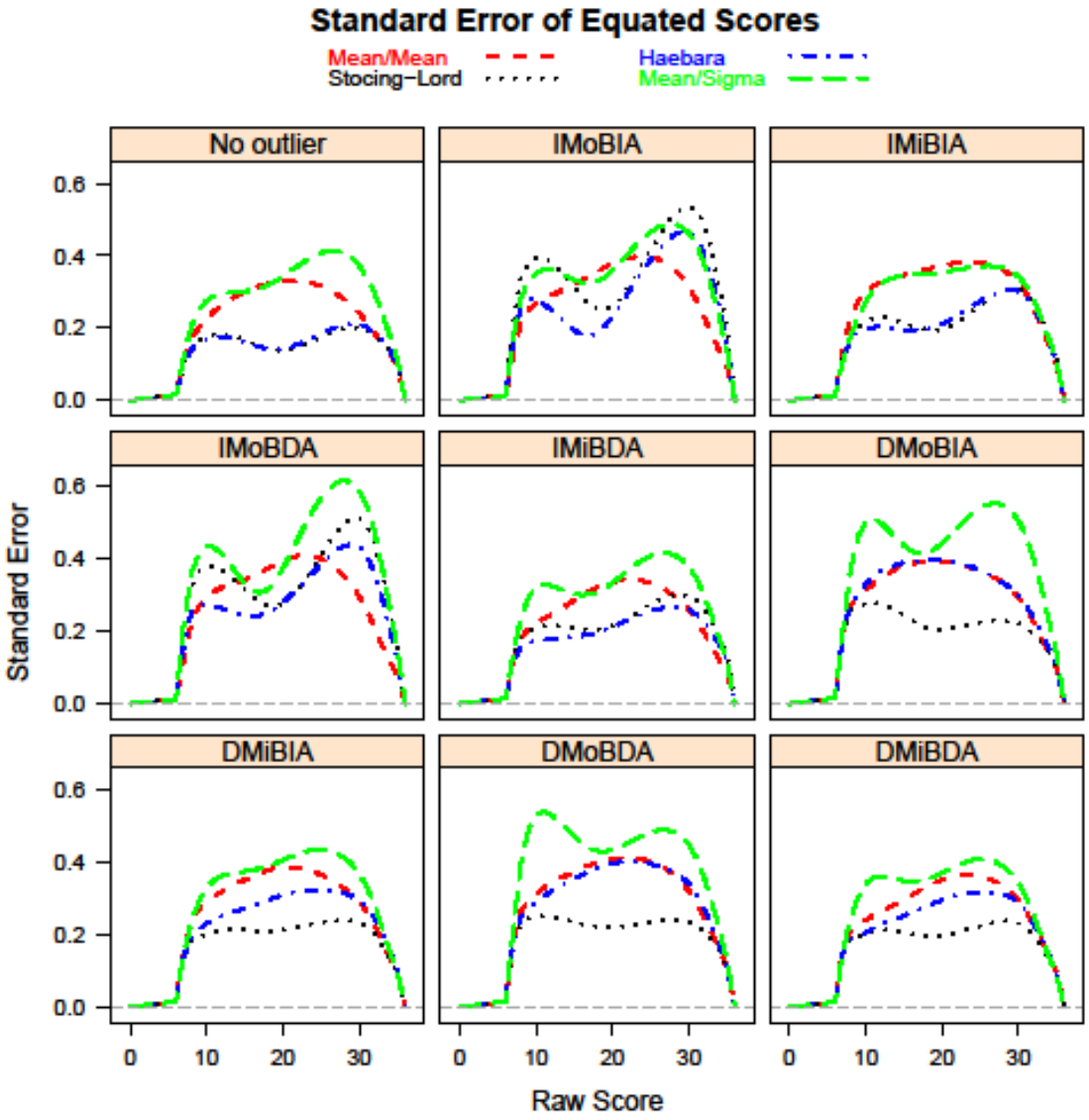
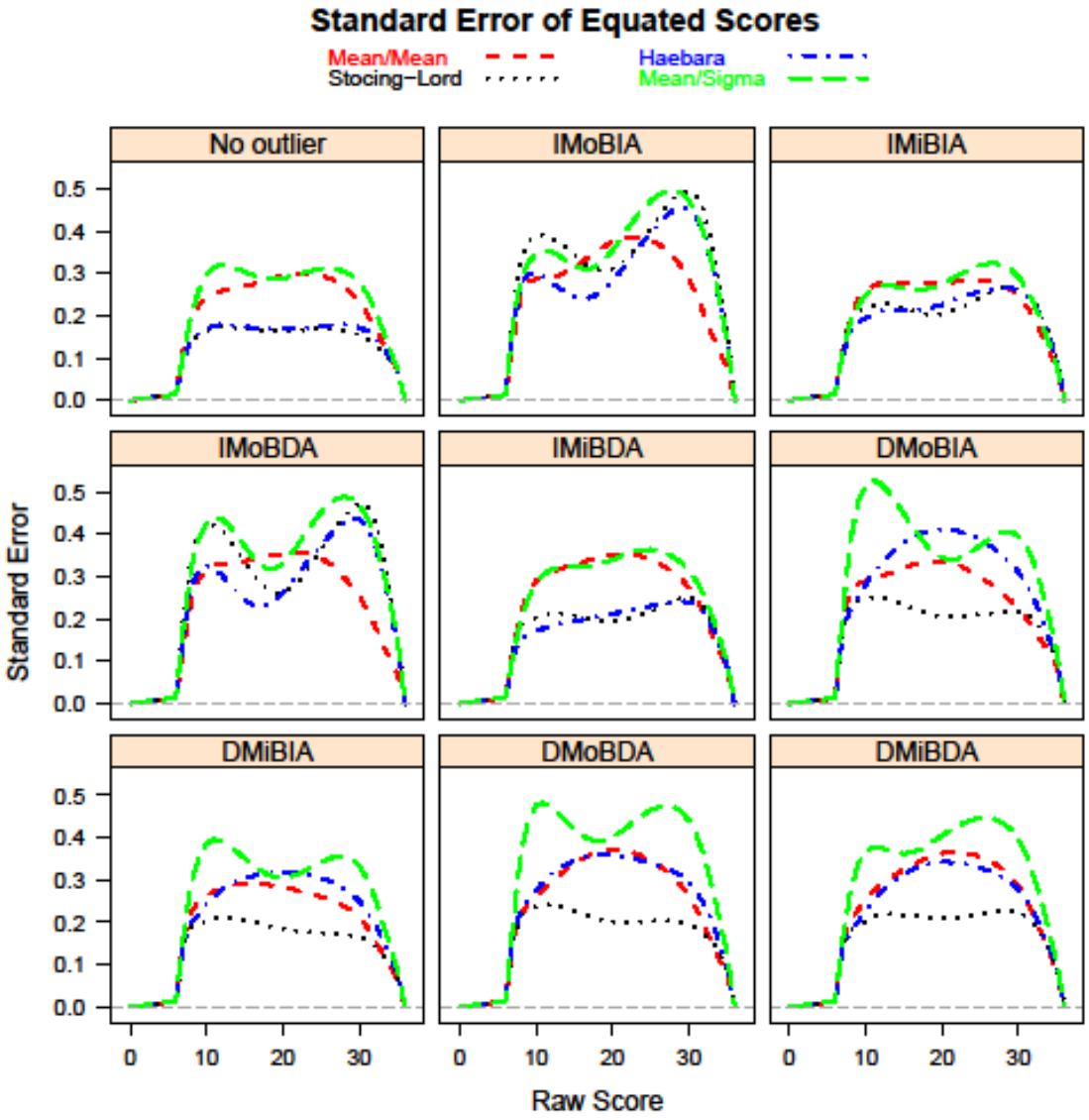


Figure 9. The Standard Error statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)



Due to the fact that the Stocking-Lord method consistently produced smaller equating errors, it was used as a baseline for the following studies when robust methods were introduced.

4.1.2 Performance of the Proposed Robust Methods of Scale Transformation

In this section, performances in terms of scale transformation as well as subsequent equating of the proposed robust methods and the selected traditional method of Stocking-Lord, were compared under various conditions of outlier manipulation and ability distributions of examinees. The results for the coefficients of scale transformation are presented first, followed by the results for the equating errors including both weighted indices and individual indices.

Coefficients of Scale Transformation

The coefficients of scale transformation were estimated by the proposed robust methods and the Stocking-Lord method. The mean and standard deviation of the scale transformation coefficients were obtained from 100 sets of the estimates of the scale transformation coefficients.

In the simulated study, the population values were assumed to be known. The observed values of item responses were simulated based on the 3PL IRT model with various ability distributions: $\theta \sim$ N(0, 1), $\theta \sim$ N(0.25, 1.1^2), $\theta \sim$ N(0. 5, 1.2^2), $\theta \sim$ N(-0.25, 1.1^2), $\theta \sim$ N(-0. 5, 1.2^2). If estimation is accurate, the estimated A coefficients should be the value of standard deviation of a given ability distribution and B the value of mean of that ability distribution. For instance, the estimation of scale transformation coefficients should be A = 1.0 and B = 0 for $\theta \sim$ N(0, 1), providing the estimation is perfect. Similarly, A = 1.1 and B = 0.25 should be obtained if the estimation is perfect for $\theta \sim$ N(0.25, 1.1^2).

If a method has a larger discrepancy between the estimated coefficients and the true ones, it produced larger bias.

Table 6. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0,1)$

Outlier		Stocking-Lord		Robust Deming		LAV		Robust Haebara	
		A	B	A	B	A	B	A	B
1) No outlier	*M*	1.007	-0.008	1.014	-0.005	1.006	-0.007	1.003	-0.008
	SD	0.035	0.037	0.041	0.047	0.037	0.039	0.036	0.038
2) IMoBIA	*M*	1.016	-0.065	1.034	-0.040	1.015	-0.003	1.020	-0.008
	SD	0.074	0.055	0.058	0.047	0.047	0.042	0.060	0.047
3) IMiBIA	*M*	1.018	-0.035	1.030	-0.022	1.010	-0.007	1.011	-0.012
	SD	0.045	0.051	0.044	0.053	0.045	0.051	0.054	0.053
4) DMoBIA	*M*	1.013	-0.071	1.027	-0.043	1.005	-0.006	1.011	-0.011
	SD	0.066	0.055	0.050	0.052	0.042	0.046	0.053	0.047
5) DMiBIA	*M*	1.018	-0.031	1.024	-0.026	1.010	-0.005	1.011	-0.011
	SD	0.046	0.052	0.048	0.056	0.041	0.047	0.049	0.050
6) IMoBDA	*M*	0.949	0.079	1.004	0.042	1.001	0.007	0.989	0.015
	SD	0.049	0.073	0.044	0.060	0.037	0.048	0.047	0.049
7) IMiBDA	*M*	0.992	0.031	1.014	0.017	1.011	0.005	1.005	0.005
	SD	0.040	0.058	0.046	0.050	0.043	0.048	0.045	0.044
8) DMoBDA	*M*	0.965	0.065	1.017	0.033	1.011	0.001	1.006	0.003
	SD	0.044	0.073	0.040	0.052	0.035	0.041	0.044	0.044
9) DMiBDA	*M*	0.994	0.027	1.022	0.020	1.009	0.001	1.009	0.004
	SD	0.046	0.059	0.046	0.051	0.040	0.040	0.044	0.041

Table 7. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0.25,1.1^2)$

Outlier		Stocking-Lord		Robust Deming		LAV		Robust Haebara	
		A	B	A	B	A	B	A	B
1) No outlier	*M*	1.112	0.252	1.116	0.254	1.111	0.254	1.108	0.254
	SD	0.042	0.039	0.045	0.051	0.049	0.042	0.047	0.040
2) IMoBIA	*M*	1.132	0.182	1.135	0.207	1.110	0.241	1.120	0.238
	SD	0.084	0.060	0.051	0.049	0.043	0.043	0.053	0.049
3) IMiBIA	*M*	1.115	0.226	1.118	0.233	1.100	0.247	1.103	0.243
	SD	0.047	0.048	0.045	0.050	0.044	0.047	0.044	0.046
4) DMoBIA	*M*	1.137	0.177	1.136	0.206	1.113	0.240	1.120	0.236
	SD	0.089	0.058	0.052	0.049	0.044	0.044	0.051	0.051
5) DMiBIA	*M*	1.108	0.222	1.114	0.231	1.099	0.241	1.102	0.239
	SD	0.051	0.050	0.048	0.051	0.042	0.046	0.049	0.047
6) IMoBDA	*M*	1.047	0.289	1.108	0.279	1.099	0.250	1.093	0.249
	SD	0.049	0.078	0.053	0.051	0.040	0.045	0.049	0.042
7) IMiBDA	*M*	1.078	0.267	1.105	0.264	1.101	0.253	1.096	0.257
	SD	0.048	0.062	0.052	0.047	0.051	0.042	0.057	0.042
8) DMoBDA	*M*	1.043	0.298	1.106	0.284	1.101	0.260	1.091	0.260
	SD	0.056	0.077	0.051	0.052	0.048	0.040	0.050	0.042
9) DMiBDA	*M*	1.079	0.265	1.111	0.263	1.102	0.250	1.092	0.254
	SD	0.044	0.060	0.045	0.055	0.040	0.046	0.041	0.046

Table 8. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0.5,1.2^2)$

Outlier		Stocking-Lord		Robust Deming		LAV		Robust Haebara	
		A	B	A	B	A	B	A	B
1) No outlier	*M*	1.199	0.496	1.203	0.499	1.198	0.496	1.193	0.496
	SD	0.038	0.041	0.042	0.045	0.045	0.044	0.044	0.042
2) IMoBIA	*M*	1.211	0.442	1.233	0.472	1.210	0.502	1.208	0.504
	SD	0.089	0.057	0.059	0.046	0.054	0.042	0.060	0.047
3) IMiBIA	*M*	1.210	0.471	1.224	0.490	1.205	0.494	1.200	0.491
	SD	0.048	0.055	0.047	0.054	0.049	0.052	0.052	0.055
4) DMoBIA	*M*	1.204	0.437	1.224	0.473	1.201	0.497	1.205	0.501
	SD	0.076	0.059	0.050	0.061	0.047	0.048	0.055	0.053
5) DMiBIA	*M*	1.215	0.480	1.223	0.487	1.205	0.503	1.206	0.500
	SD	0.054	0.055	0.051	0.059	0.052	0.052	0.057	0.055
6) IMoBDA	*M*	1.128	0.550	1.199	0.536	1.189	0.509	1.181	0.508
	SD	0.065	0.075	0.052	0.064	0.050	0.053	0.055	0.054
7) IMiBDA	*M*	1.182	0.522	1.207	0.516	1.205	0.503	1.199	0.505
	SD	0.048	0.061	0.048	0.055	0.050	0.052	0.048	0.049
8) DMoBDA	*M*	1.145	0.537	1.212	0.527	1.199	0.498	1.193	0.500
	SD	0.055	0.075	0.047	0.052	0.040	0.044	0.047	0.048
9) DMiBDA	*M*	1.185	0.518	1.212	0.517	1.205	0.503	1.204	0.505
	SD	0.055	0.061	0.051	0.051	0.052	0.041	0.055	0.041

Table 9. Mean (M) and Standard Deviation (SD) of Scale Transformation Coefficients, 100 replications, $\theta \sim N(-0.25,1.1^2)$

Outlier		Stocking-Lord		Robust Deming		LAV		Robust Haebara	
		A	B	A	B	A	B	A	B
1) No outlier	*M*	1.100	-0.258	1.107	-0.258	1.097	-0.254	1.096	-0.255
	SD	0.038	0.041	0.047	0.051	0.039	0.048	0.041	0.043
2) IMoBIA	*M*	1.109	-0.310	1.129	-0.295	1.109	-0.251	1.121	-0.260
	SD	0.082	0.063	0.066	0.049	0.050	0.046	0.061	0.053
3) IMiBIA	*M*	1.112	-0.281	1.123	-0.274	1.104	-0.252	1.105	-0.258
	SD	0.050	0.056	0.051	0.056	0.051	0.055	0.060	0.058
4) DMoBIA	*M*	1.105	-0.318	1.122	-0.299	1.101	-0.256	1.108	-0.259
	SD	0.073	0.064	0.056	0.056	0.044	0.051	0.052	0.053
5) DMiBIA	*M*	1.110	-0.277	1.116	-0.276	1.104	-0.250	1.107	-0.256
	SD	0.050	0.057	0.054	0.056	0.045	0.049	0.047	0.051
6) IMoBDA	*M*	1.036	-0.151	1.099	-0.205	1.098	-0.238	1.088	-0.235
	SD	0.056	0.076	0.047	0.059	0.040	0.052	0.050	0.052
7) IMiBDA	*M*	1.081	-0.209	1.107	-0.230	1.106	-0.243	1.100	-0.238
	SD	0.045	0.061	0.052	0.054	0.042	0.051	0.049	0.054
8) DMoBDA	*M*	1.051	-0.166	1.108	-0.215	1.103	-0.244	1.092	-0.240
	SD	0.049	0.076	0.048	0.051	0.040	0.045	0.050	0.047
9) DMiBDA	*M*	1.086	-0.215	1.119	-0.230	1.103	-0.247	1.100	-0.242
	SD	0.053	0.065	0.053	0.055	0.049	0.046	0.052	0.051

Table 10. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(-0.5,1.2^2)$

Outlier		Stocking-Lord		Robust Deming		LAV		Robust Haebara	
		A	B	A	B	A	B	A	B
1) No outlier	*M*	1.198	-0.495	1.205	-0.501	1.198	-0.495	1.196	-0.495
	SD	0.043	0.042	0.053	0.044	0.046	0.044	0.044	0.041
2) IMoBIA	*M*	1.210	-0.569	1.214	-0.559	1.190	-0.500	1.204	-0.506
	SD	0.100	0.095	0.072	0.064	0.051	0.057	0.064	0.060
3) IMiBIA	*M*	1.209	-0.523	1.212	-0.524	1.197	-0.498	1.196	-0.500
	SD	0.058	0.056	0.055	0.053	0.049	0.048	0.058	0.049
4) DMoBIA	*M*	1.211	-0.569	1.207	-0.546	1.186	-0.499	1.194	-0.502
	SD	0.094	0.082	0.070	0.060	0.050	0.054	0.061	0.056
5) DMiBIA	*M*	1.207	-0.531	1.214	-0.530	1.196	-0.508	1.201	-0.516
	SD	0.057	0.053	0.064	0.052	0.052	0.053	0.056	0.051
6) IMoBDA	*M*	1.124	-0.401	1.192	-0.462	1.194	-0.497	1.179	-0.486
	SD	0.057	0.084	0.067	0.056	0.052	0.051	0.060	0.053
7) IMiBDA	*M*	1.165	-0.449	1.191	-0.474	1.190	-0.479	1.183	-0.476
	SD	0.055	0.066	0.064	0.051	0.054	0.052	0.057	0.053
8) DMoBDA	*M*	1.123	-0.388	1.187	-0.442	1.186	-0.479	1.172	-0.469
	SD	0.055	0.082	0.065	0.063	0.052	0.049	0.062	0.054
9) DMiBDA	*M*	1.169	-0.461	1.192	-0.485	1.193	-0.496	1.186	-0.491
	SD	0.055	0.074	0.061	0.057	0.051	0.061	0.057	0.058

Patterns of biases in scale transformation coefficients are similar for the different ability distributions. Tables 6 through 10 indicate that the estimated coefficients of scale transformation by the studied methods are almost identical to each other and slightly biased under the condition of *No outlier*. The robust Deming method, however, has a larger bias than the others. In addition, the standard deviations of the coefficients obtained by the methods are similar, except that of the robust Deming method is relatively larger than others. When a single outlier was simulated, both scale transformation coefficients, particularly *B*, obtained by the Stocking-Lord method dramatically deviated from the population values. In addition, the Stocking-Lord method usually had larger standard deviations when a single outlier was included in the scale

transformation. On the other hand, the scale transformation coefficient *A* obtained by the robust Deming method deviated from the population values more than the other two robust methods. When *a*-parameter increased (*IMiBIA*, *DMiBIA*, *IMoBIA*, and *DMoBIA*), the coefficient *A* usually also increased but the *B* decreased. When *a*-parameter decreased (*IMiBDA*, *DMiBDA*, *IMoBDA*, and *DMoBDA*), the coefficient *A* decreased but the *B* increased. The scale transformation coefficients obtained by the LAV method and the robust Haebara method were similar, and they were closer to the population values.

Weighted Indices of Equating Errors

Tables 11 through 13 summarize the weighted root mean square error, weighted absolute bias, and weighted standard error of equating. Table 11 clearly indicates that the robust methods produced relatively larger overall weighted RMSE than the Stocking-Lord method when there was no outlier simulated. The robust Deming method produced larger weighted RMSE than the other robust methods. When a single outlier was simulated, the LAV method consistently produced the least overall weighted RMSE among the investigated scale transformation methods, regardless of the direction of *a*- and *b*- parameter changes and the magnitude of the *b*-parameter change. The weighted RMSEs obtained by the robust Haebara method were slightly larger than the LAV method. When a single item's *b*-parameter was mildly changed (*IMiBIA*, *DMiBIA*, *IMiBDA*, and *DMiBDA*), the robust Deming method produced the largest weighted RMSE. When the *b*-parameter was moderately change (*IMoBIA*, *DMoBIA*, *IMoBDA*, and *DMoBDA*), the Stocking-Lord method had the largest weighted RMSE. If two groups of

examinees are different, the pattern mostly held except that the Stocking-Lord method also produced the largest weighted RMSE for the condition of *DMiBDA*.

Table 11a. Weighted RMSE Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Stocking-Lord	Robust Deming	LAV	Robust Haebara
N(0,1)	*1) No outlier*	0.172	0.266	0.197	0.190
	2) IMoBIA	0.410	0.357	0.250	0.295
	3) IMiBIA	0.253	0.306	0.229	0.259
	4) DMoBIA	0.418	0.347	0.244	0.285
	5) DMiBIA	0.273	0.352	0.226	0.267
	6) IMoBDA	0.452	0.342	0.250	0.259
	7) IMiBDA	0.292	0.315	0.239	0.248
	8) DMoBDA	0.415	0.345	0.269	0.292
	9) DMiBDA	0.307	0.310	0.224	0.231
$N(0.25,1.1^2)$	*1) No outlier*	0.161	0.248	0.195	0.191
	2) IMoBIA	0.439	0.330	0.219	0.280
	3) IMiBIA	0.241	0.291	0.212	0.213
	4) DMoBIA	0.478	0.333	0.241	0.274
	5) DMiBIA	0.250	0.282	0.209	0.233
	6) IMoBDA	0.421	0.344	0.275	0.297
	7) IMiBDA	0.291	0.304	0.227	0.246
	8) DMoBDA	0.469	0.333	0.260	0.276
	9) DMiBDA	0.288	0.280	0.209	0.228
$N(0.5,1.2^2)$	*1) No outlier*	0.162	0.235	0.204	0.186
	2) IMoBIA	0.408	0.347	0.240	0.294
	3) IMiBIA	0.248	0.286	0.231	0.251
	4) DMoBIA	0.413	0.336	0.248	0.302
	5) DMiBIA	0.248	0.308	0.227	0.247
	6) IMoBDA	0.459	0.327	0.265	0.272
	7) IMiBDA	0.282	0.288	0.241	0.242
	8) DMoBDA	0.411	0.333	0.263	0.290
	9) DMiBDA	0.306	0.279	0.233	0.258

Table 11b. Weighted RMSE Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Stocking-Lord	Robust Deming	LAV	Robust Haebara
$N(-0.25,1.1^2)$	*1) No outlier*	0.186	0.277	0.217	0.204
	2) IMoBIA	0.398	0.358	0.248	0.294
	3) IMiBIA	0.249	0.301	0.230	0.274
	4) DMoBIA	0.409	0.358	0.259	0.309
	5) DMiBIA	0.274	0.346	0.236	0.261
	6) IMoBDA	0.458	0.337	0.264	0.296
	7) IMiBDA	0.307	0.317	0.231	0.258
	8) DMoBDA	0.434	0.354	0.278	0.311
	9) DMiBDA	0.315	0.306	0.235	0.246
$N(-0.5,1.2^2)$	*1) No outlier*	0.175	0.276	0.221	0.192
	2) IMoBIA	0.448	0.420	0.266	0.307
	3) IMiBIA	0.265	0.306	0.218	0.227
	4) DMoBIA	0.422	0.369	0.261	0.307
	5) DMiBIA	0.260	0.342	0.241	0.248
	6) IMoBDA	0.469	0.379	0.308	0.336
	7) IMiBDA	0.293	0.332	0.241	0.272
	8) DMoBDA	0.461	0.403	0.280	0.323
	9) DMiBDA	0.313	0.288	0.241	0.254

Table 12 indicates that the robust Deming method of scale transformation had the largest weighted bias and the LAV method yielded the smallest weighted bias under the condition of *No outlier*. When a single outlying common item was simulated, the equating bias significantly increased even for a single outlier with mild *b*-parameter changes. The traditional Stocking-Lord method constantly produced the largest weighted absolute bias among the investigated methods. The proposed robust methods generally reduced the bias, and the robust Haebara method usually produced the least weighted absolute bias. The general pattern is found in almost all ability distributions with a few exceptions. Interestingly, the LAV method had the least weighted bias under the conditions of increasing *a*-parameters with $\theta \sim N(-0.5, 1.2^2)$.

Table 12a. Weighted Bias Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Stocking-Lord	Robust Deming	LAV	Robust Haebara
N(0,1)	*1) No outlier*	0.016	0.036	0.010	0.021
	2) IMoBIA	0.271	0.142	0.095	0.070
	3) IMiBIA	0.126	0.088	0.032	0.016
	4) DMoBIA	0.278	0.134	0.101	0.076
	5) DMiBIA	0.143	0.110	0.019	0.045
	6) IMoBDA	0.296	0.118	0.107	0.075
	7) IMiBDA	0.125	0.071	0.036	0.021
	8) DMoBDA	0.254	0.137	0.107	0.093
	9) DMiBDA	0.123	0.111	0.022	0.011
$N(0.25,1.1^2)$	*1) No outlier*	0.027	0.034	0.020	0.014
	2) IMoBIA	0.278	0.157	0.072	0.057
	3) IMiBIA	0.109	0.084	0.026	0.018
	4) DMoBIA	0.315	0.161	0.062	0.041
	5) DMiBIA	0.110	0.069	0.014	0.020
	6) IMoBDA	0.208	0.140	0.088	0.076
	7) IMiBDA	0.076	0.055	0.042	0.012
	8) DMoBDA	0.231	0.122	0.082	0.055
	9) DMiBDA	0.090	0.083	0.036	0.021
$N(0.5,1.2^2)$	*1) No outlier*	0.015	0.020	0.014	0.020
	2) IMoBIA	0.266	0.149	0.097	0.110
	3) IMiBIA	0.131	0.086	0.023	0.013
	4) DMoBIA	0.274	0.132	0.090	0.106
	5) DMiBIA	0.130	0.106	0.023	0.014
	6) IMoBDA	0.293	0.098	0.086	0.074
	7) IMiBDA	0.114	0.052	0.052	0.024
	8) DMoBDA	0.229	0.116	0.115	0.093
	9) DMiBDA	0.114	0.071	0.032	0.024

Table 12b. Weighted Bias Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Stocking-Lord	Robust Deming	LAV	Robust Haebara
$N(-0.25,1.1^2)$	*1) No outlier*	0.023	0.030	0.017	0.023
	2) IMoBIA	0.256	0.150	0.099	0.085
	3) IMiBIA	0.114	0.081	0.044	0.017
	4) DMoBIA	0.264	0.146	0.097	0.095
	5) DMiBIA	0.139	0.119	0.024	0.036
	6) IMoBDA	0.299	0.111	0.112	0.108
	7) IMiBDA	0.140	0.076	0.046	0.021
	8) DMoBDA	0.271	0.127	0.109	0.091
	9) DMiBDA	0.127	0.114	0.032	0.011
$N(-0.5,1.2^2)$	*1) No outlier*	0.030	0.049	0.030	0.032
	2) IMoBIA	0.264	0.207	0.076	0.086
	3) IMiBIA	0.113	0.113	0.036	0.039
	4) DMoBIA	0.262	0.147	0.069	0.076
	5) DMiBIA	0.113	0.099	0.036	0.055
	6) IMoBDA	0.250	0.120	0.134	0.099
	7) IMiBDA	0.104	0.037	0.027	0.018
	8) DMoBDA	0.269	0.147	0.101	0.074
	9) DMiBDA	0.127	0.058	0.036	0.023

Table 13 shows that the weighted standard errors had a similar pattern to that of weighted RMSE. It indicates that the Stocking-Lord method had the least weighted standard errors of equating and the proposed robust methods introduced more random errors to some extent under the condition of *No outlier*. When a single outlying common item was included in the scale transformation, random errors produced by the Stocking-Lord method dramatically increased. The proposed LAV method normally produced the least weighted standard error among the investigated methods. The robust Haebara method yielded similar amount of random errors. When the *b*-parameter change of an outlying common item was mild (*IMiBIA*, *DMiBIA*, *IMiBDA*, and *DMiBDA*), the robust Deming method generally produced the largest weighted standard error. When the *b*-

parameter change was moderate (*IMoBIA*, *DMoBIA*, *IMoBDA*, and *DMoBDA*), the Stocking-Lord method had the largest weighted RMSE.

Table 13a. Weighted Standard Error Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Stocking-Lord	Robust Deming	LAV	Robust Haebara
N(0,1)	*1) No outlier*	0.172	0.266	0.197	0.189
	2) IMoBIA	0.410	0.357	0.250	0.278
	3) IMiBIA	0.253	0.306	0.229	0.258
	4) DMoBIA	0.418	0.347	0.244	0.270
	5) DMiBIA	0.273	0.352	0.226	0.263
	6) IMoBDA	0.452	0.342	0.250	0.247
	7) IMiBDA	0.292	0.315	0.239	0.247
	8) DMoBDA	0.415	0.345	0.269	0.273
	9) DMiBDA	0.307	0.310	0.224	0.231
$N(0.25,1.1^2)$	*1) No outlier*	0.161	0.248	0.195	0.190
	2) IMoBIA	0.439	0.330	0.219	0.268
	3) IMiBIA	0.241	0.291	0.212	0.212
	4) DMoBIA	0.478	0.333	0.241	0.267
	5) DMiBIA	0.250	0.282	0.209	0.232
	6) IMoBDA	0.421	0.344	0.275	0.285
	7) IMiBDA	0.291	0.304	0.227	0.245
	8) DMoBDA	0.469	0.333	0.260	0.269
	9) DMiBDA	0.288	0.280	0.209	0.227
$N(0.5,1.2^2)$	*1) No outlier*	0.162	0.235	0.204	0.185
	2) IMoBIA	0.408	0.347	0.240	0.270
	3) IMiBIA	0.248	0.286	0.231	0.250
	4) DMoBIA	0.413	0.336	0.248	0.276
	5) DMiBIA	0.248	0.308	0.227	0.246
	6) IMoBDA	0.459	0.327	0.265	0.261
	7) IMiBDA	0.282	0.288	0.241	0.241
	8) DMoBDA	0.411	0.333	0.263	0.272
	9) DMiBDA	0.306	0.279	0.233	0.257

Table 13b. Weighted Standard Error Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Stocking-Lord	Robust Deming	LAV	Robust Haebara
$N(-0.25,1.1^2)$	*1) No outlier*	0.186	0.277	0.217	0.202
	2) IMoBIA	0.398	0.358	0.248	0.269
	3) IMiBIA	0.249	0.301	0.230	0.273
	4) DMoBIA	0.409	0.358	0.259	0.287
	5) DMiBIA	0.274	0.346	0.236	0.258
	6) IMoBDA	0.458	0.337	0.264	0.273
	7) IMiBDA	0.307	0.317	0.231	0.256
	8) DMoBDA	0.434	0.354	0.278	0.295
	9) DMiBDA	0.315	0.306	0.235	0.246
$N(-0.5,1.2^2)$	*1) No outlier*	0.175	0.276	0.221	0.189
	2) IMoBIA	0.448	0.420	0.266	0.287
	3) IMiBIA	0.265	0.306	0.218	0.223
	4) DMoBIA	0.422	0.369	0.261	0.296
	5) DMiBIA	0.260	0.342	0.241	0.241
	6) IMoBDA	0.469	0.379	0.308	0.320
	7) IMiBDA	0.293	0.332	0.241	0.271
	8) DMoBDA	0.461	0.403	0.280	0.313
	9) DMiBDA	0.313	0.288	0.241	0.252

Indices of Equating Errors

The plots of RMSE at each score point for the IRT true score equating for both test forms are shown in Figures 10 to 14, the plots of bias at each score point for the IRT true score equating for both test forms are shown in Figures 15 to 19, and the plots of standard error at each score point for the IRT true score equating for both test forms are shown in Figures 20 to 24. The ability differences do not have much impact on the comparison among the scale transformation methods in terms of the equating errors. Although shapes of the error curves are somewhat different with various ability distributions, the differences in equating errors are similar among the scale transformation methods, and they are similar to what was observed when $\theta \sim N(0,1)$. As a consequence, results with

large ability differences are not separately discussed in detail, although the figures are presented.

Figure 10. The RMSE statistics of IRT equating ($\theta \sim N(0,1)$)

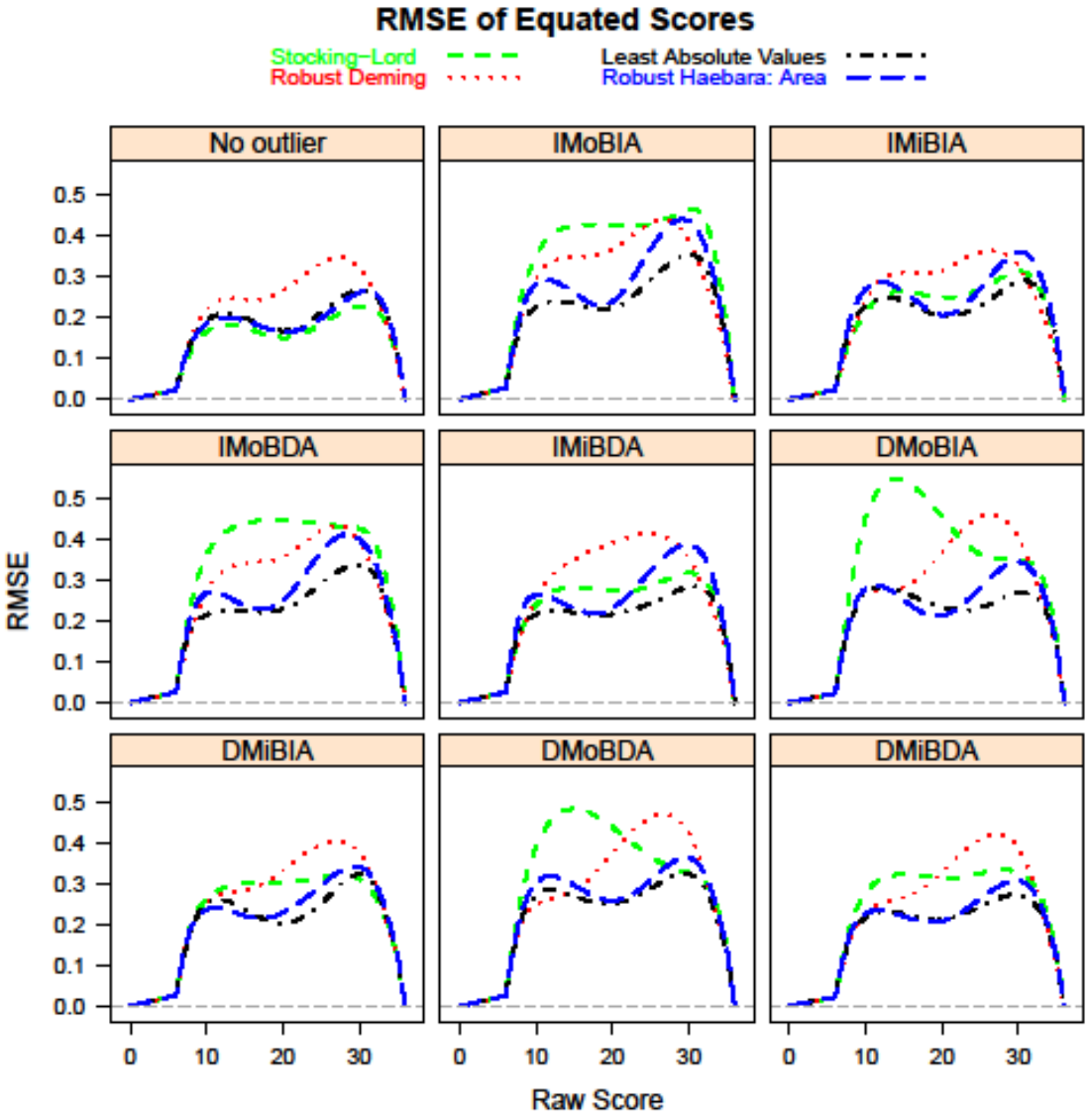


Figure 11. The RMSE statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

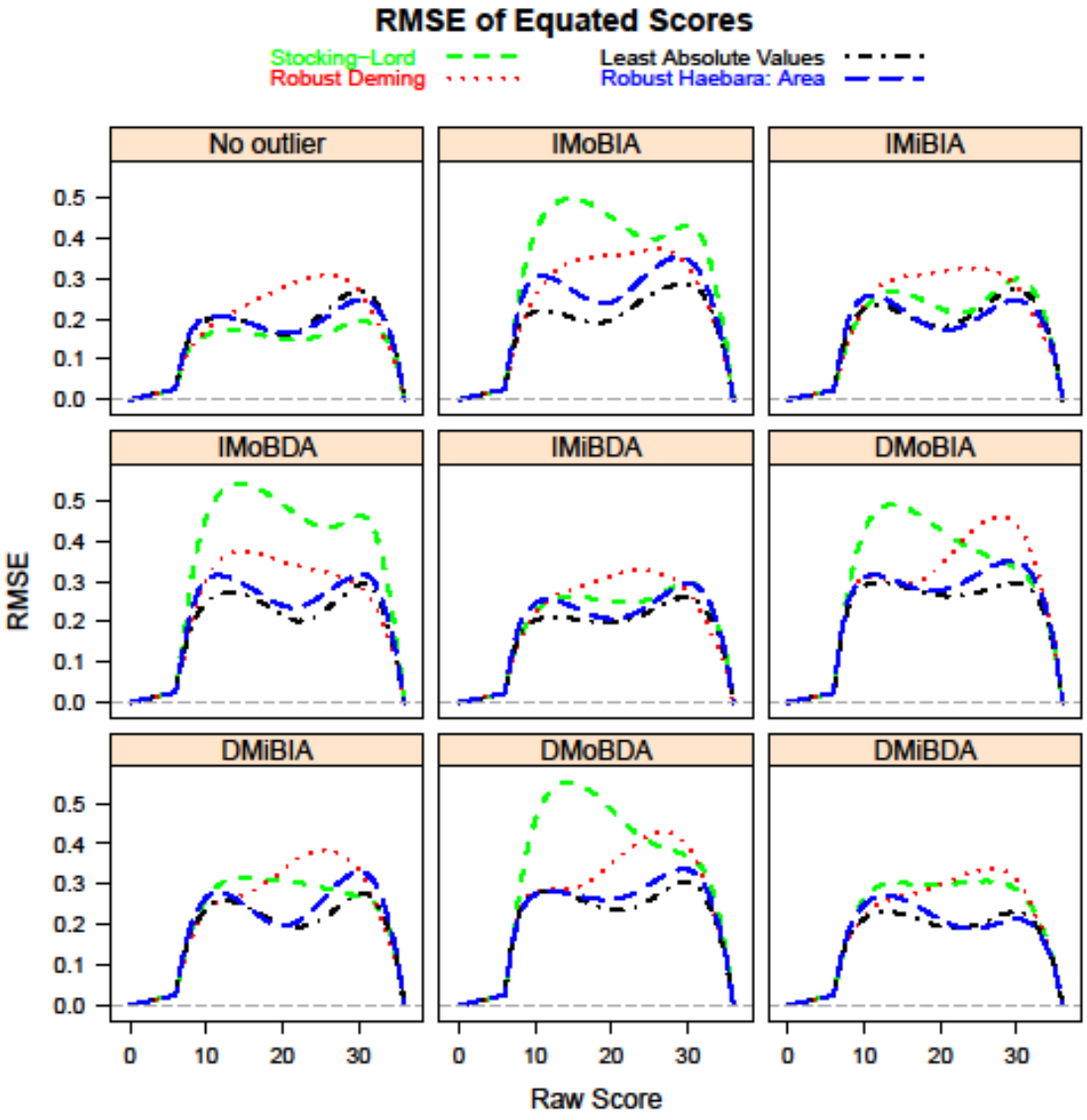


Figure 12. The RMSE statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)

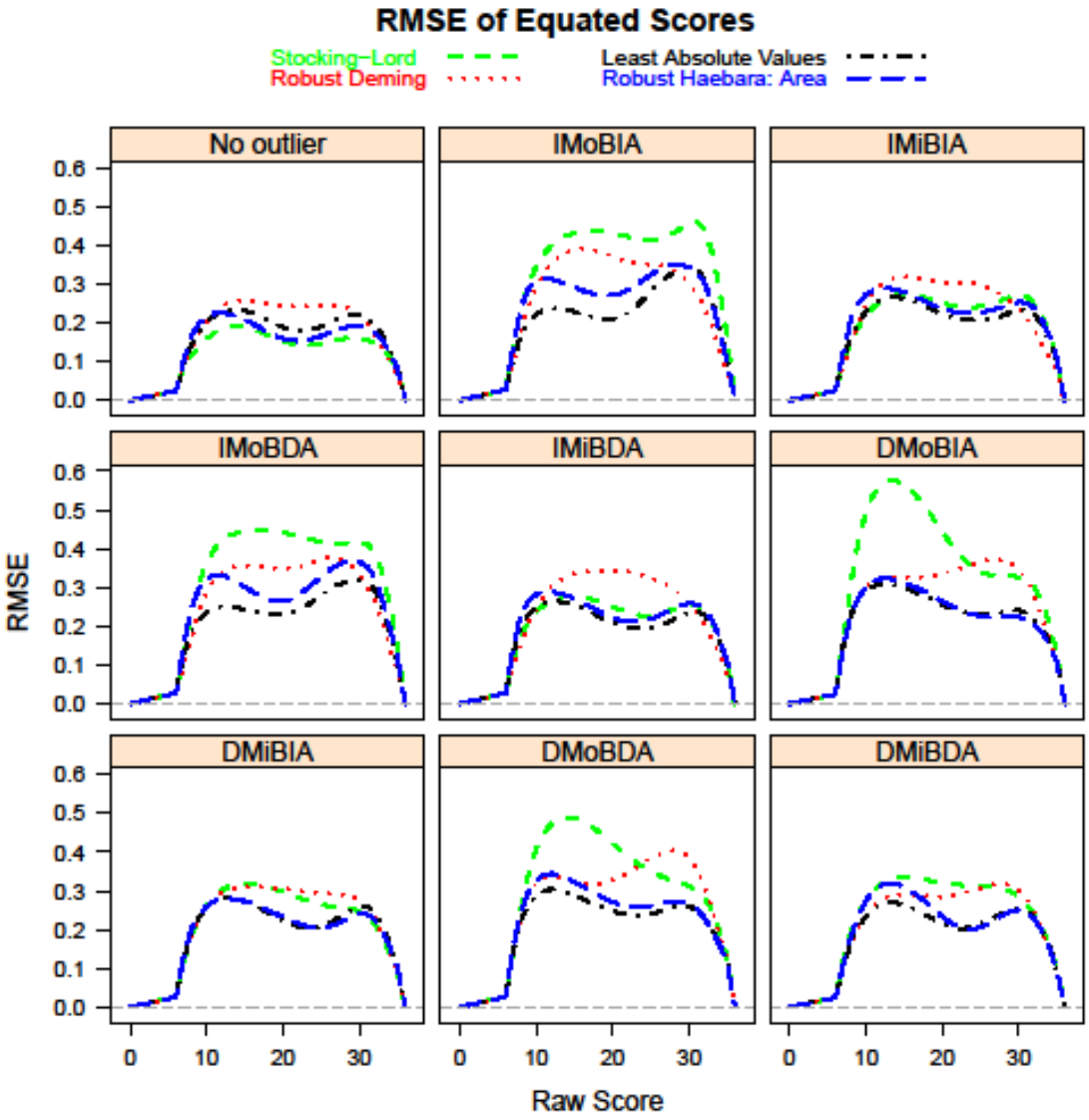


Figure 13. The RMSE statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$)

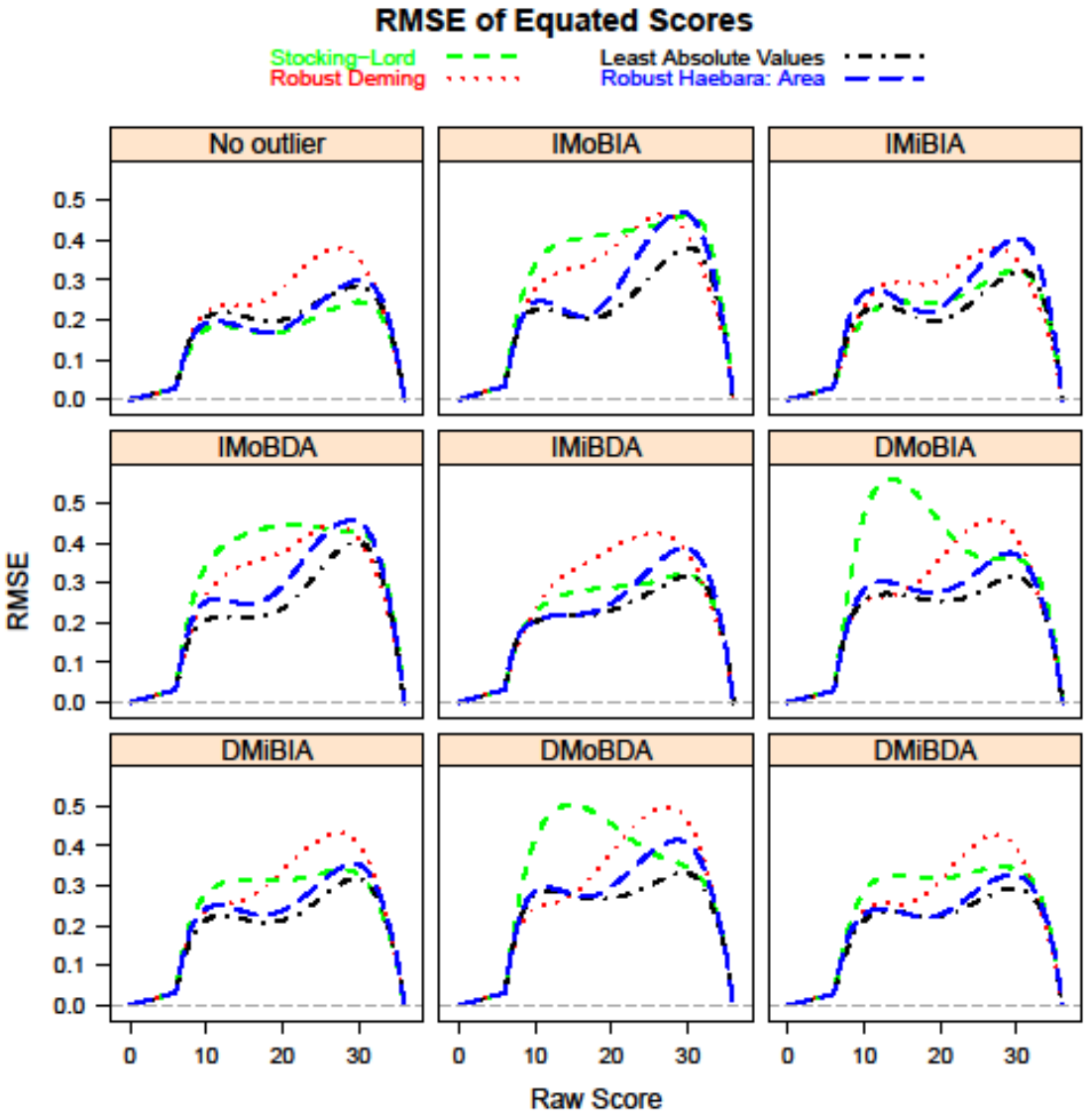
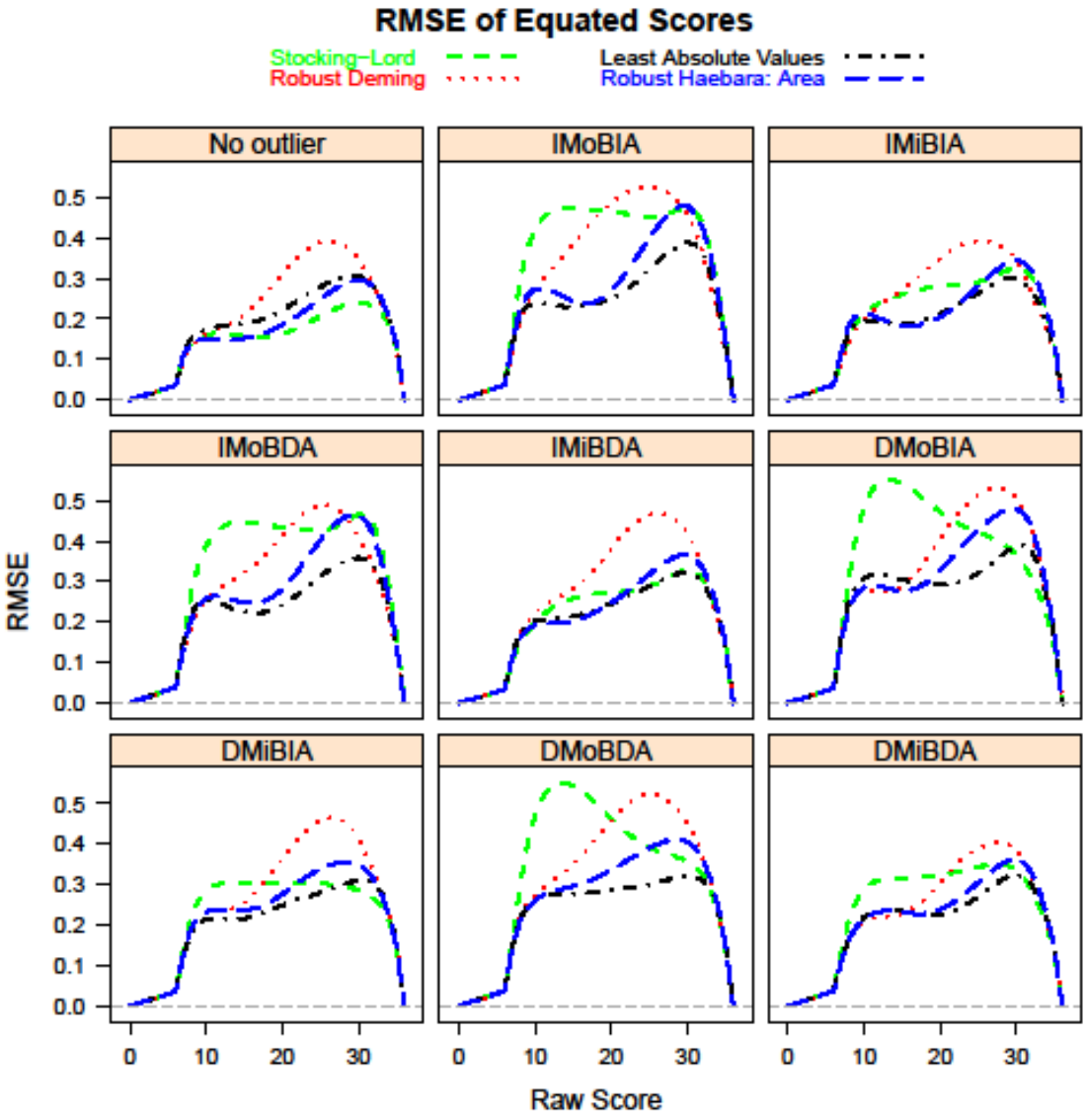


Figure 14. The RMSE statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$)



Figures 10 to 14 plots the RMSE results of equated scores under different outlier situations and methods of scale transformation and different ability distributions of examinees.

The figures indicate that when no outlier was simulated, the equating errors obtained by the proposed robust Deming method are apparently larger than the traditional

Stocking-Lord method, almost throughout the entire scale. The LAV method and the robust Haebara method generally yielded slightly larger errors than the Stocking-Lord method throughout the entire scale, except that the curves were almost identical in the middle of the scale.

When a single outlying item with mild change of b-parameter was simulated in the common item set, the equating errors produced by the Stocking-Lord method increased The robust Deming method had larger errors than the Stocking-Lord method throughout the entire scale when b-parameter increased (*IMIBIA* and *IMIBDA*), and particular for the high scores when b-parameter decreased (*DMIBIA* and *DMIBDA*). The LAV method had smaller errors than the Stocking-Lord method throughout the entire scale and the difference is more evident when b-parameter decreased (*DMIBIA* and *DMIBDA*). Equating errors obtained by the robust Haebara method were not very consistent. It produced smaller equating errors than the Stocking-Lord method when b-parameter decreased (*DMIBIA* and *DMIBDA*). However, it was larger than the Stocking-Lord method for the higher scores when b-parameter increased (*IMIBIA* and *IMIBDA*).

When the magnitude of b-parameter change was increased, the equating errors dramatically increased, particularly for the Stocking-Lord method. Generally, decreased *b*-parameter led to larger equating error than increased *b*-parameter. In addition, differences among the tested scale transformation methods are more evident. The robust Deming method had slightly smaller errors than the Stocking-Lord method almost throughout the entire scale when b-parameter increased (*IMIBIA* and *IMIBDA*). When the *b*-parameter decreased (*DMIBIA* and *DMIBDA*), the RMSE of the robust Deming method

was smaller for scores below approximately 20. The LAV method had the least equating errors among all the studied methods throughout the entire scale with a few exceptions. The robust Haebara method generally produced smaller RMSE than the Stocking-Lord method, except that they were nearly identical for higher scores.

Figure 15. The Bias statistics of IRT equating ($\theta \sim N(0,1)$)

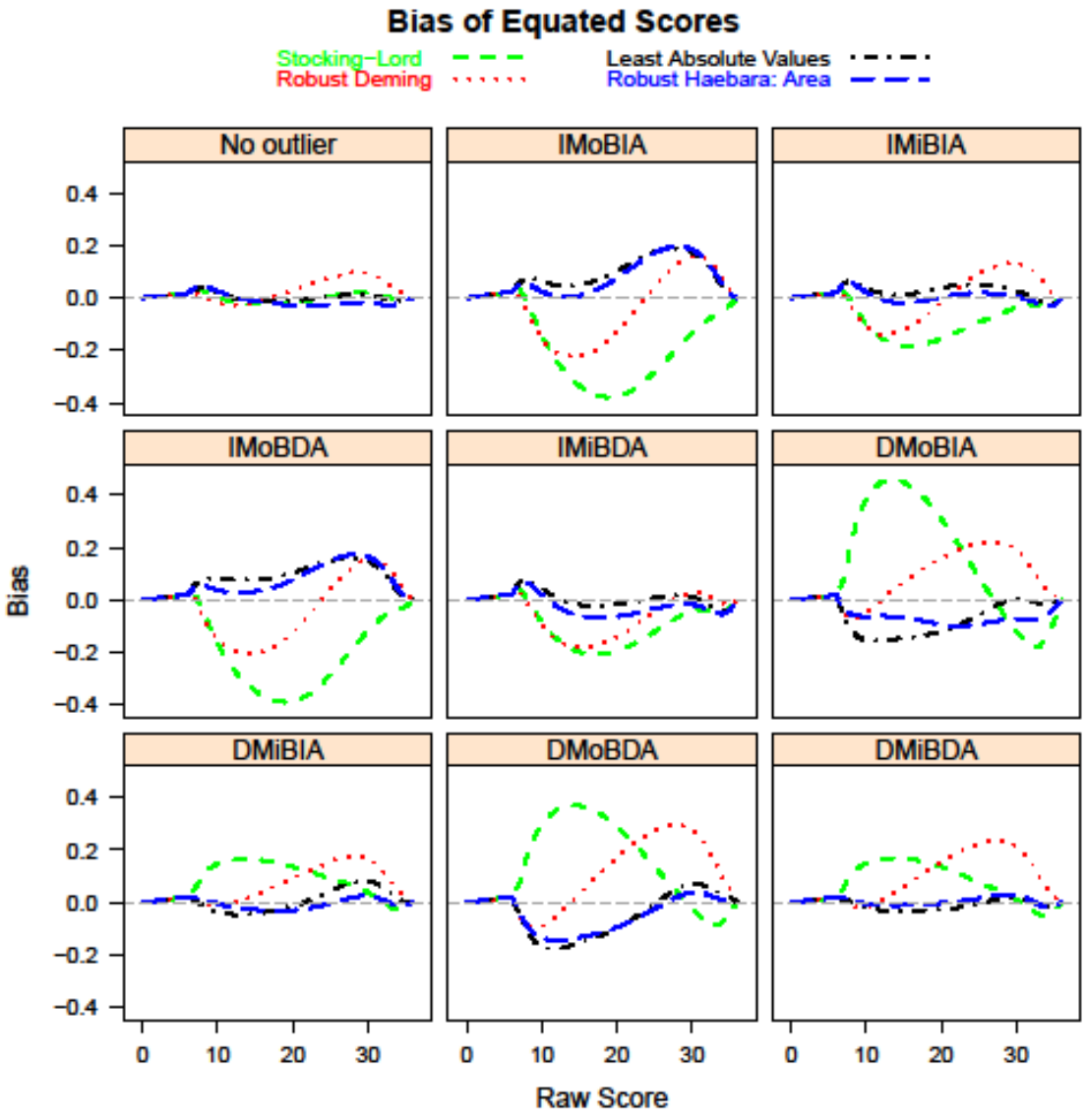


Figure 16. The Bias statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

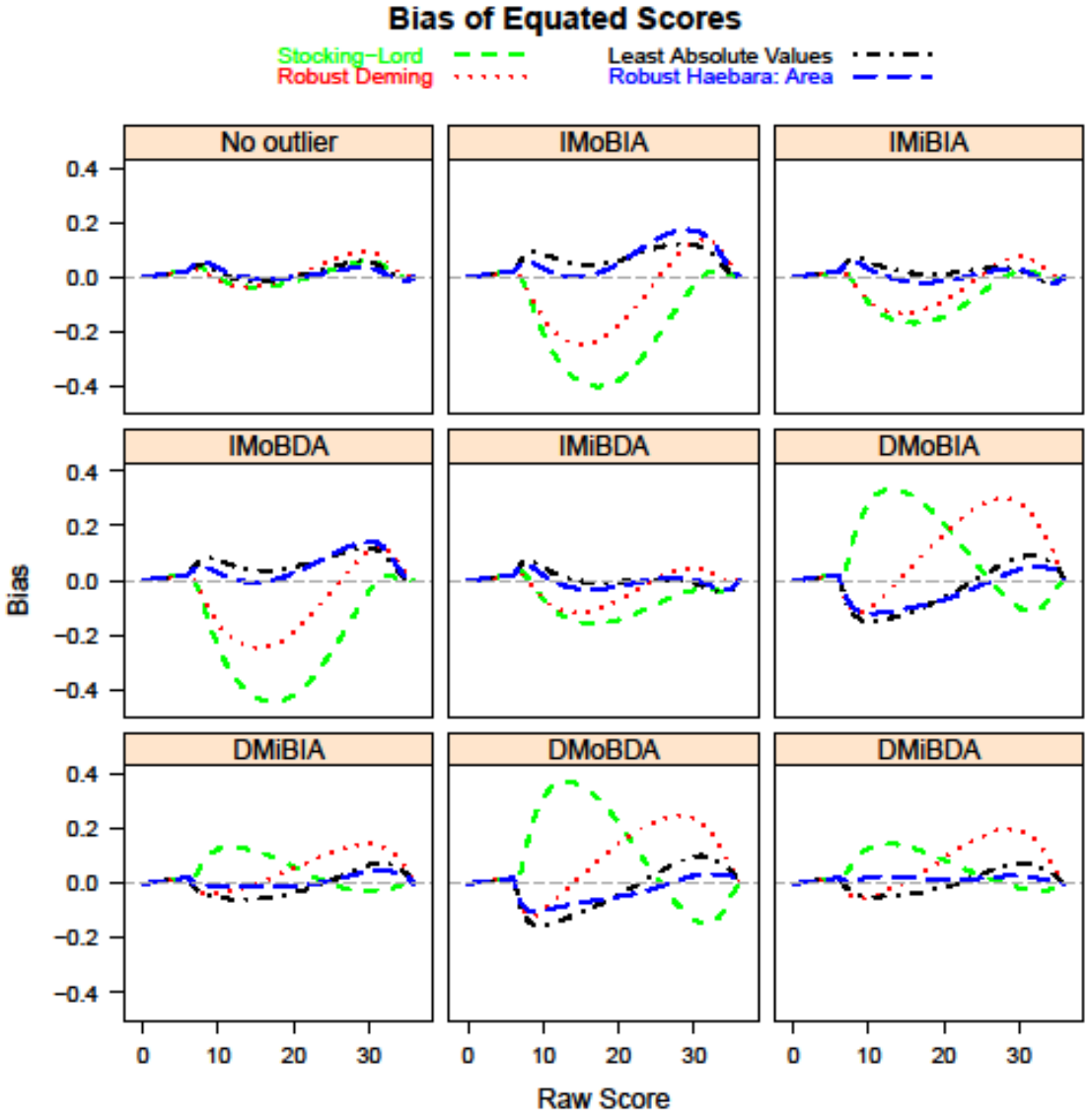


Figure 17. The Bias statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)

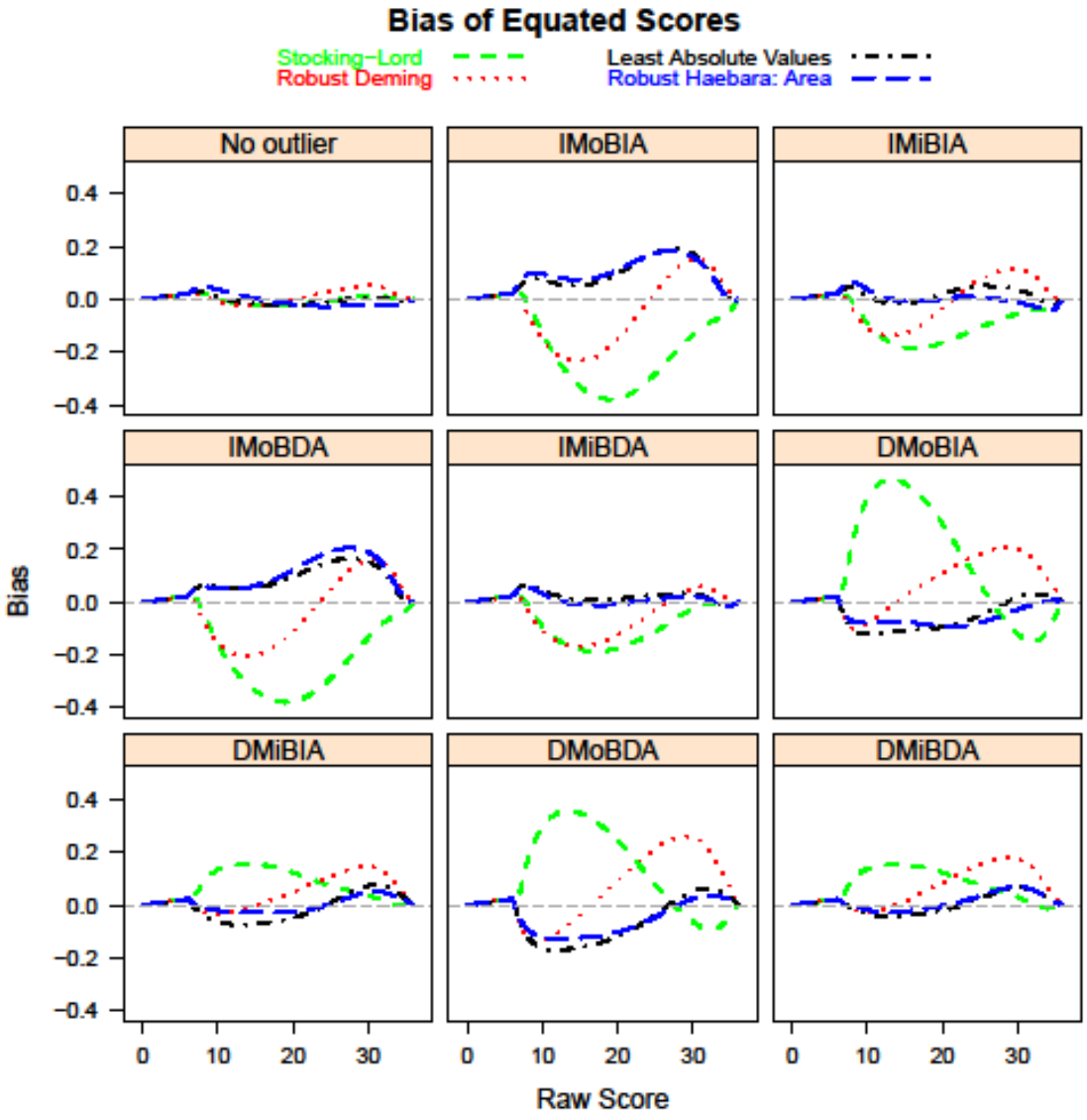


Figure 18. The Bias statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$)

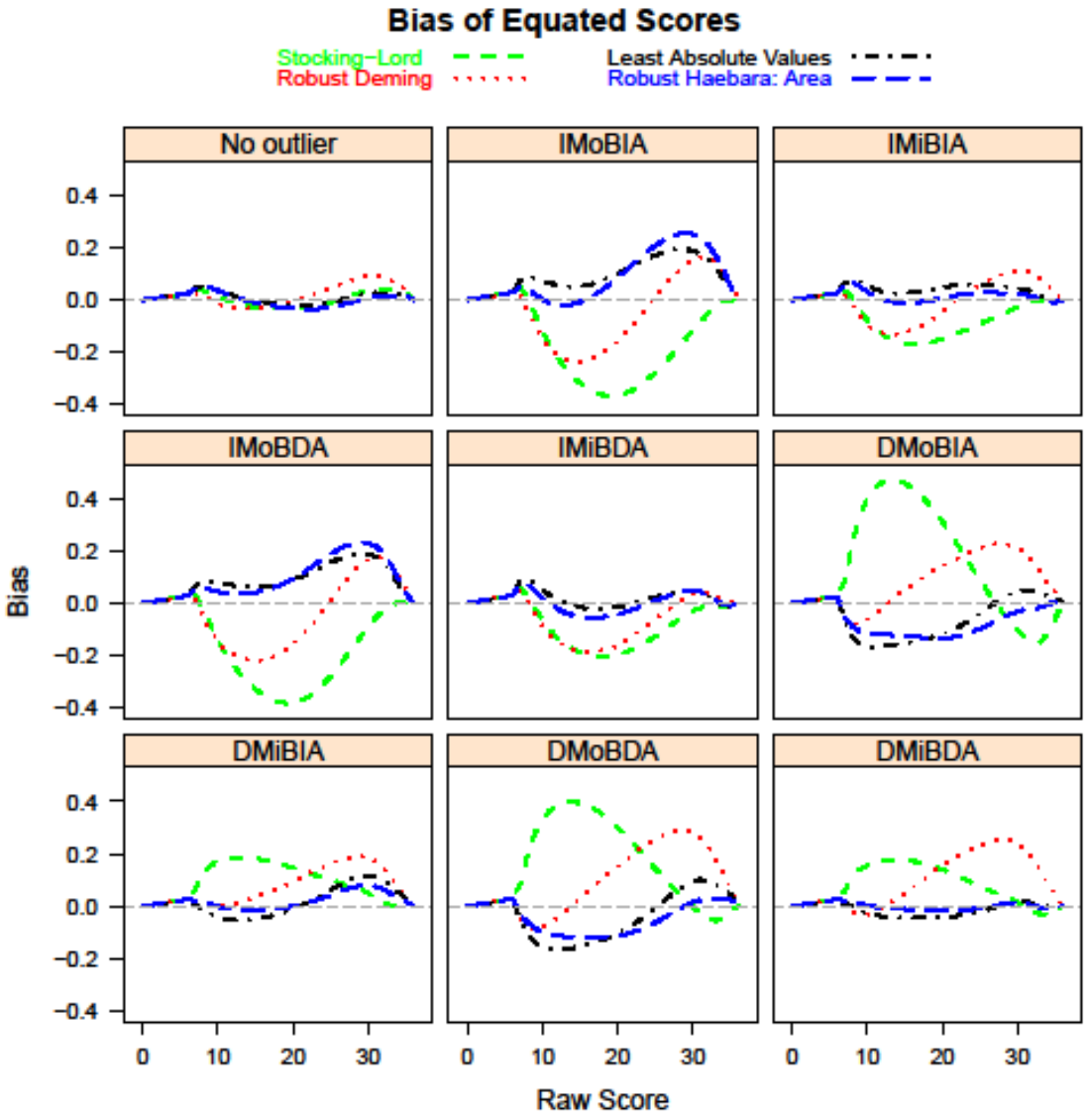
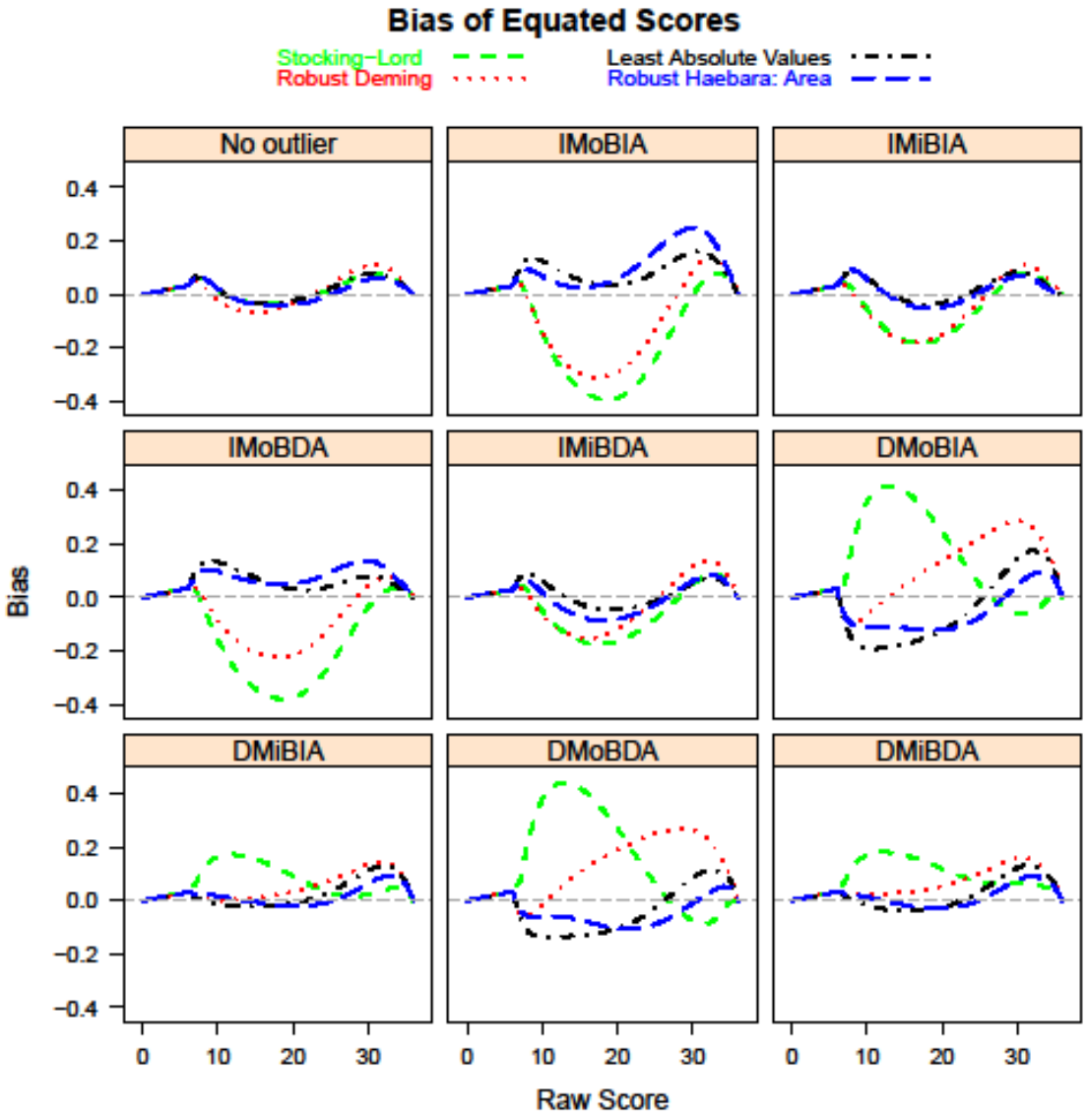


Figure 19. The Bias statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$)



Figures 15 to 19 show the bias plots for equated scores under different outlier situations and methods of the traditional scale transformation and various ability distributions of examinees.

All scale transformation methods produced approximately identical absolute biases when no outlier was simulated since the bias curves were very close to each other.

However, the bias obtained by the robust Deming method is slightly larger than the others for the scores approximately larger than 20.

When a single outlying item with mild change of *b*-parameter was simulated in the common item set, more systematic errors werer introduced to the equating procedure after transformation. Interestingly, both the LAV method and the robust Haebara method had smaller bias than the other two methods. In addition, the direction of bias caused by the LAV method is opposite to the Stocking-Lord method. The robust Deming method had smaller errors than the Stocking-Lord method throughout the entire scale when *b*-parameter increased (*IMIBIA* and *IMIBDA*). The robust Deming method did not perform very well for high scores when the b-parameter decreased (*DMIBIA* and *DMIBDA*).

When the b-parameter was moderately increased, the equating bias dramatically increased for the traditional Stocking-Lord method. Generally, decreased b-parameter led to larger equating error than increased b-parameter. Moreover, differences among the tested scale transformation methods are more apparent. The LAV method consistently had the least equating bias among all the studied methods throughout the entire scale. Generally, the LAV method and the robust Haebara method had smaller bias than the Stocking-Lord method except that the robust Deming method had larger bias for high scores when b-parameter (*DMIBIA* and *DMIBDA*).

Figures 20 to 24 show the plots of standard errors for equated scores under different outlier situations and methods of the traditional scale transformation and various ability distributions of examinees. The robust methods generally had larger standard errors than the Stocking-Lord method when no outlying common item is simulated. Particularly, the

robust Deming method had the largest random errors among the methods. When a single outlying item with mild *b*-parameter change was simulated, the robust Deming method still had the largest equating bias, especially when the scores were approximately between 12 and 30. The robust Haebara method had slightly larger equating bias than the Stocking-Lord method when *b*-parameter increased (*IMiBIA* and *IMiBDA*). It was smaller equating bias when *b*- parameter decreased. The LAV method had nearly identical equating bias to the Stocking-Lord method when *b*-parameter increased (*IMiBIA* and *IMiBDA*), but smaller bias when *b*-parameter decreased (*DMiBIA* and *DMiBDA*).

When the *b*-parameter was moderately increased, equating standard errors dramatically increased for the Stocking-Lord method. The robust Haebara method had nearly identical equating bias to the Stocking-Lord method when the *b*-parameter increased (*IMoBIA* and *IMoBDA*), but smaller bias when the *b*-parameter deceased (*DMoBIA* and *DMoBDA*). The robust Deming method generally had larger equating bias when the *b*-parameter moderately increased (*IMoBIA* and *IMoBDA*), but smaller bias when the *b*-parameter decreased (*DMoBIA* and *DMoBDA*).

Overall, the inclusion of a single simulated outlying common item enlarges the equating errors. The proposed LAV method and the robust Haebara method generally have smaller errors than the robust Deming method. In addition, reduction of errors by the robust Deming method, however, is not very consistent. Therefore, in the following study, the robust Deming method is not used.

Figure 20. The Standard Error statistics of IRT equating ($\theta \sim N(0,1)$)

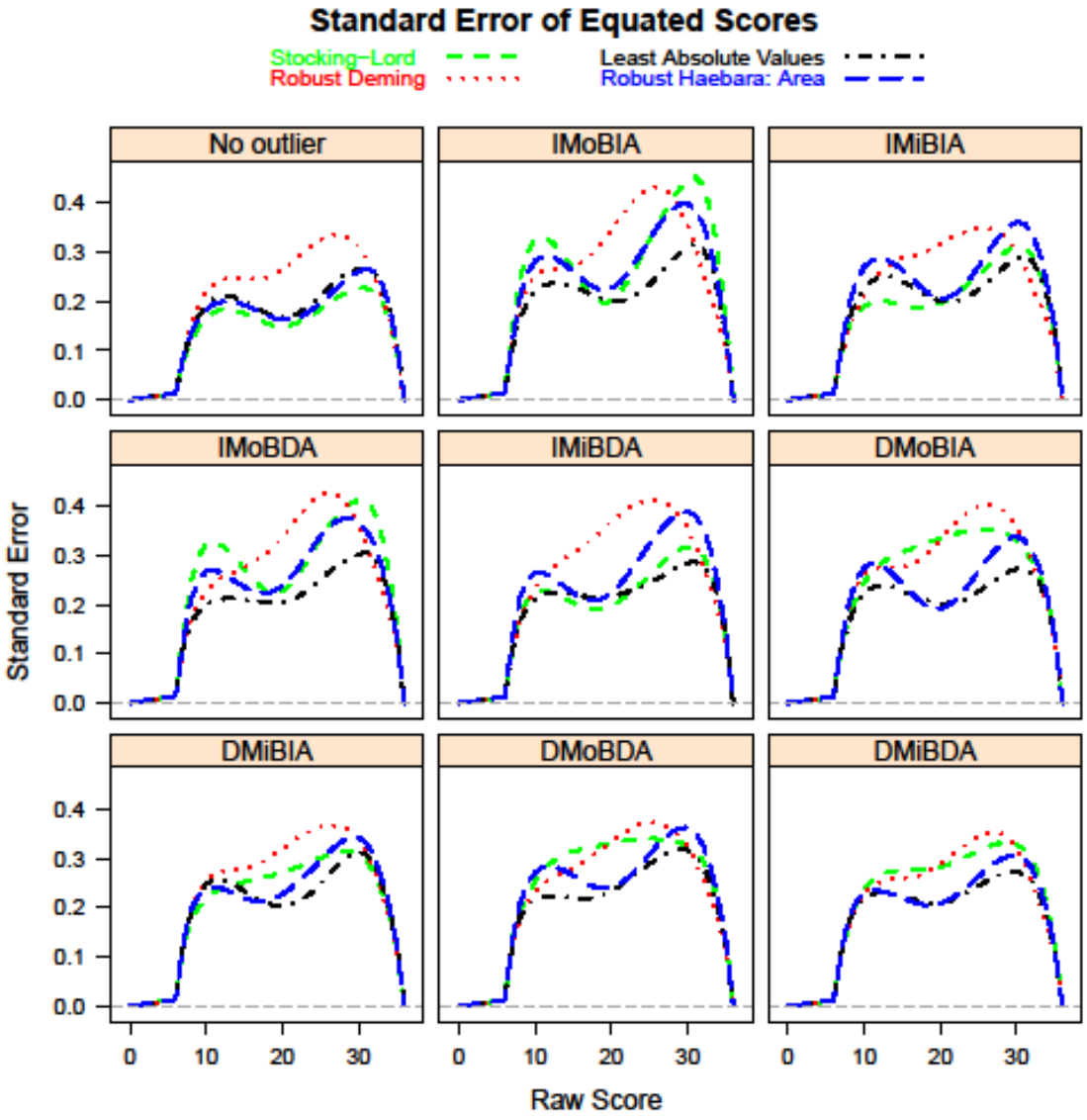


Figure 21. The Standard Error statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

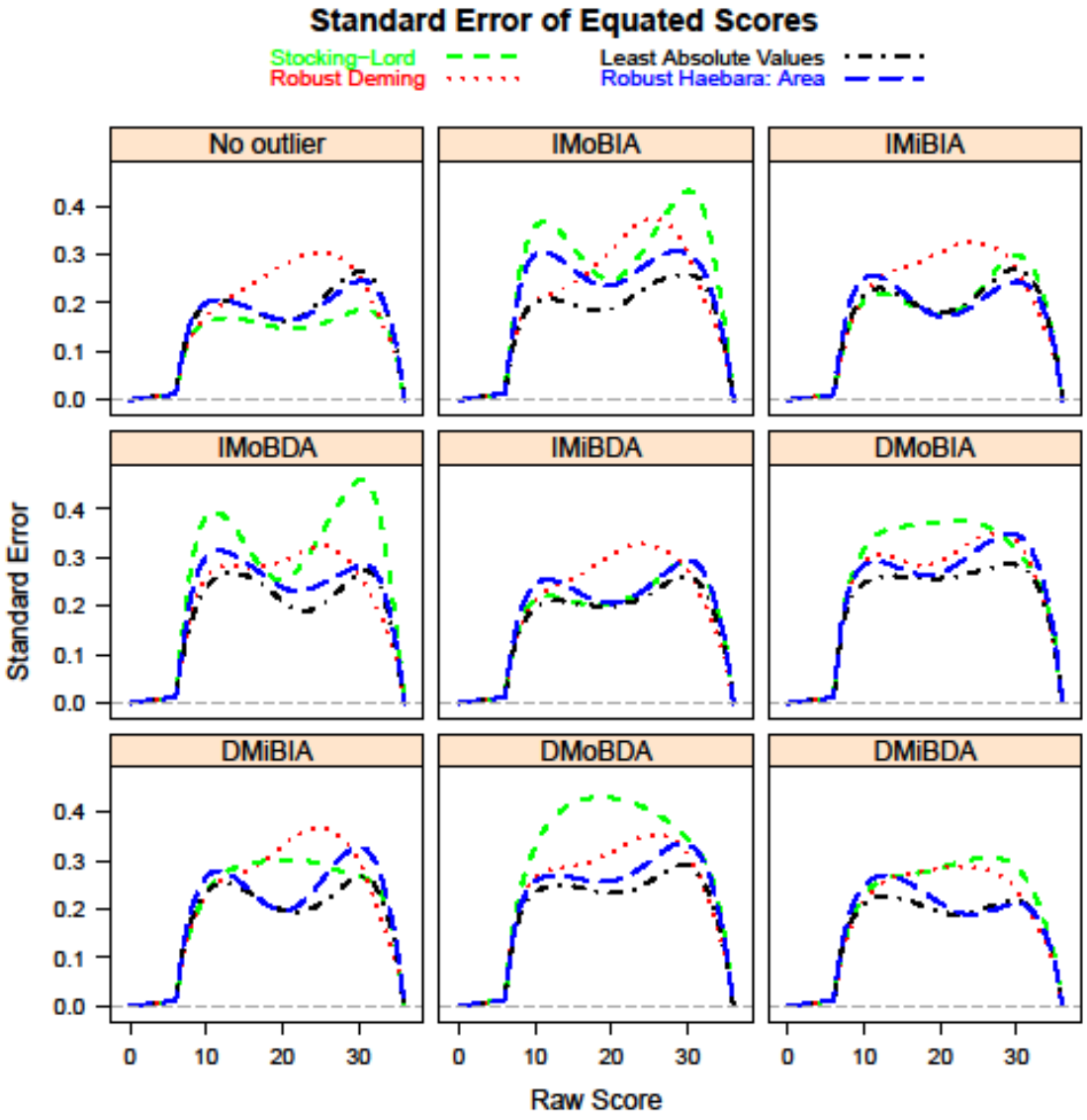


Figure 22. The Standard Error statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)

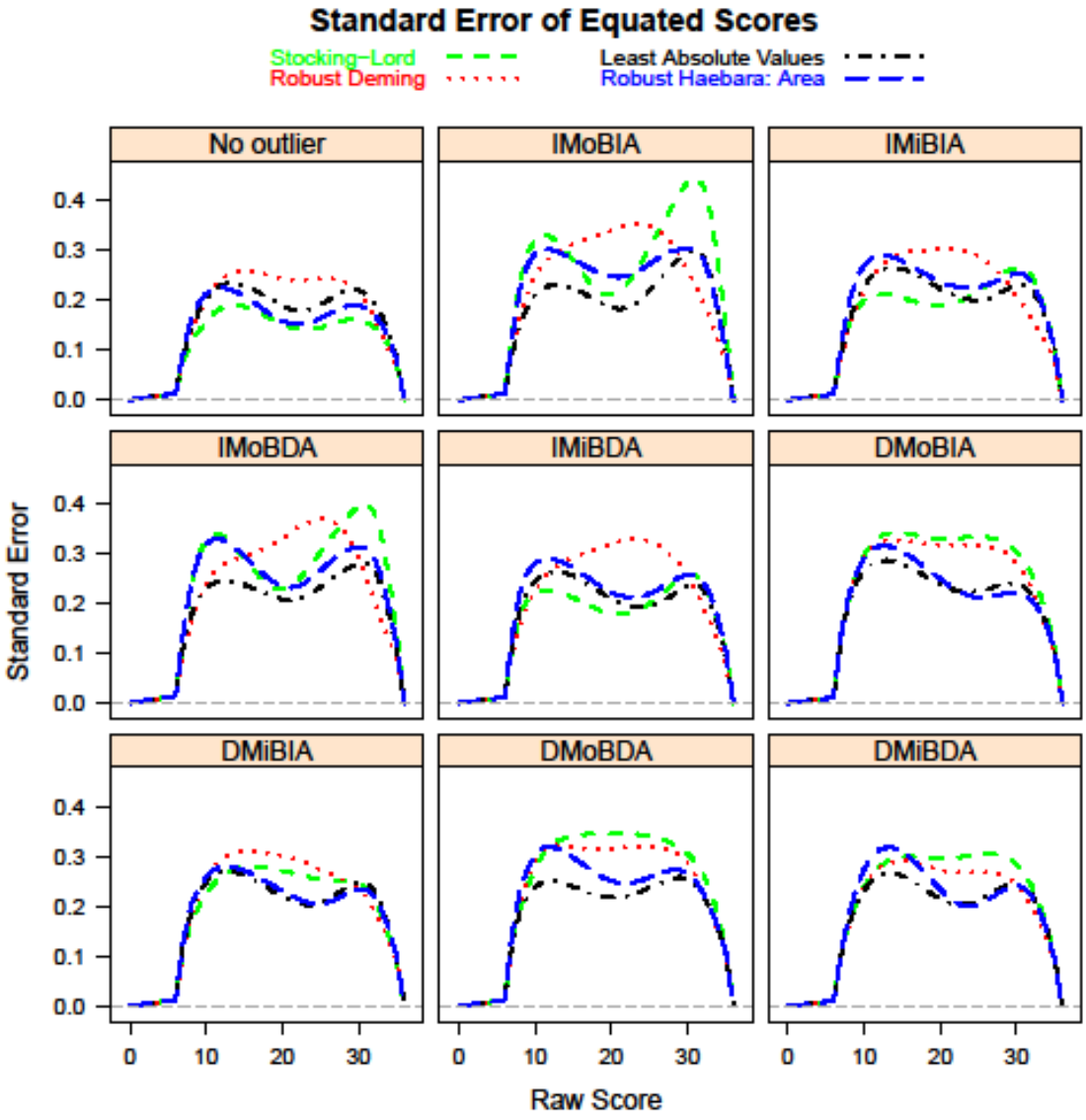


Figure 23. The Standard Error statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$)

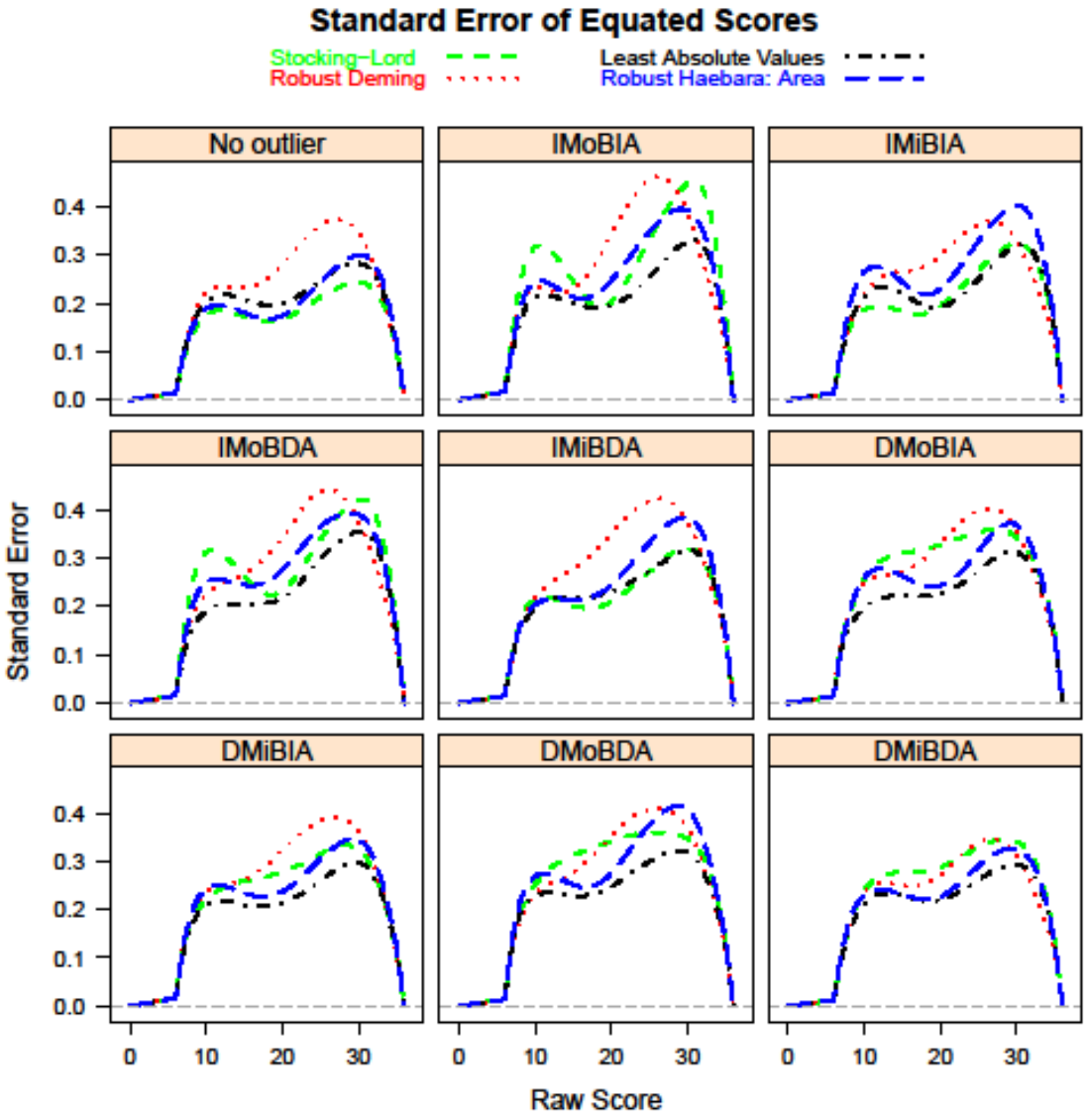
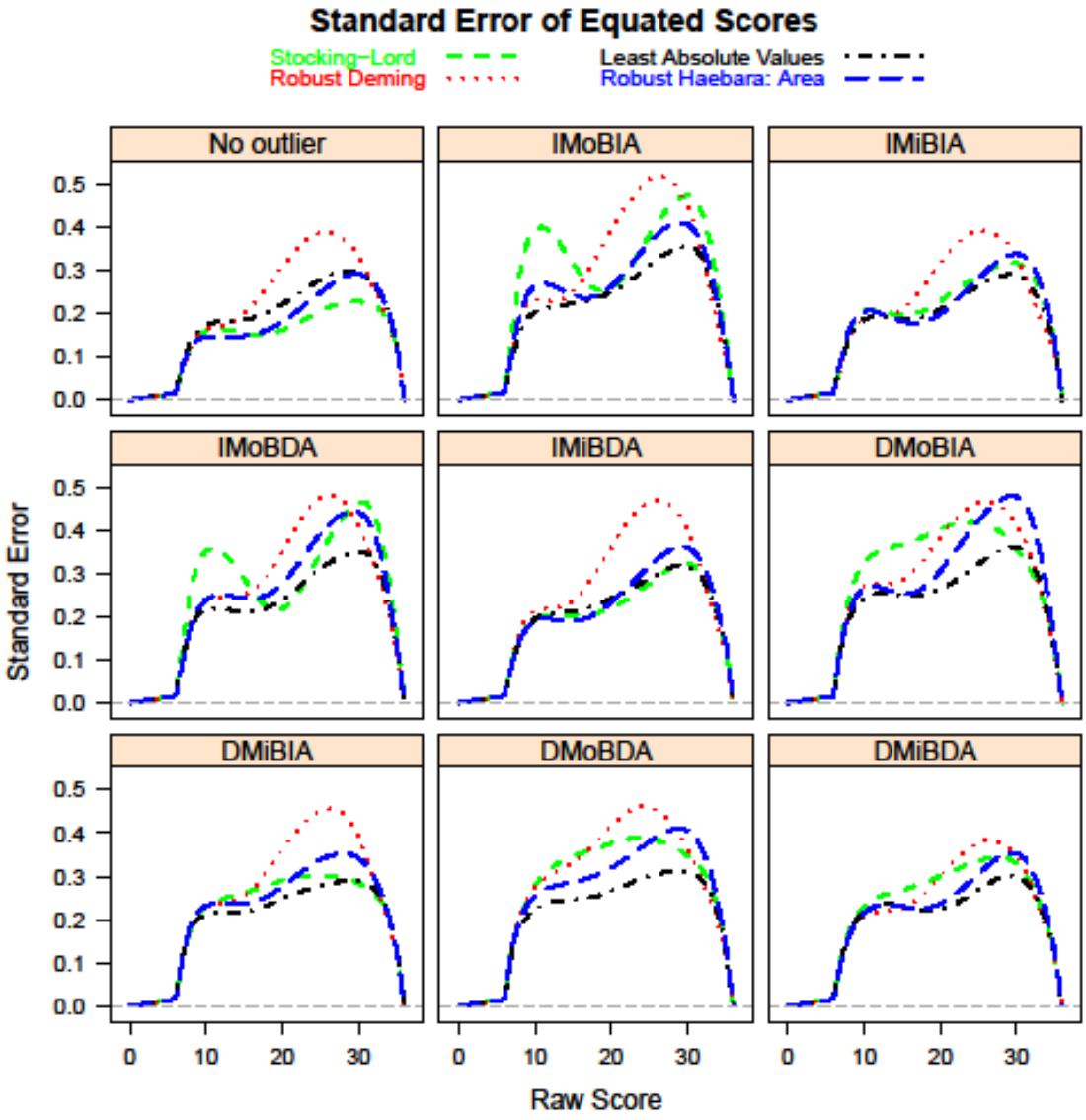


Figure 24. The Standard Error statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$)



4.1.3 Comparisons between Robust Methods of Scale Transformation and Outlier Removal

In this section, performances of scale transformation and equating are compared between the robust methods, LAV and robust Haebara, and traditional outlier elimination. Again, the Stocking-Lord method without outlier identification and elimination are

treated as the *Baseline*. Two outlier elimination methods are used: the displacement method and the direct elimination after simulation. To further investigate the performance of the robust methods, a single outlying common item under more severe conditions ($1 < |\Delta b| < 1.5$) was simulated in the common item set. The results for the coefficients of scale transformation are presented first, followed by the results for the equating errors including both weighted indices and individual indices.

Coefficients of Scale Transformation

The coefficients of scale transformation were estimated by the proposed robust methods, the outlier removal methods, and the *Baseline* with the Stocking-Lord method without outlier detection and elimination (Tables 14 to 18) for various ability distributions. The mean and standard deviation in the table are obtained from 100 sets of the estimates of the scale transformation coefficients.

In the simulated study, the population values were assumed to be known. The observed values of item responses were simulated based on the 3PL IRT model with various ability distributions: $\theta \sim N(0, 1)$, $\theta \sim N(0.25, 1.1^2)$, $\theta \sim N(0.5, 1.2^2)$, $\theta \sim N(-0.25, 1.1^2)$, $\theta \sim N(-0.5, 1.2^2)$. If estimation is accurate, the estimated A coefficients should be the value of standard deviation of a given ability distribution and B the value of mean of that ability distribution. If a method has a larger discrepancy between the estimated coefficients and the true ones, it produced larger bias.

Table 14. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0,1)$

Outlier		Baseline		Robust Haebara		LAV		Displacement		Exclusion	
		A	B	A	B	A	B	A	B	A	B
1) No outlier	*M*	1.016	0.001	1.006	0.005	1.012	0.004	1.013	0.003	1.016	0.001
	SD	0.031	0.040	0.035	0.040	0.033	0.040	0.033	0.040	0.031	0.040
2) IMoBIA	*M*	1.029	-0.083	1.024	-0.014	1.012	-0.013	1.017	-0.012	1.016	-0.008
	SD	0.084	0.055	0.049	0.046	0.039	0.043	0.040	0.047	0.035	0.042
3) ISBIA	*M*	1.029	-0.114	1.022	-0.005	1.014	-0.005	1.020	-0.001	1.022	-0.001
	SD	0.102	0.081	0.056	0.049	0.041	0.047	0.036	0.046	0.035	0.046
4) DMoBIA	*M*	1.032	-0.083	1.021	-0.019	1.014	-0.013	1.025	-0.015	1.021	-0.010
	SD	0.076	0.063	0.058	0.050	0.039	0.046	0.041	0.048	0.036	0.041
5) DSBIA	*M*	1.025	-0.125	1.023	-0.013	1.014	-0.015	1.019	-0.009	1.018	-0.008
	SD	0.097	0.079	0.046	0.042	0.039	0.040	0.034	0.039	0.034	0.038
6) IMoBDA	*M*	0.962	0.052	1.008	0.000	1.015	-0.001	1.008	-0.008	1.022	-0.010
	SD	0.045	0.073	0.044	0.041	0.039	0.041	0.044	0.044	0.036	0.040
7) ISBDA	*M*	0.951	0.105	0.999	0.014	1.007	0.014	1.016	0.003	1.019	0.002
	SD	0.065	0.066	0.053	0.049	0.044	0.044	0.038	0.044	0.037	0.044
8) DMoBDA	*M*	0.961	0.061	1.002	0.010	1.008	0.013	1.007	0.003	1.021	0.001
	SD	0.051	0.065	0.047	0.040	0.042	0.042	0.042	0.044	0.038	0.041
9) DSBDA	*M*	0.955	0.112	1.008	0.015	1.014	0.013	1.018	0.006	1.024	0.004
	SD	0.067	0.052	0.057	0.041	0.048	0.043	0.042	0.038	0.041	0.038

Table 15. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0.25,1.1^2)$

Outlier		Baseline		Robust Haebara		LAV		Displacement		Exclusion	
		A	B	A	B	A	B	A	B	A	B
1) No outlier	*M*	1.109	0.249	1.103	0.252	1.106	0.253	1.108	0.250	1.109	0.249
	SD	0.033	0.040	0.038	0.040	0.041	0.041	0.034	0.041	0.033	0.040
2) IMoBIA	*M*	1.129	0.185	1.120	0.239	1.112	0.240	1.118	0.238	1.115	0.244
	SD	0.082	0.050	0.057	0.047	0.048	0.045	0.050	0.049	0.043	0.043
3) ISBIA	*M*	1.132	0.145	1.114	0.241	1.106	0.241	1.113	0.244	1.111	0.246
	SD	0.118	0.071	0.053	0.049	0.042	0.048	0.038	0.046	0.038	0.044
4) DMoBIA	*M*	1.132	0.183	1.120	0.236	1.114	0.238	1.116	0.237	1.116	0.243
	SD	0.079	0.054	0.060	0.045	0.047	0.041	0.047	0.044	0.041	0.042
5) DSBIA	*M*	1.130	0.146	1.113	0.236	1.109	0.234	1.113	0.243	1.113	0.244
	SD	0.108	0.063	0.058	0.047	0.047	0.046	0.041	0.044	0.041	0.044
6) IMoBDA	*M*	1.056	0.290	1.094	0.255	1.105	0.250	1.104	0.243	1.114	0.246
	SD	0.059	0.077	0.055	0.045	0.047	0.044	0.053	0.048	0.046	0.044
7) ISBDA	*M*	1.052	0.332	1.102	0.252	1.104	0.253	1.108	0.249	1.114	0.247
	SD	0.070	0.054	0.048	0.048	0.039	0.045	0.037	0.042	0.037	0.043
8) DMoBDA	*M*	1.051	0.295	1.091	0.256	1.097	0.255	1.097	0.244	1.109	0.249
	SD	0.052	0.072	0.049	0.045	0.039	0.045	0.044	0.050	0.036	0.046
9) DSBDA	*M*	1.048	0.334	1.096	0.255	1.104	0.256	1.111	0.250	1.115	0.248
	SD	0.069	0.055	0.054	0.048	0.044	0.045	0.040	0.045	0.041	0.045

Table 16. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(0.5,1.2^2)$

Outlier		Baseline		Robust Haebara		LAV		Displacement		Exclusion	
		A	B	A	B	A	B	A	B	A	B
1) No outlier	*M*	1.209	0.507	1.202	0.508	1.206	0.509	1.209	0.507	1.209	0.507
	SD	0.043	0.046	0.049	0.048	0.047	0.050	0.044	0.046	0.043	0.046
2) IMoBIA	*M*	1.234	0.450	1.212	0.501	1.208	0.500	1.218	0.499	1.214	0.506
	SD	0.095	0.060	0.058	0.051	0.056	0.047	0.054	0.050	0.050	0.045
3) ISBIA	*M*	1.225	0.398	1.202	0.495	1.201	0.492	1.209	0.497	1.207	0.498
	SD	0.123	0.064	0.058	0.048	0.046	0.050	0.043	0.047	0.042	0.048
4) DMoBIA	*M*	1.225	0.444	1.211	0.492	1.205	0.497	1.211	0.499	1.208	0.502
	SD	0.091	0.056	0.057	0.046	0.051	0.042	0.047	0.044	0.042	0.040
5) DSBIA	*M*	1.236	0.396	1.210	0.491	1.211	0.493	1.218	0.497	1.216	0.498
	SD	0.126	0.073	0.059	0.055	0.047	0.050	0.045	0.050	0.043	0.050
6) IMoBDA	*M*	1.143	0.535	1.199	0.502	1.207	0.504	1.204	0.496	1.214	0.500
	SD	0.054	0.077	0.051	0.049	0.046	0.050	0.048	0.053	0.044	0.048
7) ISBDA	*M*	1.136	0.578	1.200	0.506	1.204	0.507	1.214	0.501	1.215	0.500
	SD	0.075	0.056	0.055	0.046	0.044	0.043	0.038	0.045	0.037	0.045
8) DMoBDA	*M*	1.140	0.535	1.188	0.507	1.201	0.508	1.198	0.502	1.207	0.503
	SD	0.062	0.077	0.060	0.047	0.051	0.050	0.049	0.051	0.042	0.049
9) DSBDA	*M*	1.136	0.580	1.194	0.509	1.206	0.512	1.215	0.507	1.217	0.506
	SD	0.074	0.058	0.057	0.050	0.042	0.049	0.040	0.048	0.039	0.047

Table 17. Mean (M) and Standard Deviation (SD) of Scale Transformation Coefficients, 100 replications, $\theta \sim N(-0.25,1.1^2)$

Outlier		Baseline		Robust Haebara		LAV		Displacement		Exclusion	
		A	B	A	B	A	B	A	B	A	B
1) No outlier	*M*	1.114	-0.259	1.110	-0.256	1.113	-0.256	1.113	-0.258	1.114	-0.259
	SD	0.041	0.041	0.041	0.041	0.043	0.043	0.041	0.040	0.041	0.041
2) IMoBIA	*M*	1.118	-0.325	1.124	-0.272	1.105	-0.259	1.112	-0.262	1.112	-0.257
	SD	0.075	0.070	0.060	0.062	0.044	0.047	0.043	0.052	0.039	0.046
3) ISBIA	*M*	1.114	-0.355	1.121	-0.261	1.107	-0.257	1.112	-0.251	1.113	-0.251
	SD	0.105	0.097	0.065	0.057	0.050	0.047	0.044	0.045	0.043	0.046
4) DMoBIA	*M*	1.119	-0.325	1.118	-0.267	1.107	-0.258	1.112	-0.259	1.114	-0.255
	SD	0.069	0.071	0.048	0.052	0.038	0.048	0.035	0.053	0.036	0.048
5) DSBIA	*M*	1.110	-0.359	1.113	-0.258	1.103	-0.256	1.108	-0.251	1.110	-0.251
	SD	0.100	0.108	0.050	0.054	0.038	0.049	0.036	0.047	0.036	0.046
6) IMoBDA	*M*	1.039	-0.157	1.084	-0.231	1.094	-0.236	1.088	-0.235	1.100	-0.244
	SD	0.054	0.073	0.052	0.048	0.046	0.045	0.048	0.051	0.039	0.044
7) ISBDA	*M*	1.044	-0.124	1.088	-0.236	1.101	-0.236	1.107	-0.245	1.111	-0.247
	SD	0.073	0.062	0.047	0.046	0.043	0.043	0.043	0.043	0.041	0.043
8) DMoBDA	*M*	1.048	-0.165	1.098	-0.244	1.104	-0.244	1.102	-0.250	1.114	-0.254
	SD	0.060	0.076	0.052	0.050	0.042	0.050	0.047	0.051	0.041	0.050
9) DSBDA	*M*	1.037	-0.131	1.092	-0.239	1.098	-0.240	1.103	-0.251	1.107	-0.255
	SD	0.066	0.070	0.054	0.048	0.044	0.046	0.042	0.048	0.041	0.048

Table 18. Mean (M) and Standard Deviation (SD) of the Scale Transformation Coefficients, 100 replications, $\theta \sim N(-0.5,1.2^2)$

Outlier		Baseline		Robust Haebara		LAV		Displacement		Exclusion	
		A	B	A	B	A	B	A	B	A	B
1) No outlier	*M*	1.195	-0.488	1.201	-0.490	1.197	-0.490	1.194	-0.487	1.195	-0.488
	SD	0.051	0.051	0.051	0.049	0.054	0.051	0.053	0.052	0.051	0.051
2) IMoBIA	*M*	1.229	-0.581	1.219	-0.517	1.204	-0.510	1.213	-0.511	1.209	-0.505
	SD	0.091	0.090	0.056	0.065	0.051	0.062	0.049	0.064	0.045	0.056
3) ISBIA	*M*	1.215	-0.608	1.208	-0.508	1.194	-0.502	1.200	-0.498	1.200	-0.498
	SD	0.112	0.115	0.064	0.056	0.053	0.053	0.050	0.052	0.048	0.052
4) DMoBIA	*M*	1.219	-0.570	1.218	-0.509	1.205	-0.502	1.202	-0.498	1.202	-0.495
	SD	0.089	0.086	0.065	0.056	0.050	0.051	0.050	0.047	0.045	0.043
5) DSBIA	*M*	1.215	-0.605	1.211	-0.501	1.195	-0.492	1.199	-0.489	1.199	-0.490
	SD	0.111	0.119	0.069	0.062	0.048	0.052	0.048	0.055	0.046	0.054
6) IMoBDA	*M*	1.134	-0.407	1.182	-0.485	1.195	-0.491	1.185	-0.493	1.199	-0.498
	SD	0.064	0.081	0.053	0.054	0.052	0.053	0.056	0.055	0.046	0.051
7) ISBDA	*M*	1.143	-0.364	1.184	-0.477	1.196	-0.478	1.199	-0.489	1.202	-0.490
	SD	0.090	0.084	0.069	0.052	0.059	0.046	0.056	0.046	0.054	0.046
8) DMoBDA	*M*	1.136	-0.397	1.195	-0.492	1.199	-0.491	1.193	-0.493	1.204	-0.496
	SD	0.059	0.073	0.050	0.049	0.046	0.045	0.054	0.045	0.046	0.041
9) DSBDA	*M*	1.142	-0.369	1.183	-0.481	1.195	-0.481	1.196	-0.490	1.200	-0.492
	SD	0.075	0.095	0.055	0.057	0.046	0.059	0.042	0.057	0.041	0.058

The tables indicate that the estimated coefficients of scale transformation by the studied methods are very close to each other and slightly biased under the condition of *No outlier*. The robust methods had slightly larger standard deviations of the coefficients. When a single outlier was simulated, the scale transformation coefficients, particularly *B*, obtained by the *Baseline* method dramatically deviated from the population values. In addition, the *Baseline* method usually had larger standard deviations when a single outlier was included in the scale transformation. The scale transformation coefficients obtained by the robust methods were similar to the outlier removal methods. The patterns of bias in scale transformation coefficients are similar among the different ability distributions, and the biases seem to be smaller when the examinee abilities. The displacement method had larger bias than the robust methods, but the deviation was smaller when the *a*-parameter decreased. It is noteworthy that even the direct elimination cannot exactly reproduce the population values of the scale transformation coefficients.

Weighted Indices of Equating Errors

Tables 19 through 21 summarize the weighted root mean square error, weighted absolute bias, and weighted standard error of equating. Table 19 indicates that the robust methods produced slightly larger overall weighted RMSE under the condition of *No outlier*. In addition, the displacement method also had slightly larger weighted RMSE than the *Baseline* method and the exclusion method. When a single outlier was simulated, the *Baseline* method constantly had the largest weighted RMSE. The LAV method usually produced the least overall weighted RMSE among the investigated methods, regardless of the direction of *a*- and *b*- parameter changes and the magnitude of the *b*-

parameter change. The outlier removal methods produced substantive smaller weighted RMSE than the Stocking-Lord method, as the robust Haebara method. Interestingly, the weighted RMSEs obtained by the displacement method were somewhat smaller than the exclusion method when moderate *b*-parameter changes were simulated. While the weighted RMSEs obtained by the displacement method were slightly larger than the exclusion method when severe *b*-parameter changes were simulated.

Table 19a. Weighted RMSE Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Baseline	Robust Haebara	LAV	Displacement	Exclusion
N(0,1)	*1) No outlier*	0.176	0.190	0.189	0.179	0.176
	2) IMoBIA	0.446	0.300	0.252	0.268	0.259
	3) ISBIA	0.615	0.379	0.332	0.349	0.361
	4) DMoBIA	0.435	0.306	0.256	0.278	0.271
	5) DSBIA	0.627	0.348	0.305	0.334	0.343
	6) IMoBDA	0.423	0.276	0.260	0.289	0.276
	7) ISBDA	0.530	0.314	0.274	0.294	0.301
	8) DMoBDA	0.420	0.272	0.252	0.276	0.269
	9) DSBDA	0.538	0.323	0.305	0.294	0.307
$N(0.25,1.1^2)$	*1) No outlier*	0.156	0.180	0.199	0.156	0.156
	2) IMoBIA	0.396	0.289	0.247	0.283	0.263
	3) ISBIA	0.610	0.342	0.314	0.302	0.310
	4) DMoBIA	0.403	0.287	0.237	0.277	0.247
	5) DSBIA	0.577	0.319	0.278	0.301	0.306
	6) IMoBDA	0.448	0.274	0.253	0.272	0.252
	7) ISBDA	0.480	0.316	0.289	0.298	0.312
	8) DMoBDA	0.445	0.264	0.232	0.274	0.247
	9) DSBDA	0.509	0.310	0.274	0.291	0.303
$N(0.5,1.2^2)$	*1) No outlier*	0.177	0.209	0.209	0.180	0.177
	2) IMoBIA	0.397	0.297	0.249	0.271	0.261
	3) ISBIA	0.617	0.346	0.308	0.312	0.322
	4) DMoBIA	0.415	0.273	0.241	0.254	0.251
	5) DSBIA	0.619	0.359	0.333	0.326	0.335
	6) IMoBDA	0.438	0.298	0.290	0.291	0.291
	7) ISBDA	0.537	0.322	0.289	0.310	0.315
	8) DMoBDA	0.424	0.312	0.289	0.296	0.291
	9) DSBDA	0.528	0.309	0.279	0.298	0.304

Table 19b. Weighted RMSE Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Baseline	Robust Haebara	LAV	Displacement	Exclusion
N(-0.25,1.1^2)	*1) No outlier*	0.187	0.198	0.215	0.192	0.187
	2) IMoBIA	0.420	0.299	0.259	0.267	0.266
	3) ISBIA	0.572	0.373	0.310	0.318	0.335
	4) DMoBIA	0.392	0.264	0.261	0.286	0.280
	5) DSBIA	0.617	0.361	0.317	0.331	0.345
	6) IMoBDA	0.479	0.282	0.267	0.284	0.268
	7) ISBDA	0.567	0.305	0.271	0.278	0.286
	8) DMoBDA	0.458	0.295	0.242	0.257	0.250
	9) DSBDA	0.541	0.331	0.293	0.313	0.324
N(-0.5,1.2^2)	*1) No outlier*	0.189	0.206	0.212	0.194	0.189
	2) IMoBIA	0.425	0.313	0.266	0.281	0.262
	3) ISBIA	0.600	0.388	0.341	0.337	0.339
	4) DMoBIA	0.416	0.303	0.264	0.270	0.271
	5) DSBIA	0.599	0.414	0.336	0.341	0.345
	6) IMoBDA	0.458	0.295	0.275	0.305	0.282
	7) ISBDA	0.560	0.346	0.292	0.300	0.303
	8) DMoBDA	0.491	0.319	0.271	0.278	0.270
	9) DSBDA	0.545	0.321	0.280	0.297	0.302

Table 20 shows that the robust methods have slightly larger weighted bias than the traditional scale transformation methods under the conditions of *No outlier*. When a single outlier was simulated, the *Baseline* method constantly had the largest weighted bias. The robust Haebara method generally produced the least overall weighted bias when moderate *b*-parameter changes were simulated, and the LAV method produced the least overall weighted bias when severe *b*-parameter changes were simulated with $\theta \sim$ N(0, 1), $\theta \sim$ N(0.25, 1.1^2) and $\theta \sim$ N(-0.25, 1.1^2). The robust Haebara method produced smaller weighted bias under most outlier conditions with $\theta \sim$ N(0. 5, 1.2^2), and the LAV method produced the least weighted bias with $\theta \sim$ N(-0. 5, 1.2^2). The outlier removal methods produced substantive smaller weighted bias than the *Baseline* method, and they generally

had similar amount of bias to the robust methods. Interestingly, the displacement method usually had slightly smaller weighted bias than the direct exclusion method.

Table 20a. Weighted Bias Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Baseline	Robust Haebara	LAV	Displacement	Exclusion
N(0,1)	*1) No outlier*	0.025	0.034	0.034	0.027	0.025
	2) IMoBIA	0.303	0.100	0.093	0.103	0.126
	3) ISBIA	0.434	0.198	0.189	0.220	0.224
	4) DMoBIA	0.280	0.084	0.111	0.113	0.142
	5) DSBIA	0.451	0.199	0.178	0.219	0.222
	6) IMoBDA	0.220	0.076	0.093	0.117	0.140
	7) ISBDA	0.405	0.117	0.109	0.172	0.177
	8) DMoBDA	0.234	0.055	0.056	0.096	0.124
	9) DSBDA	0.415	0.135	0.136	0.177	0.191
$N(0.25,1.1^2)$	*1) No outlier*	0.027	0.015	0.015	0.024	0.027
	2) IMoBIA	0.244	0.079	0.087	0.078	0.112
	3) ISBIA	0.411	0.154	0.154	0.171	0.179
	4) DMoBIA	0.260	0.062	0.074	0.064	0.102
	5) DSBIA	0.398	0.141	0.129	0.181	0.184
	6) IMoBDA	0.212	0.035	0.079	0.110	0.118
	7) ISBDA	0.360	0.151	0.147	0.176	0.187
	8) DMoBDA	0.228	0.044	0.057	0.120	0.112
	9) DSBDA	0.376	0.135	0.136	0.176	0.190
$N(0.5,1.2^2)$	*1) No outlier*	0.017	0.023	0.018	0.018	0.017
	2) IMoBIA	0.250	0.090	0.090	0.062	0.116
	3) ISBIA	0.439	0.159	0.143	0.162	0.172
	4) DMoBIA	0.258	0.048	0.086	0.087	0.113
	5) DSBIA	0.432	0.158	0.167	0.182	0.189
	6) IMoBDA	0.232	0.086	0.104	0.132	0.140
	7) ISBDA	0.409	0.148	0.152	0.207	0.211
	8) DMoBDA	0.210	0.060	0.097	0.117	0.139
	9) DSBDA	0.408	0.131	0.141	0.189	0.197

Table 20b. Weighted Bias Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Baseline	Robust Haebara	LAV	Displacement	Exclusion
$N(-0.25,1.1^2)$	*1) No outlier*	0.032	0.025	0.023	0.030	0.032
	2) IMoBIA	0.260	0.076	0.101	0.105	0.132
	3) ISBIA	0.392	0.183	0.168	0.217	0.223
	4) DMoBIA	0.249	0.087	0.116	0.125	0.153
	5) DSBIA	0.414	0.185	0.172	0.212	0.216
	6) IMoBDA	0.295	0.064	0.078	0.082	0.115
	7) ISBDA	0.446	0.134	0.118	0.158	0.168
	8) DMoBDA	0.281	0.095	0.094	0.125	0.142
	9) DSBDA	0.408	0.135	0.130	0.182	0.197
$N(-0.5,1.2^2)$	*1) No outlier*	0.040	0.047	0.035	0.042	0.040
	2) IMoBIA	0.270	0.085	0.069	0.093	0.117
	3) ISBIA	0.396	0.173	0.158	0.203	0.203
	4) DMoBIA	0.251	0.108	0.102	0.117	0.139
	5) DSBIA	0.418	0.181	0.180	0.209	0.206
	6) IMoBDA	0.235	0.091	0.107	0.126	0.135
	7) ISBDA	0.421	0.136	0.107	0.157	0.160
	8) DMoBDA	0.290	0.105	0.097	0.114	0.118
	9) DSBDA	0.409	0.135	0.104	0.148	0.152

Table 21 indicates that the *Baseline* method had the least weighted standard errors of equating under the condition of *No outlier*. When a single outlying common item was included in the scale transformation, considerable random errors were produced. The outlier detection and elimination procedures remarkably reduced the random errors. The weighted standard errors obtained by the direct exclusion method were the least among the investigated methods when moderate *b*-parameter changes were simulated. The displacement method produced the least weighted standard errors when severe *b*-parameter changes were simulated. The proposed robust methods also produced smaller weighted standard error than the *Baseline* method, but the amount was similar to those produced by the outlier removal methods.

Table 21a. Weighted Standard Error Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Baseline	Robust Haebara	LAV	Displacement	Exclusion
N(0,1)	*1) No outlier*	0.172	0.184	0.185	0.175	0.172
	2) IMoBIA	0.304	0.276	0.234	0.244	0.223
	3) ISBIA	0.402	0.312	0.268	0.258	0.267
	4) DMoBIA	0.310	0.290	0.229	0.246	0.224
	5) DSBIA	0.399	0.267	0.241	0.235	0.247
	6) IMoBDA	0.342	0.264	0.238	0.260	0.228
	7) ISBDA	0.292	0.290	0.243	0.218	0.221
	8) DMoBDA	0.327	0.266	0.244	0.255	0.230
	9) DSBDA	0.290	0.289	0.263	0.217	0.218
$N(0.25,1.1^2)$	*1) No outlier*	0.153	0.179	0.197	0.153	0.153
	2) IMoBIA	0.291	0.274	0.230	0.269	0.235
	3) ISBIA	0.417	0.303	0.272	0.246	0.251
	4) DMoBIA	0.284	0.277	0.223	0.268	0.223
	5) DSBIA	0.385	0.283	0.244	0.235	0.240
	6) IMoBDA	0.377	0.271	0.238	0.244	0.215
	7) ISBDA	0.281	0.272	0.241	0.226	0.230
	8) DMoBDA	0.360	0.260	0.225	0.243	0.215
	9) DSBDA	0.301	0.277	0.230	0.214	0.215
$N(0.5,1.2^2)$	*1) No outlier*	0.176	0.206	0.208	0.179	0.176
	2) IMoBIA	0.285	0.282	0.230	0.261	0.231
	3) ISBIA	0.399	0.305	0.271	0.262	0.269
	4) DMoBIA	0.303	0.265	0.224	0.235	0.221
	5) DSBIA	0.409	0.321	0.286	0.263	0.271
	6) IMoBDA	0.349	0.284	0.266	0.253	0.246
	7) ISBDA	0.298	0.279	0.234	0.207	0.208
	8) DMoBDA	0.350	0.305	0.269	0.268	0.247
	9) DSBDA	0.279	0.278	0.232	0.213	0.213

Table 21b. Weighted Standard Error Statistics for IRT True Score Equating with Proposed Robust Scale Transformation Methods

Theta distribution	Outlier	Baseline	Robust Haebara	LAV	Displacement	Exclusion
	1) No outlier	0.184	0.196	0.213	0.189	0.184
	2) IMoBIA	0.311	0.282	0.238	0.244	0.229
	3) ISBIA	0.381	0.318	0.259	0.229	0.246
	4) DMoBIA	0.281	0.245	0.233	0.256	0.231
$N(-0.25,1.1^2)$	*5) DSBIA*	0.424	0.303	0.265	0.248	0.261
	6) IMoBDA	0.344	0.274	0.254	0.271	0.237
	7) ISBDA	0.303	0.271	0.237	0.215	0.215
	8) DMoBDA	0.333	0.277	0.218	0.219	0.193
	9) DSBDA	0.305	0.299	0.255	0.241	0.240
	1) No outlier	0.182	0.197	0.206	0.186	0.182
	2) IMoBIA	0.304	0.292	0.254	0.259	0.231
	3) ISBIA	0.418	0.338	0.300	0.263	0.265
	4) DMoBIA	0.311	0.274	0.240	0.241	0.230
$N(-0.5,1.2^2)$	*5) DSBIA*	0.389	0.366	0.282	0.266	0.273
	6) IMoBDA	0.368	0.280	0.249	0.275	0.241
	7) ISBDA	0.332	0.315	0.267	0.246	0.247
	8) DMoBDA	0.365	0.299	0.250	0.251	0.238
	9) DSBDA	0.324	0.288	0.256	0.249	0.251

Indices of Equating Errors

The plots of RMSE at each score point for the IRT true score equating for both test forms are shown in Figures 25 to 29, the plots of bias at each score point for the IRT true score equating for both test forms are shown in Figures 30 to 34, and the plots of standard error at each score point for the IRT true score equating for both test forms are shown in Figures 35 to 39. The ability differences do not have much impact on the comparison among the scale transformation methods in terms of the equating errors. Although shapes of the error curves are a little different with various ability distributions, the differences in equating errors are similar among the scale transformation methods, and they are similar to what was observed when $\theta \sim N(0,1)$. As a consequence, results with large

ability differences are not separately discussed in detail, although the figures are presented.

Figure 25. The RMSE statistics of IRT equating ($\theta \sim N(0,1)$)

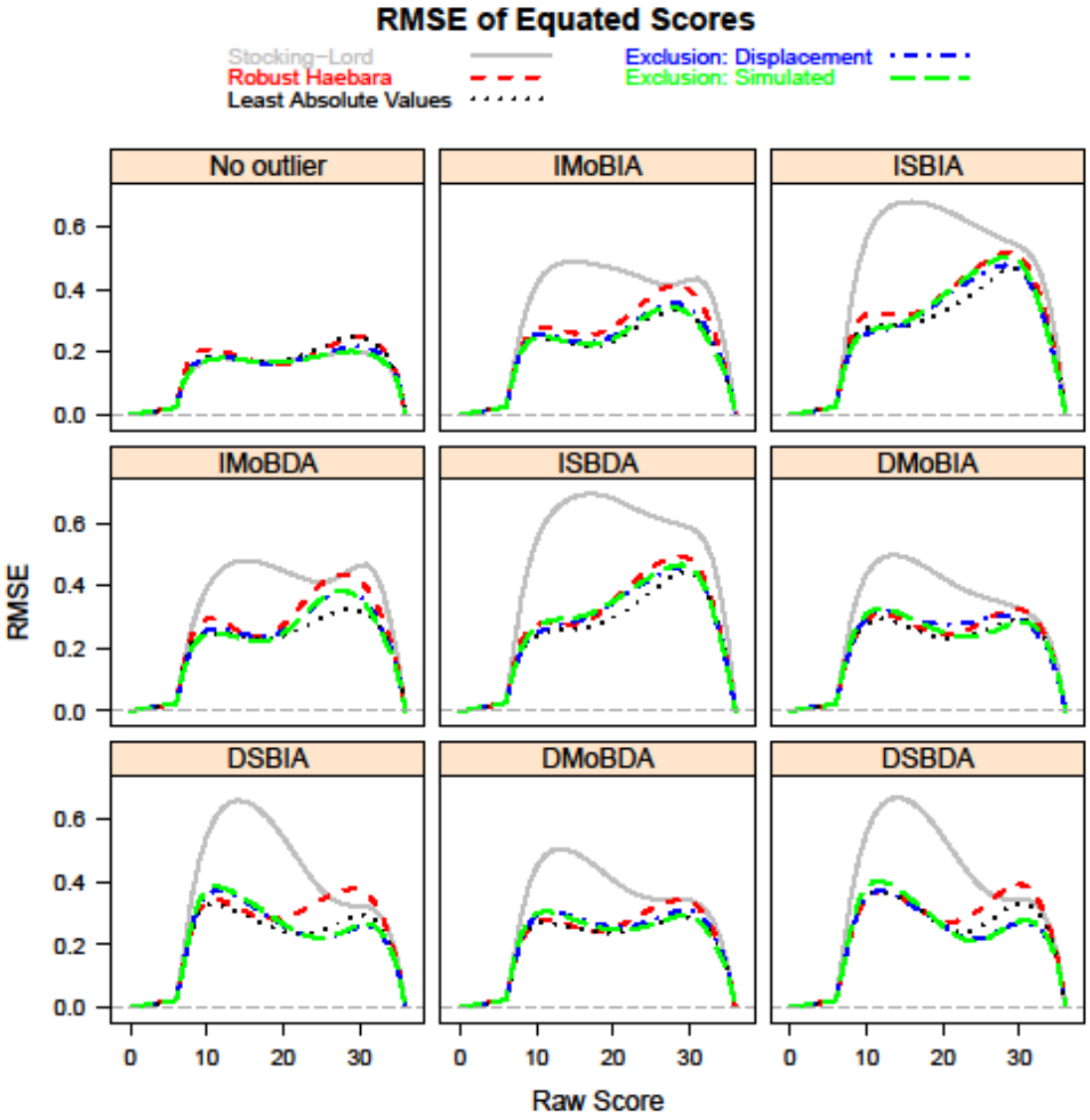


Figure 26. The RMSE statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

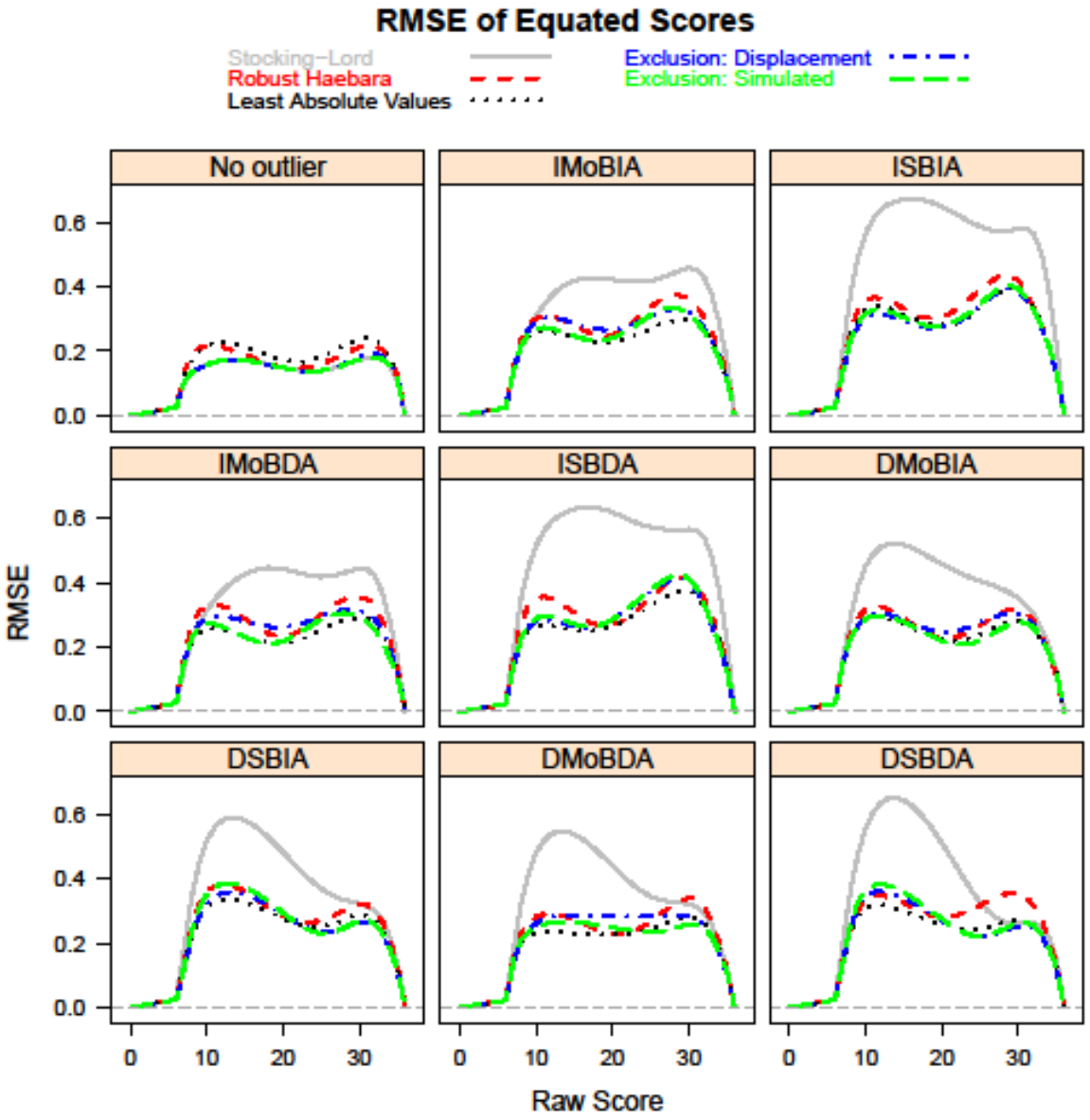


Figure 27. The RMSE statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)

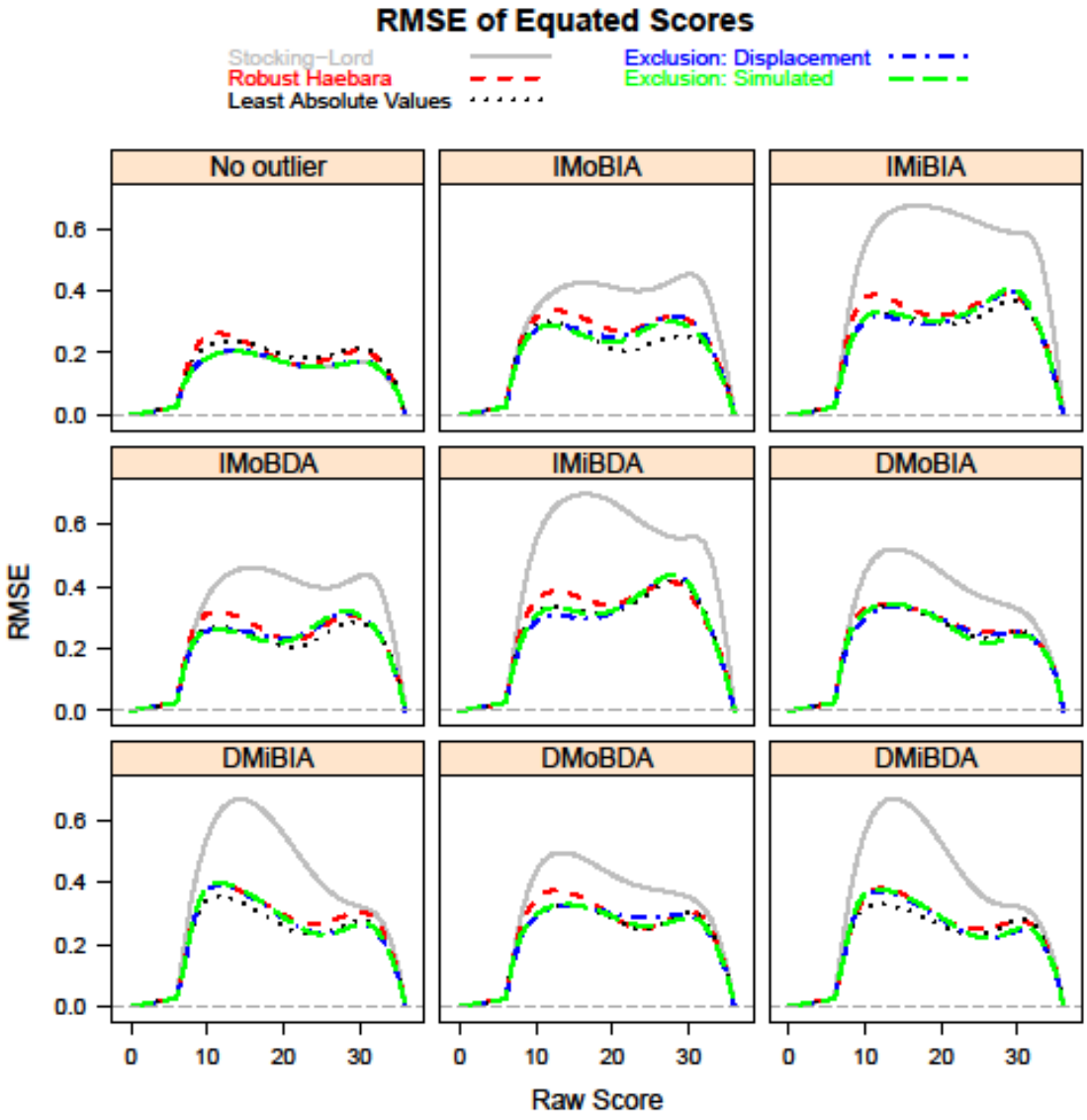


Figure 28. The RMSE statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$)

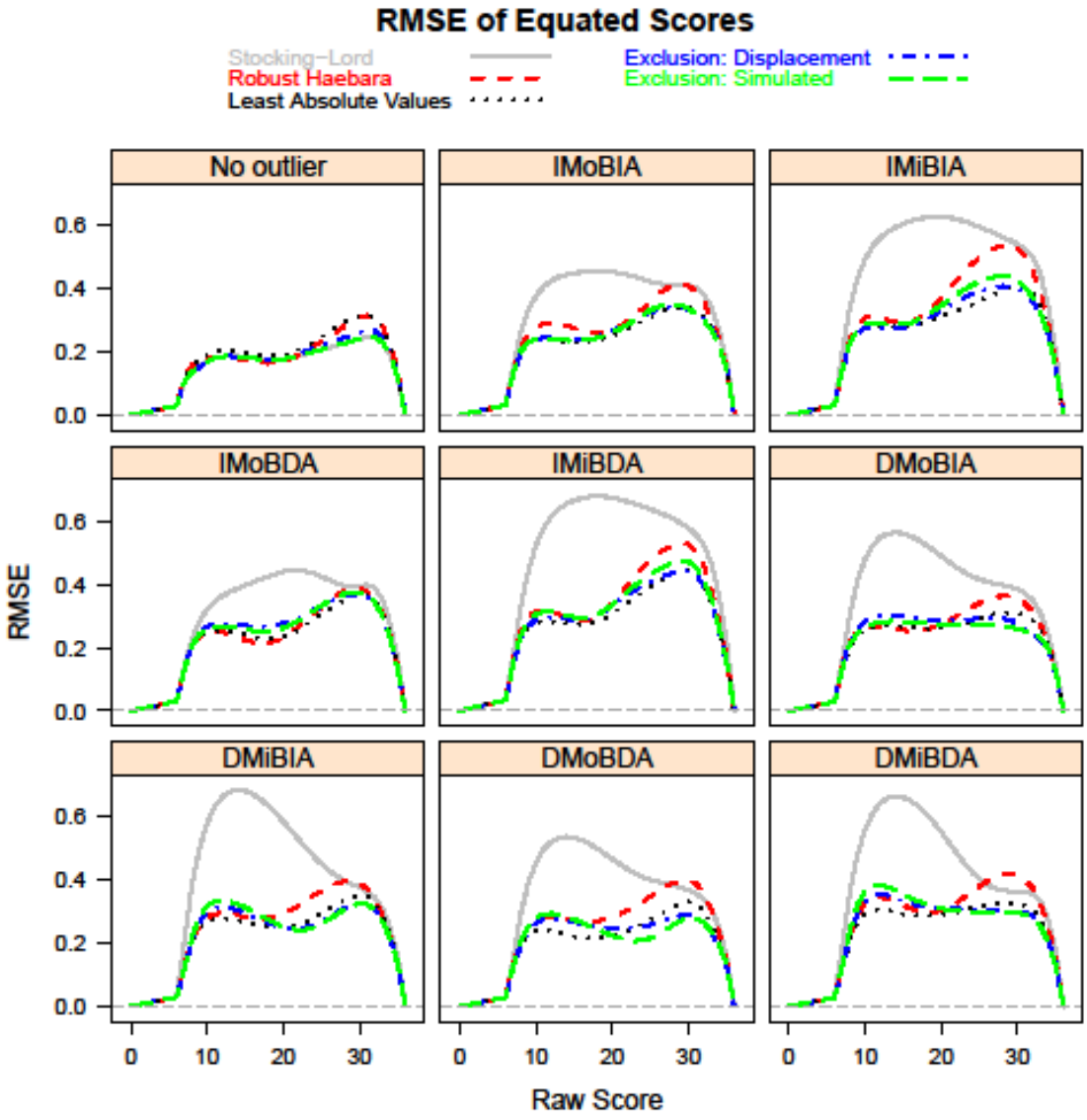
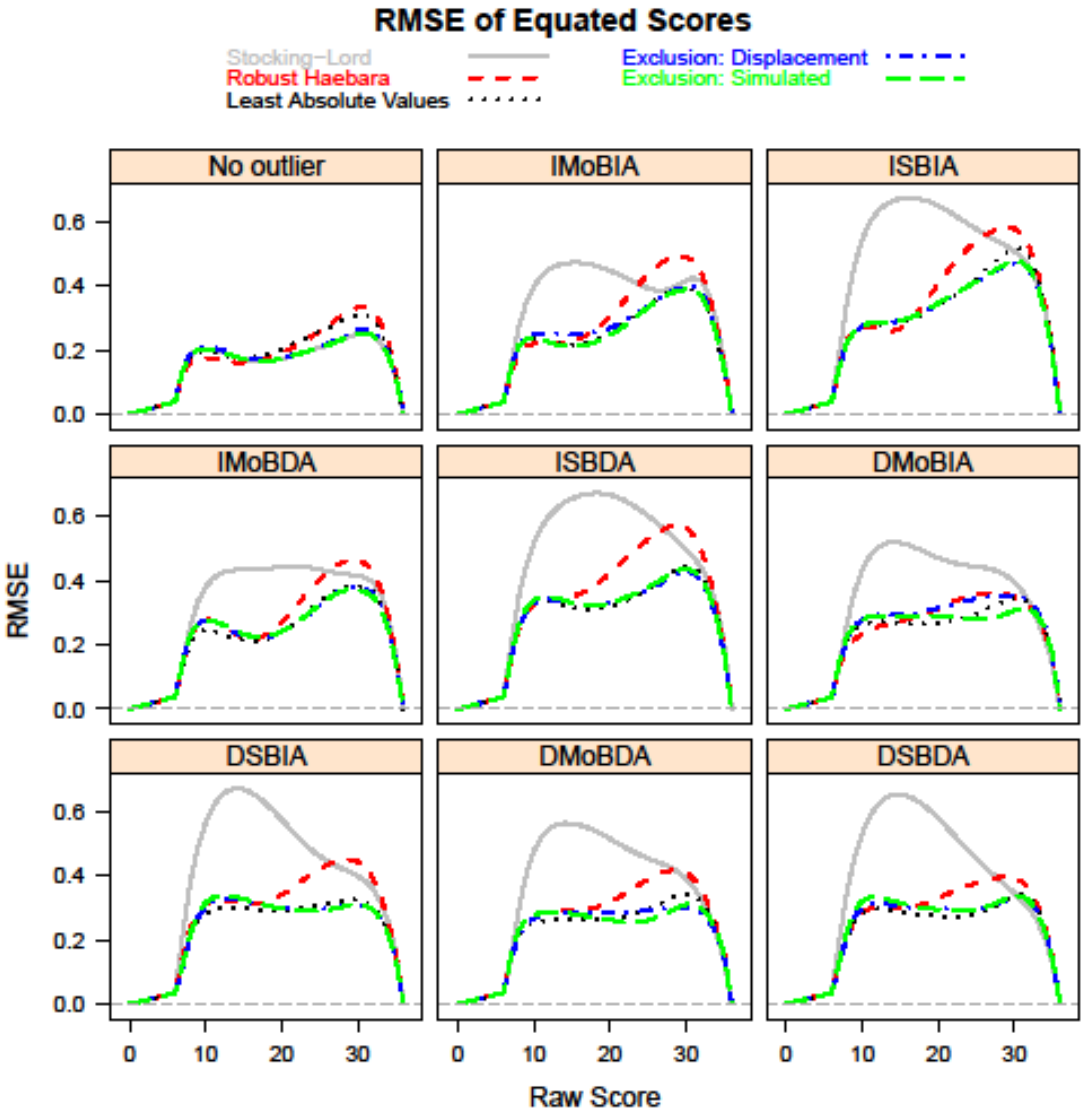


Figure 29. The RMSE statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$)



Figures 25 to 29 plots the RMSE results of equated scores under different outlier situations and methods of scale transformation and different ability distributions of examinees. The figures indicate that when no outlier was simulated, the equating errors obtained by investigated methods were slightly larger than the *Baseline* method, especially for either low scores or high scores. When a single outlying item with moderate change of b-parameter was simulated in the common item set, equating errors increased. Both the proposed robust methods and the outlier elimination methods dramatically reduced the equating errors. The robust Haebara method had larger errors than the other methods almost throughout the entire scale when b-parameter increased (*IMoBIA* and *IMoBDA*), and high scores when b-parameter decreased (*DMoBIA* and *DMoBDA*). The LAV method had slightly smaller errors than the other methods throughout the entire scale. When the magnitude of b-parameter change was increased, the equating errors remarkably increased, particularly for the traditional Stocking-Lord method. The robust methods and the outlier removal methods produced similar amount of RMSEs throughout the entire scale. The discrepancy among the methods was nearly imperceptible except when *b*-parameter decreased with $\theta \sim N(0,1)$. In addition, the robust methods had larger RMSEs than the outlier removal methods with $\theta \sim N(-0.25, 1.1^2)$ and $\theta \sim N(-0.5, 1.2^2)$,.

Figures 30 to 34 show the bias plots for equated scores under different outlier situations and methods of the traditional scale transformation and various ability distributions of examinees. All scale transformation methods produced approximately identical absolute biases when no outlier is simulated since the bias curves were very close to each other. However, the bias obtained by the robust Deming method is slightly

larger than the others. When a single outlying item was simulated in the common item set, more systematic errors were introduced to the equating procedure after transformation. The proposed robust method and the outlier removal methods led to smaller bias. The robust method generally had smaller bias values than the outlier removal methods. It is more evident if a single severe outlier was simulated.

Figure 30. The Bias statistics of IRT equating ($\theta \sim N(0,1)$)

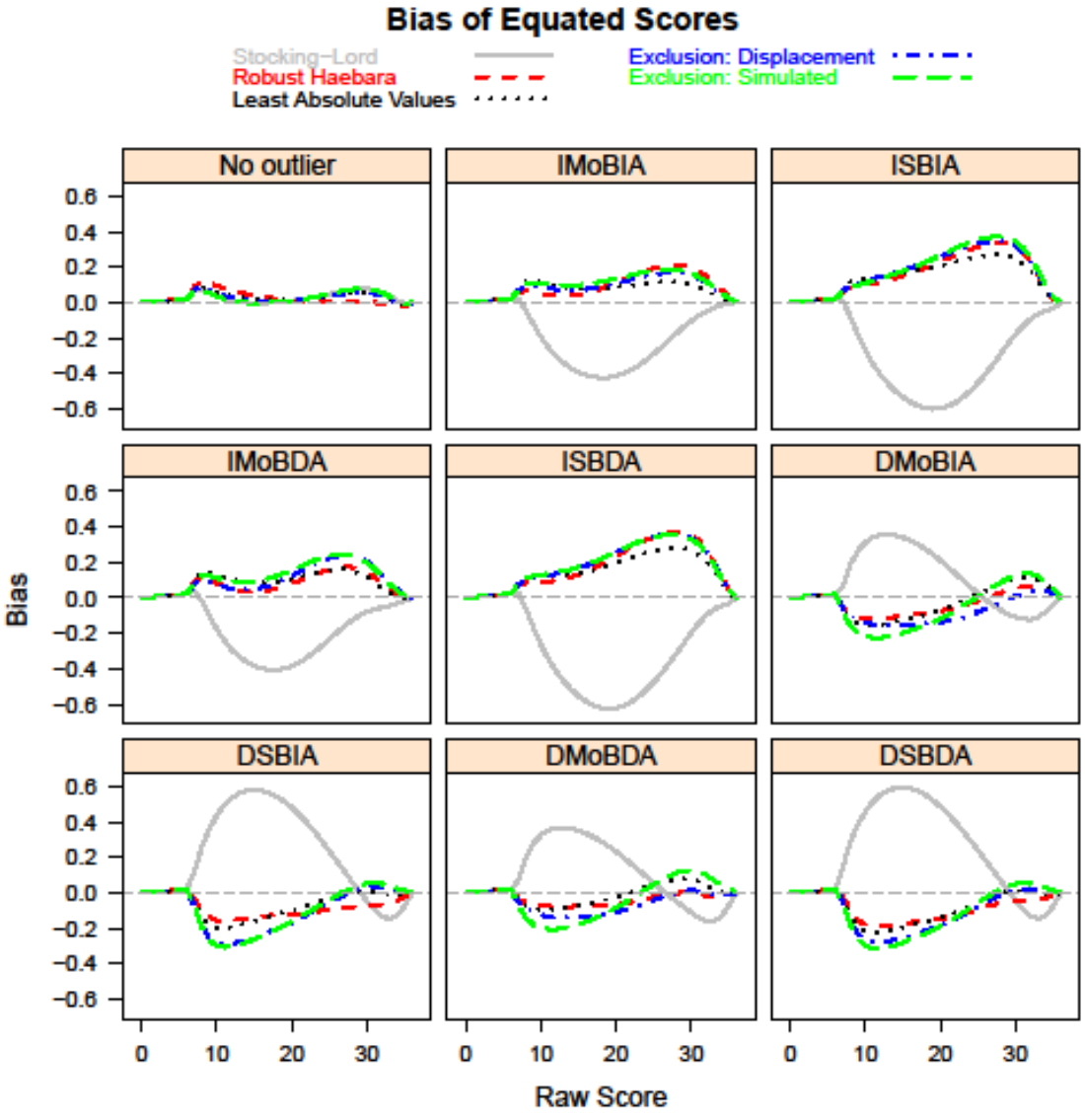


Figure 31. The Bias statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

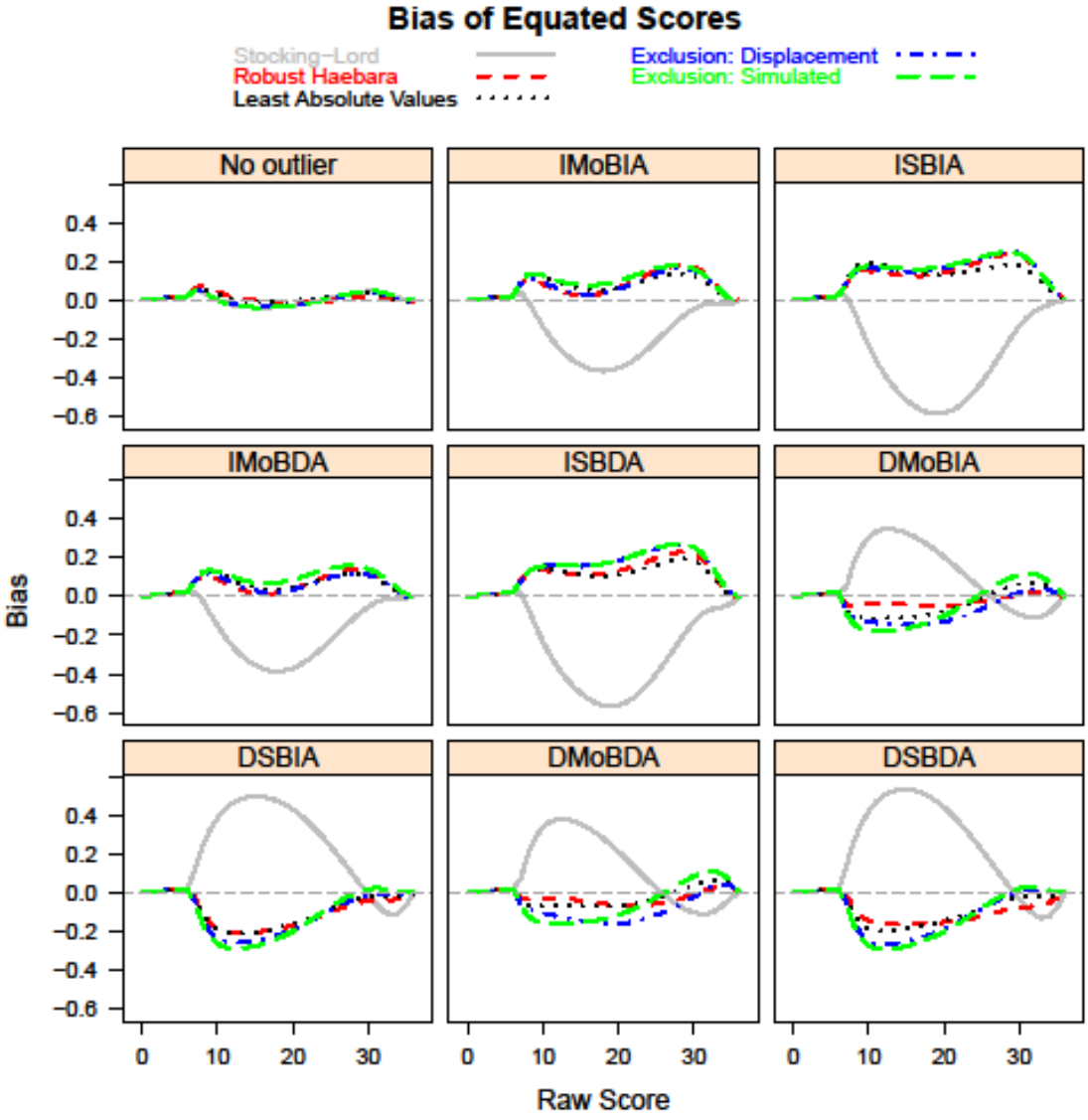


Figure 32. The Bias statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)

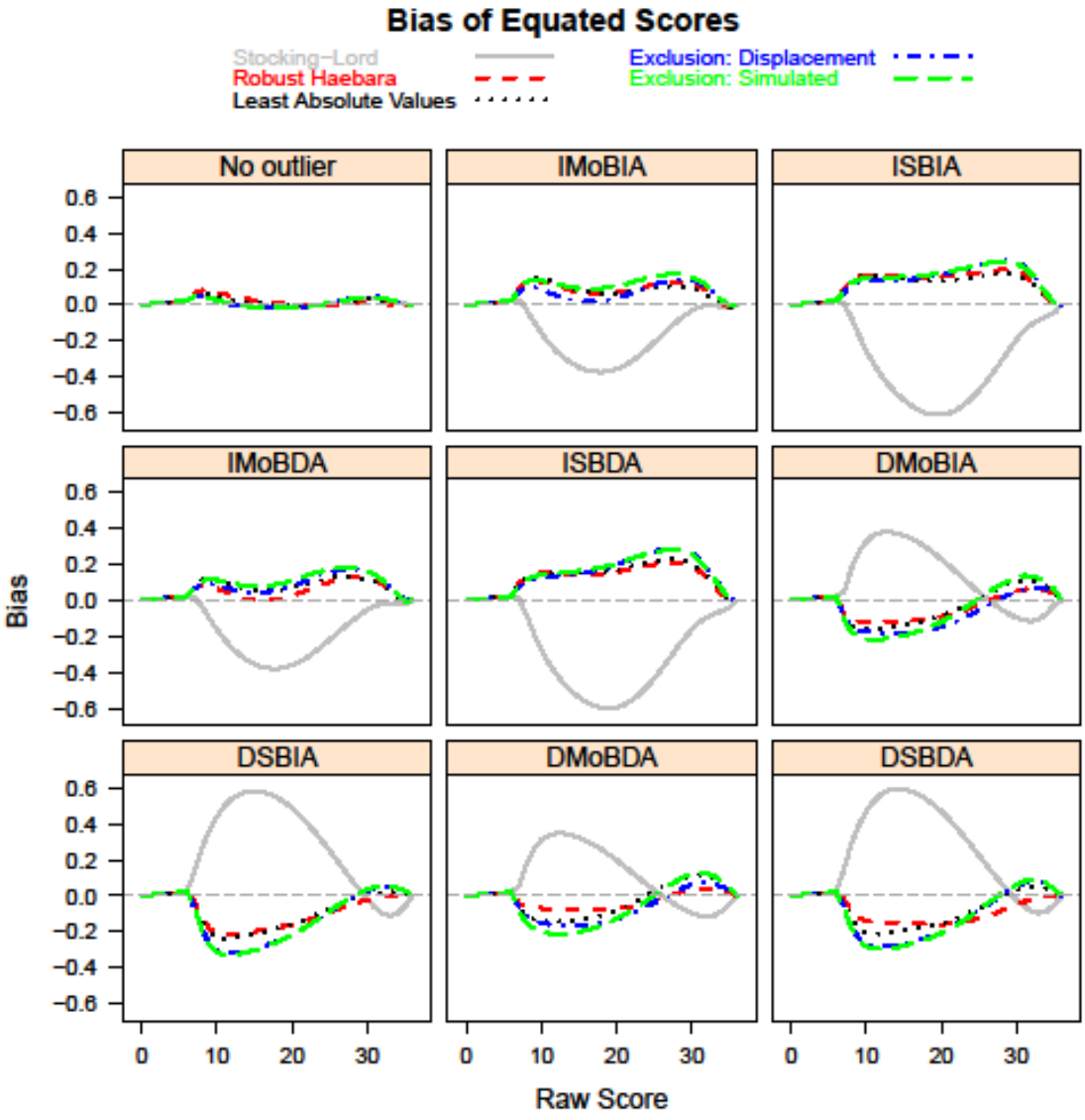


Figure 33. The Bias statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$)

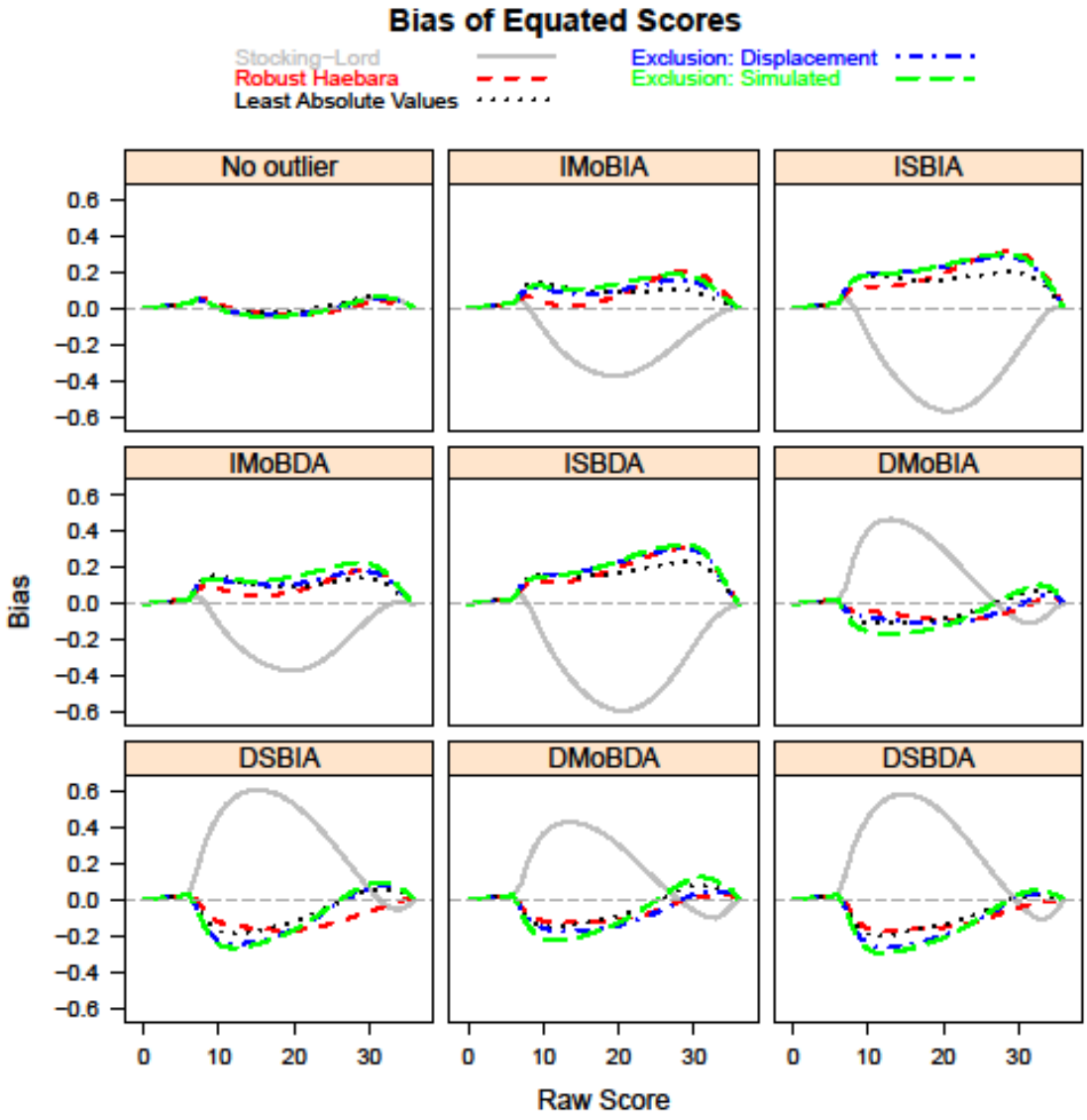
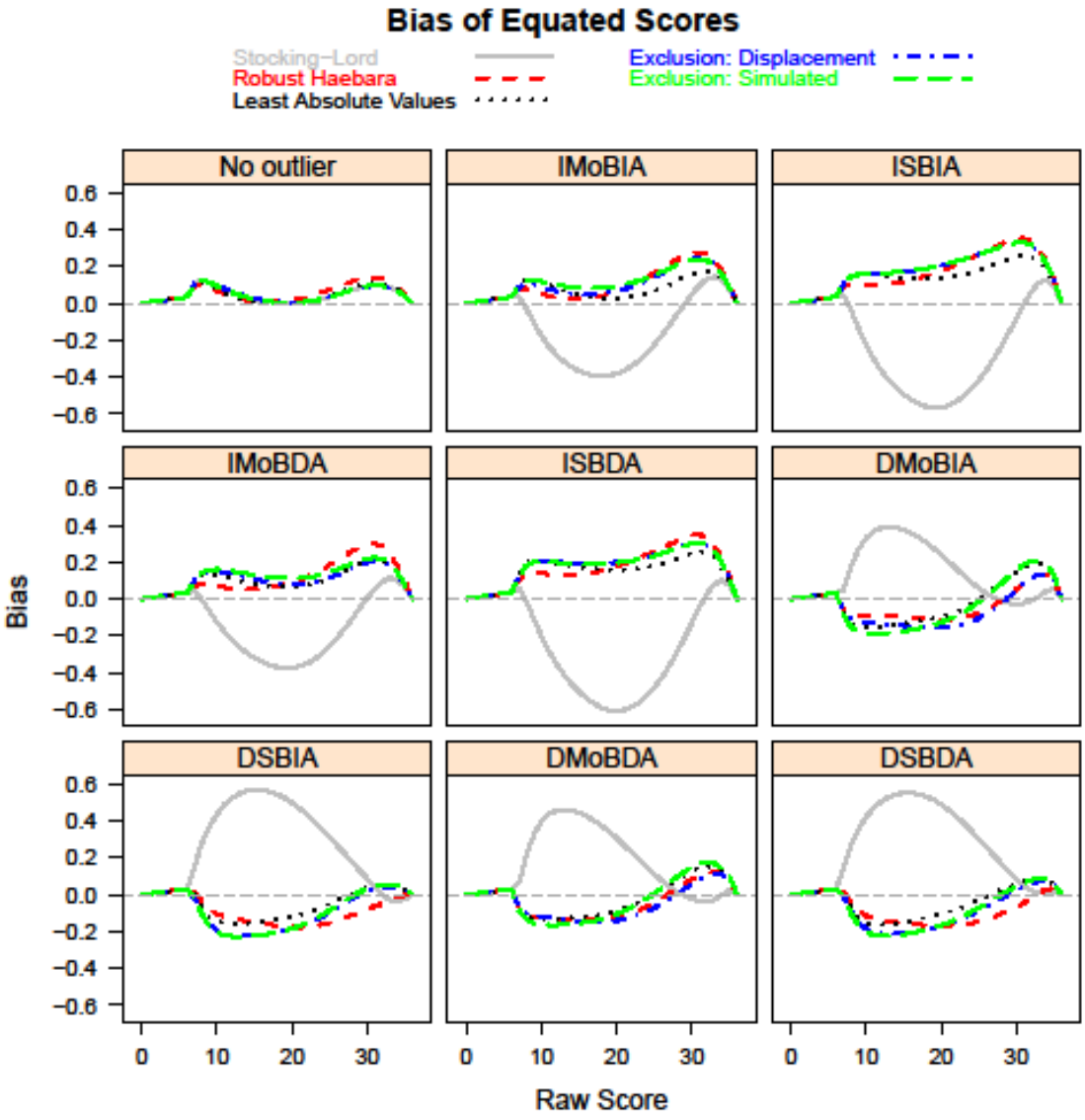


Figure 34. The Bias statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$)



Figures 35 to 39 show the plots of standard errors for equated scores under different outlier situations and methods of the traditional scale transformation and various ability distributions of examinees. The robust methods and the outlier removal methods generally had larger standard errors than the *Baseline* method for high scores when no outlying common item is simulated. When increased *b*-parameter was simulated, the

robust Haebara method had slightly larger equating bias than the other methods. It is also true when b-parameter severely decreased.

Figure 35. The Standard Error statistics of IRT equating ($\theta \sim N(0,1)$)

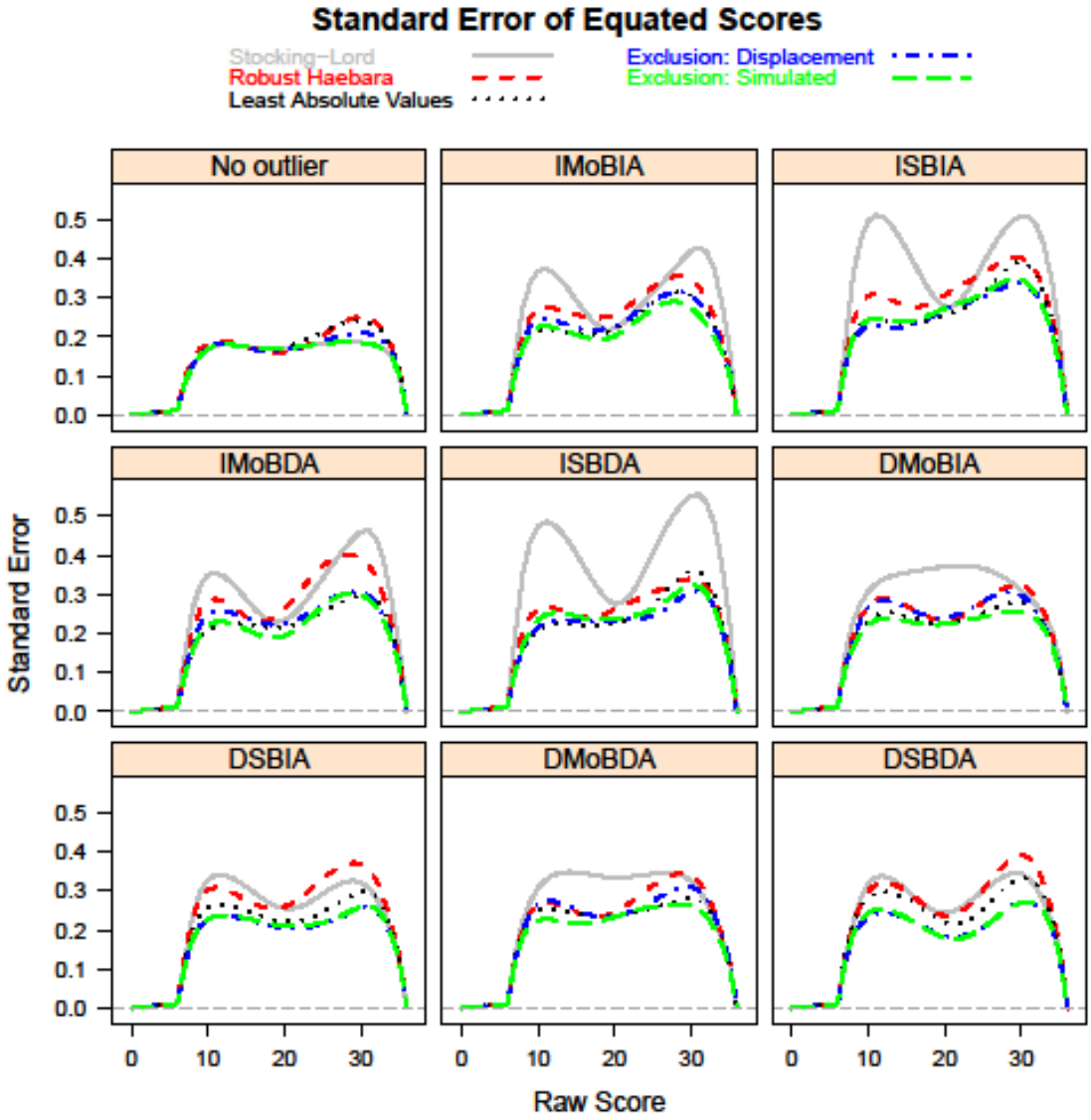


Figure 36. The Standard Error statistics of IRT equating ($\theta \sim N(0.25,1.1^2)$)

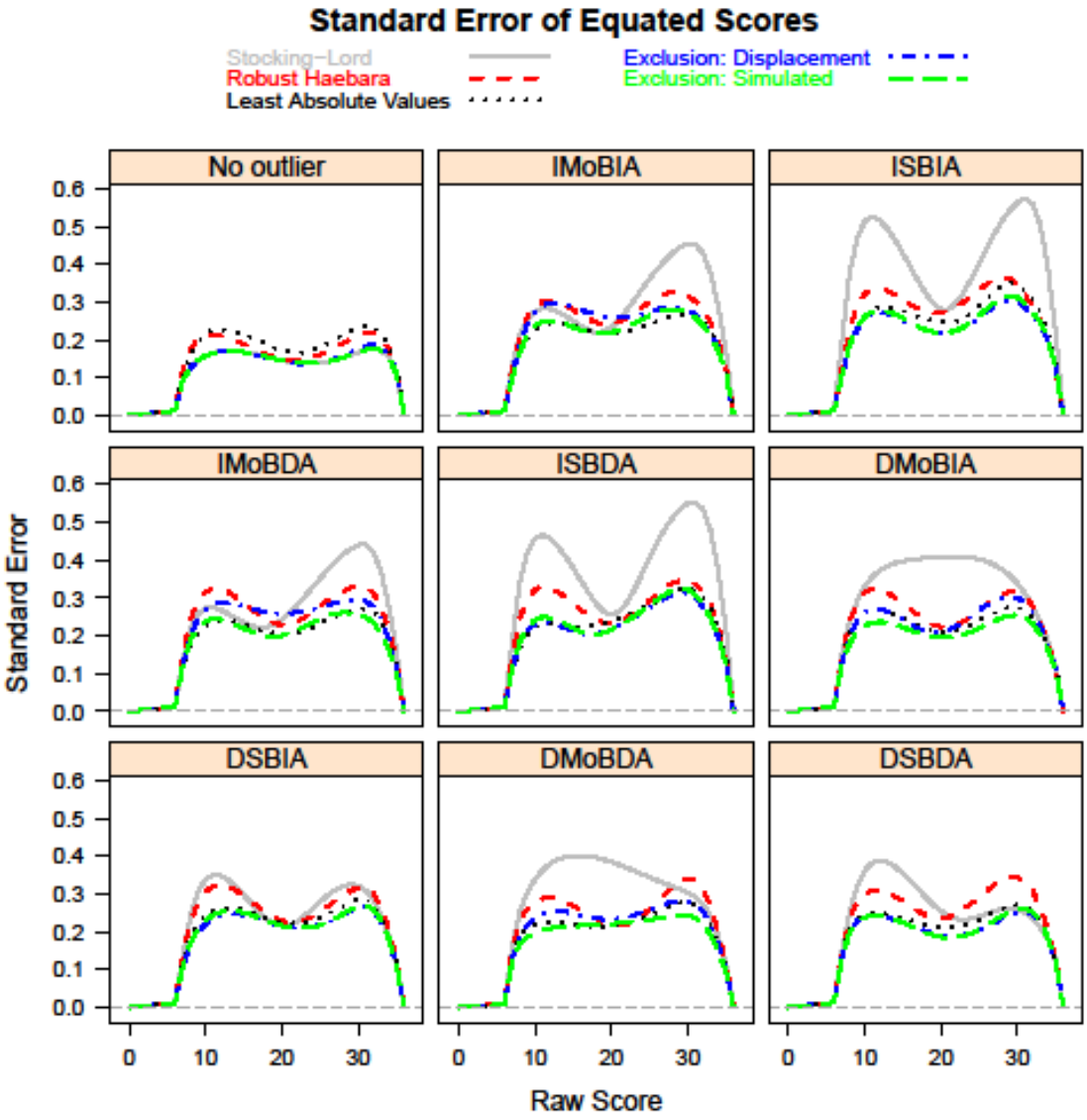


Figure 37. The Standard Error statistics of IRT equating ($\theta \sim N(0.5,1.2^2)$)

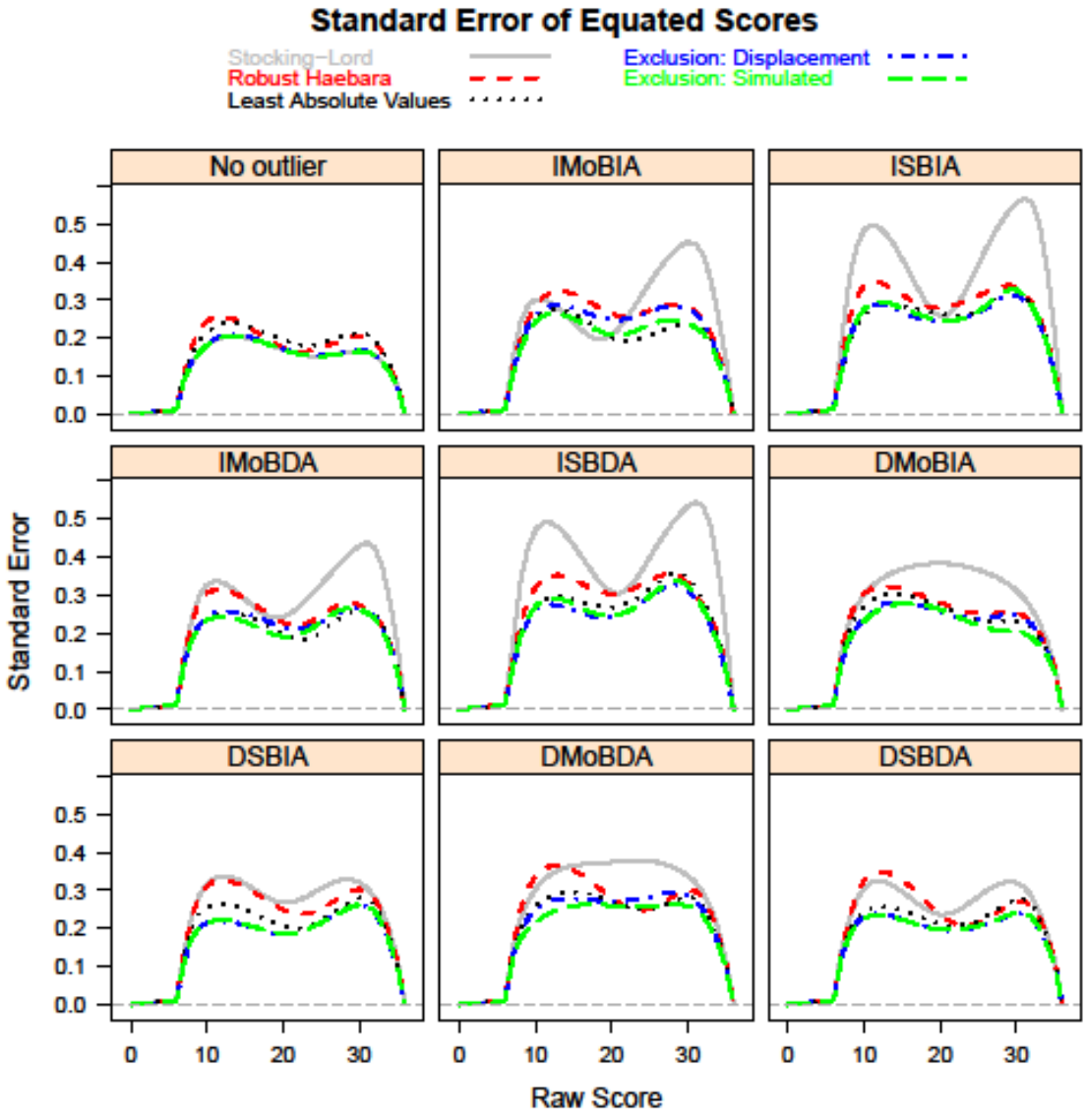


Figure 38. The Standard Error statistics of IRT equating ($\theta \sim N(-0.25,1.1^2)$)

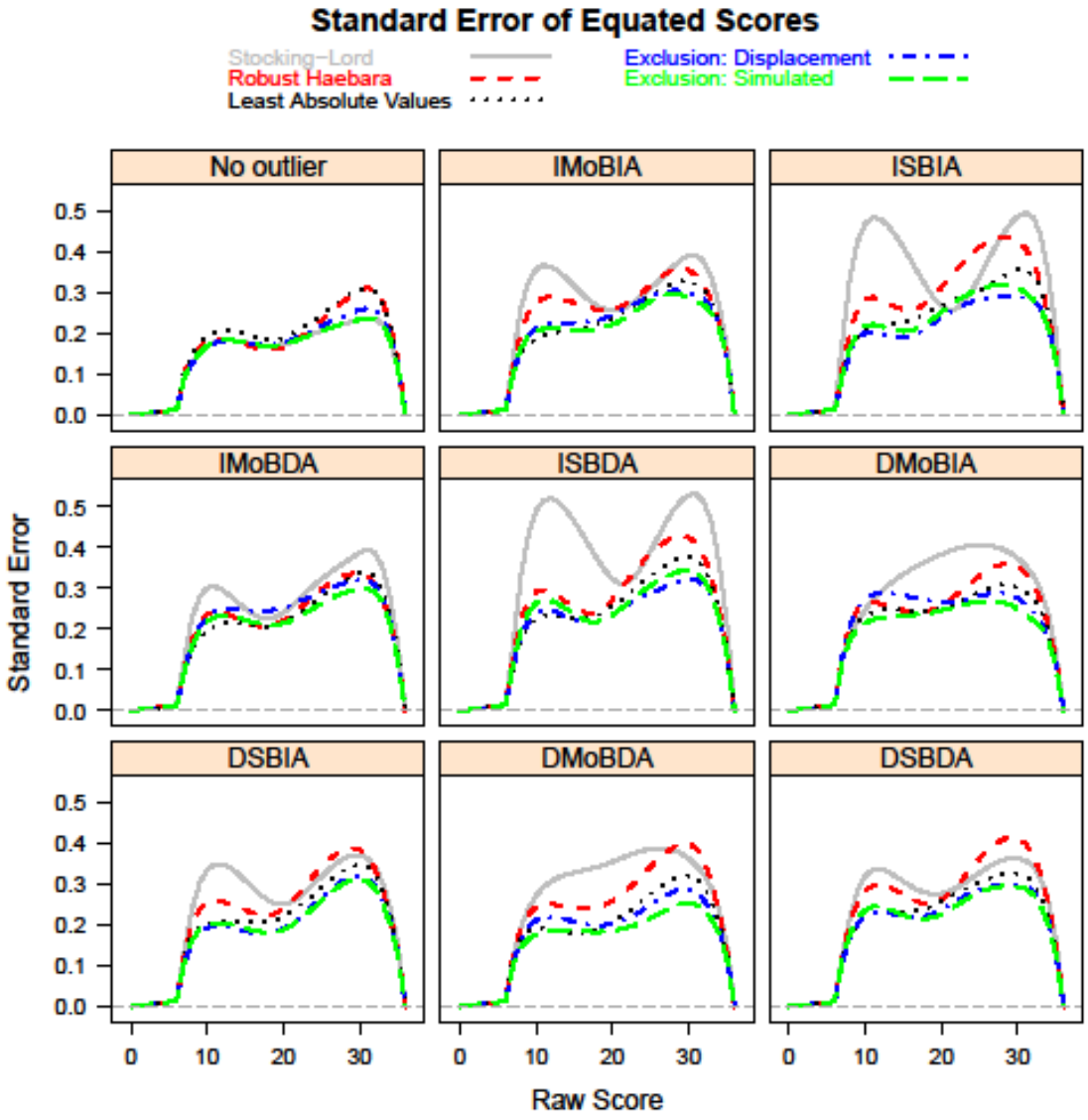
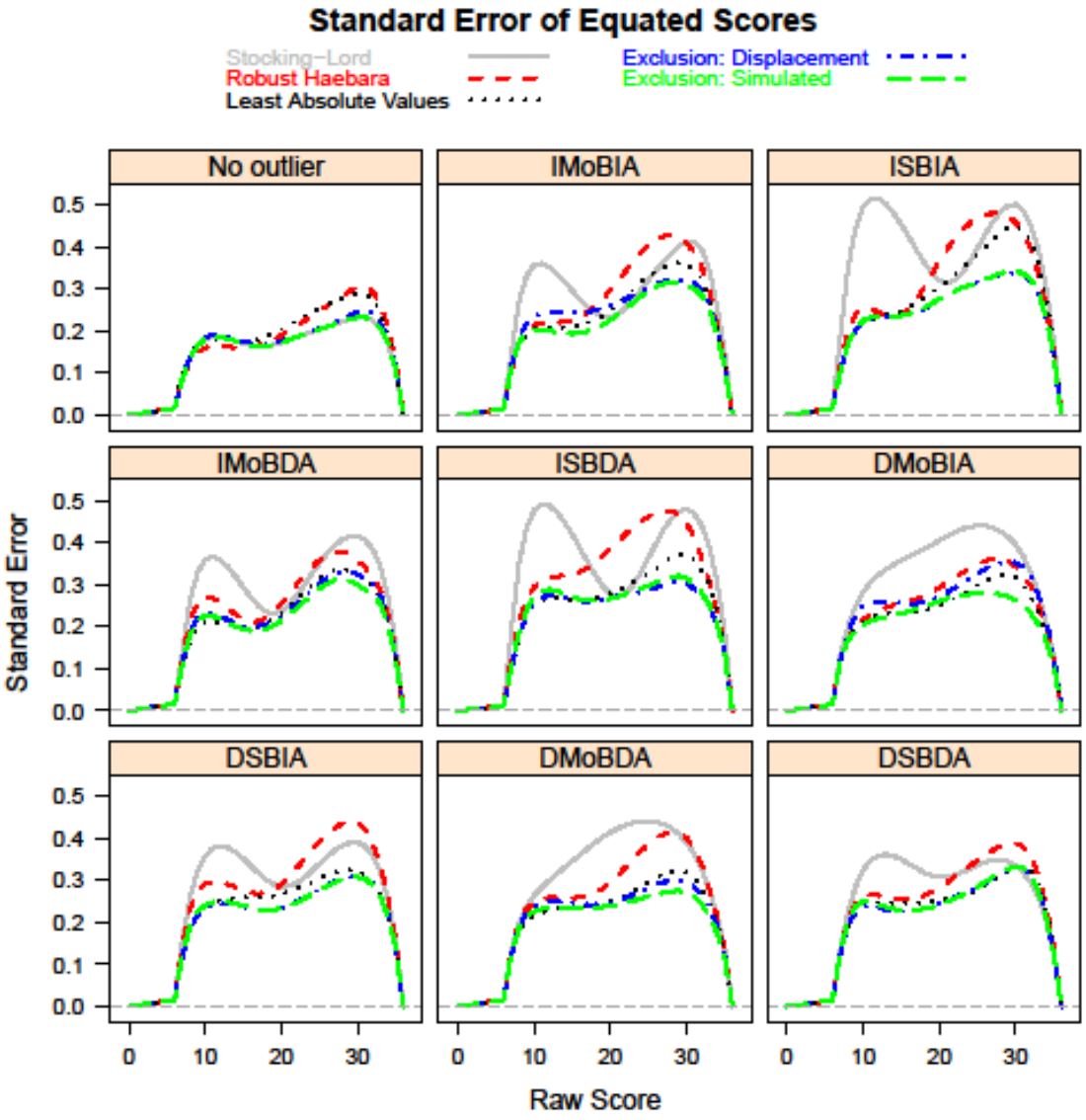


Figure 39. The Standard Error statistics of IRT equating ($\theta \sim N(-0.5,1.2^2)$)



Overall, the inclusion of a single simulated outlying common item enlarges the equating errors. The proposed robust method consistently has smaller errors than the others. Though the robust Haebara method generally reduces the equating error, it had slightly larger standard errors.

4.2 Numeric Illustration with Empirical Data

By using the *CBASE* data for subjects, English and Mathematics, the traditional characteristic-curve based scale transformation methods (the Stocking-Lord method and the Haebara method) and robust methods (the robust Deming method, the least absolute values method, and the robust Haebara method) were compared. The English subject included 41 multiple-choice items with 17 common items (41%). The Mathematics subject included 56 multiple-choice items with 14 common items (25%).

Table 22. Scale Transformation Coefficients for *CBASE* English and Mathematics

Subject	Coefficient	SL	Haebara	Robust Deming	LAV	Robust Haebara
English	*B*	0.1197	0.1312	0.0970	0.1139	0.1266
	A	0.9061	0.9189	0.8973	0.8807	0.9197
Mathematics	*B*	0.1569	0.1770	0.1274	0.1430	0.1764
	A	0.9147	0.8650	0.9001	0.8557	0.9091

The coefficients are summarized in Table 22. There was difference in the values of the scale transformation coefficients with the various scale transformation methods. Patterns of scale transformation were similar for both subjects. The robust Deming method produced smaller *B* coefficient, and had similar A coefficient to the traditional characteristic-curve methods. The scale transformation coefficients *A* and *B* obtained by the least absolute values method were usually smaller than the traditional characteristic-curve methods. On the other hand, the robust Haebara method yielded larger scale transformation coefficients *A* and *B* than the other methods.

The *b*- and *a*- parameters of the common items for the English subject were plotted in Figures 40 and 41. The rescaled item parameter estimates for the English subject were

used to estimate the true score equating functions. The resulting equivalents are shown in Figure 42.

Figure 40. The transformations of item difficulties (*b*-parameters) for English

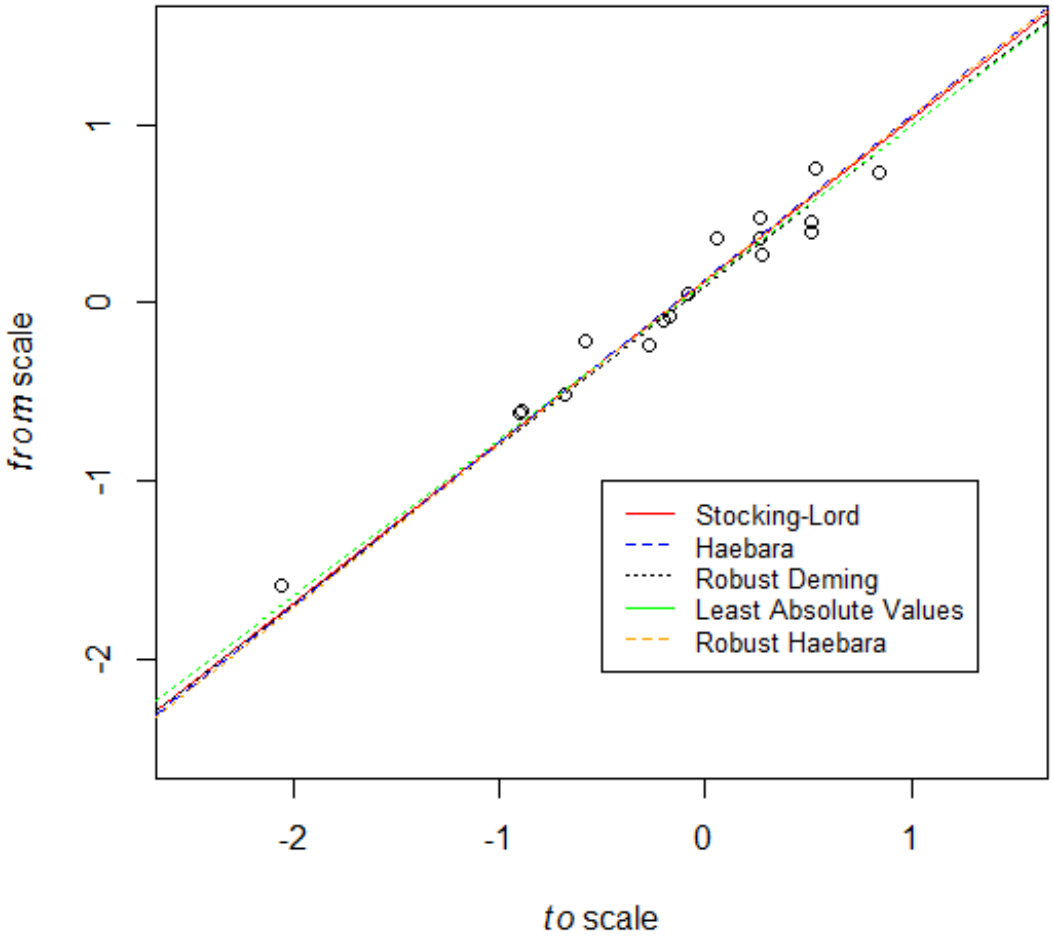


Figure 41. The transformations of item discrimination (*a*-parameters) for English

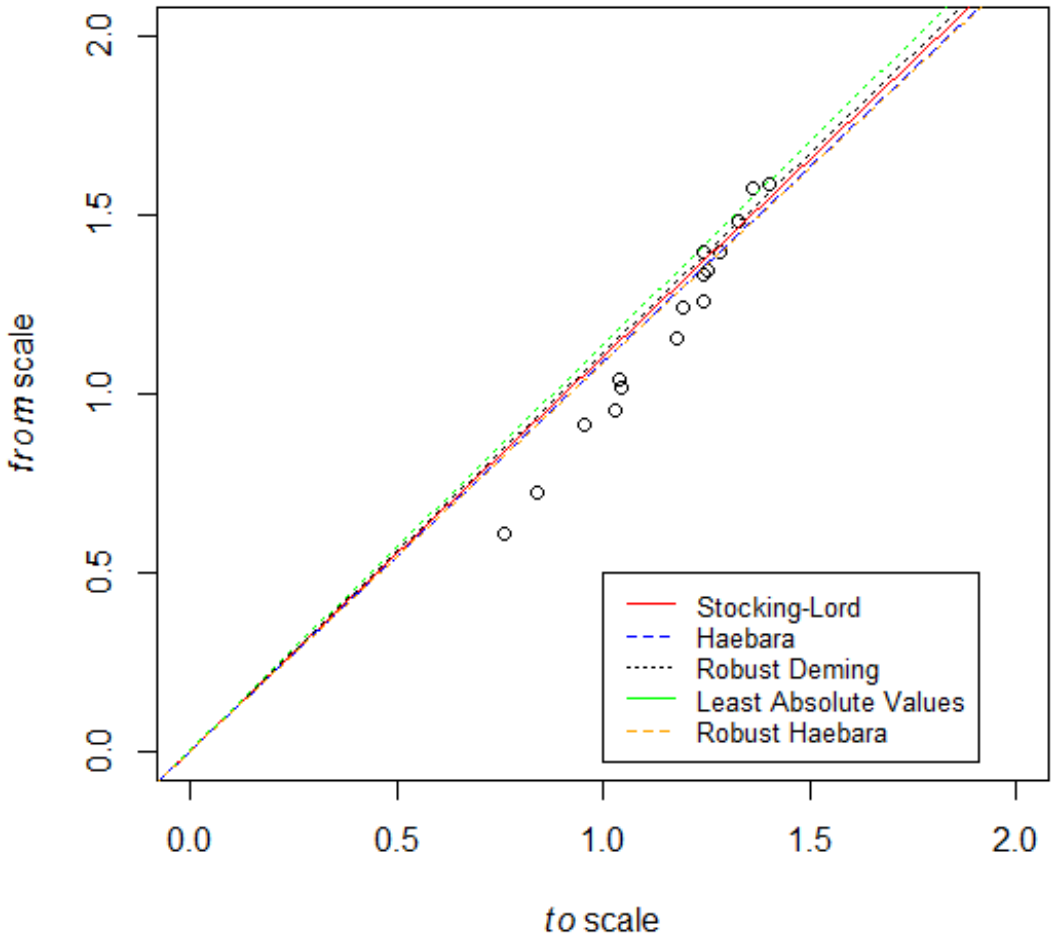


Figure 40 indicates that the *b*- parameters of common items for the English subject were closely along the line. The lines yielded by the investigated methods are almost identical to each other, except the LAV method has a slightly smaller slope. Figure 41 shows that the *a*- parameters of common items for the English subject were deviated from the lines with zero intercept. The lines produced by the investigated methods separately. Figure 42 indicates that the equated scores are different. It indicates that the true scores obtained by using the Stocking-Lord method are higher than others at a given true score on the *to* scale. The Haebara method has similar equated scores to the Stocking-Lord

method when the true scores on the *to* scale are approximately between 15 and 38. The robust methods have similar equated scores to the Haebara method when the true scores on the *to* scale are beyond the range approximately between 15 and 38. Generally, the equated scores according to the robust method are smaller than the traditional scale transformation methods.

Figure 42. Estimated *to* scale true score equivalents of *from* scale true scores using IRT true score equating (English)

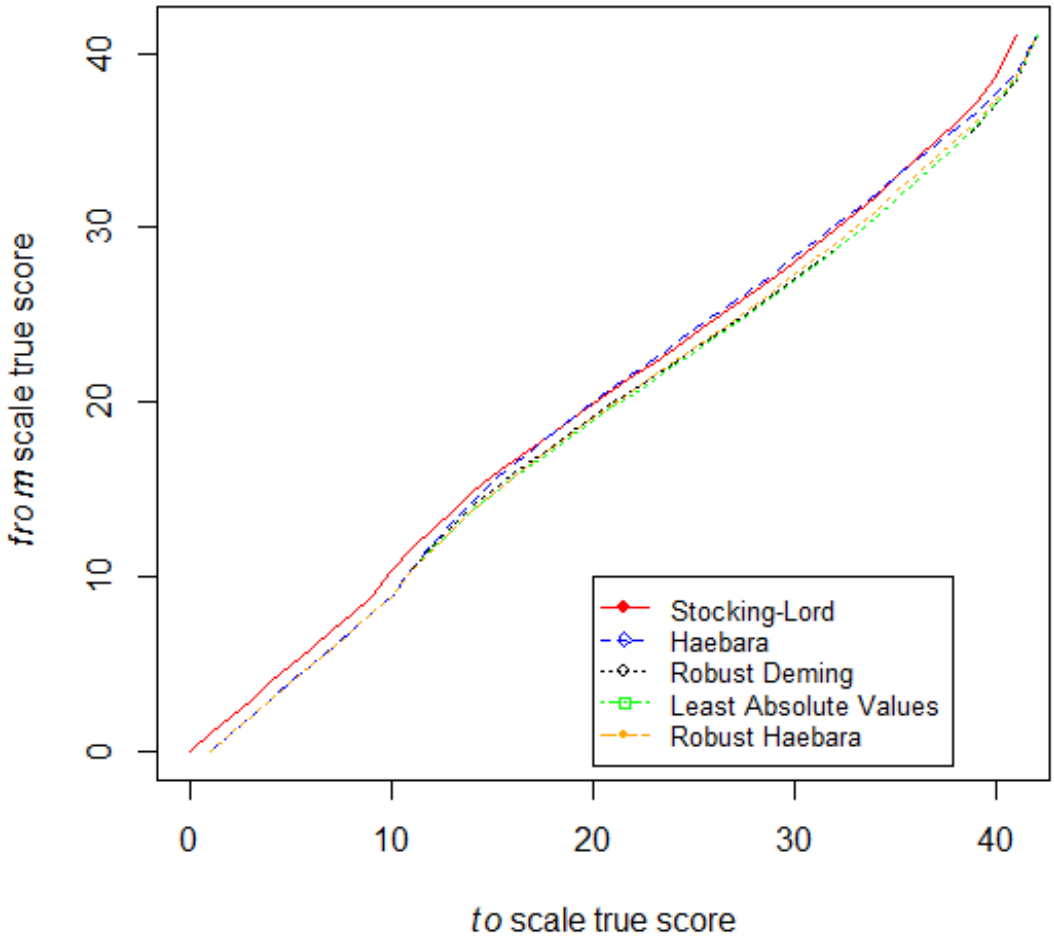


Figure 43. The transformations of item difficulties (*b*-parameters) for Mathematics

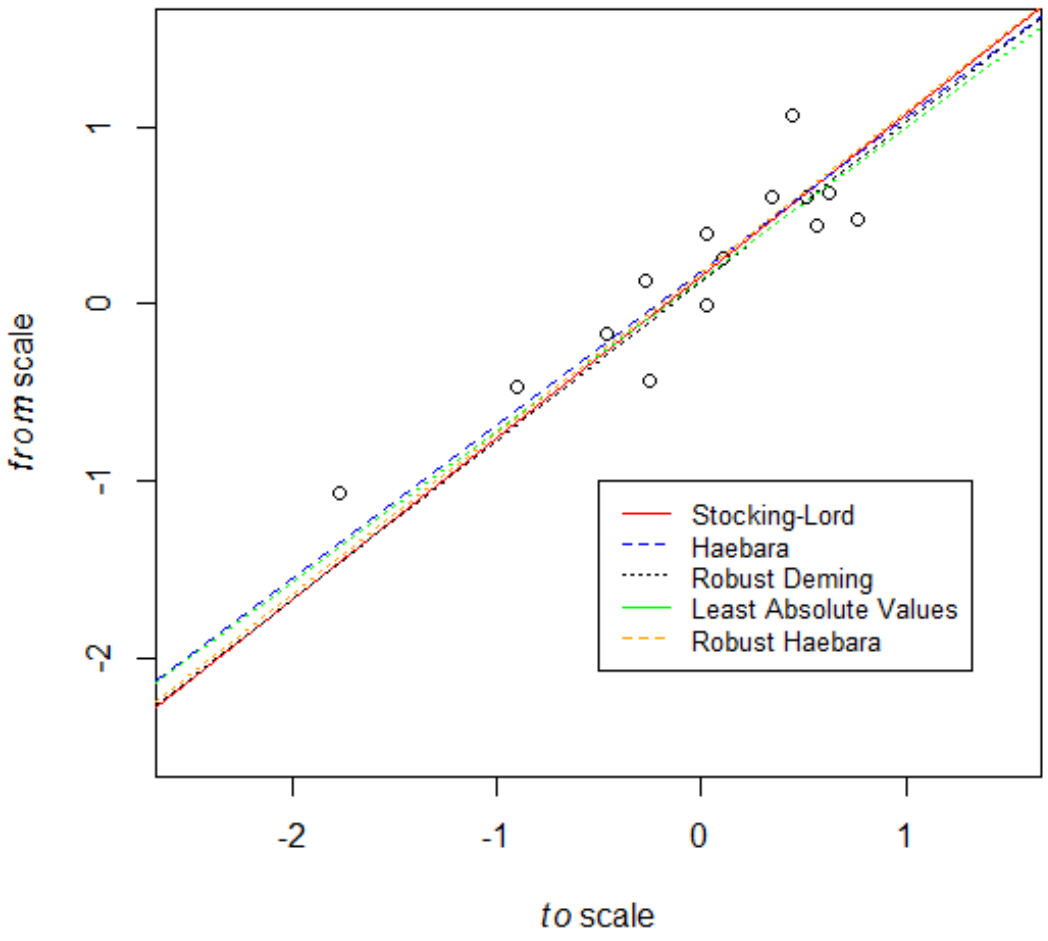


Figure 44. The transformations of item discrimination (*a*-parameters) for Mathematics

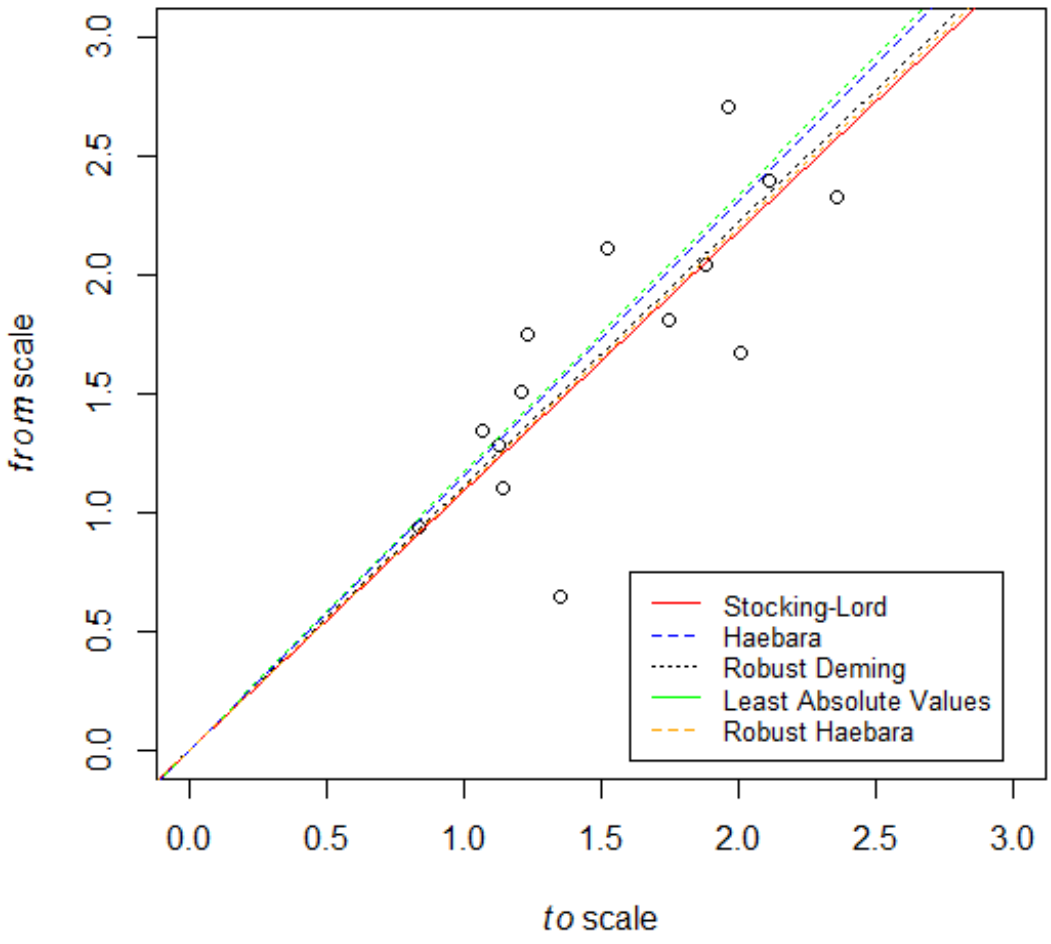
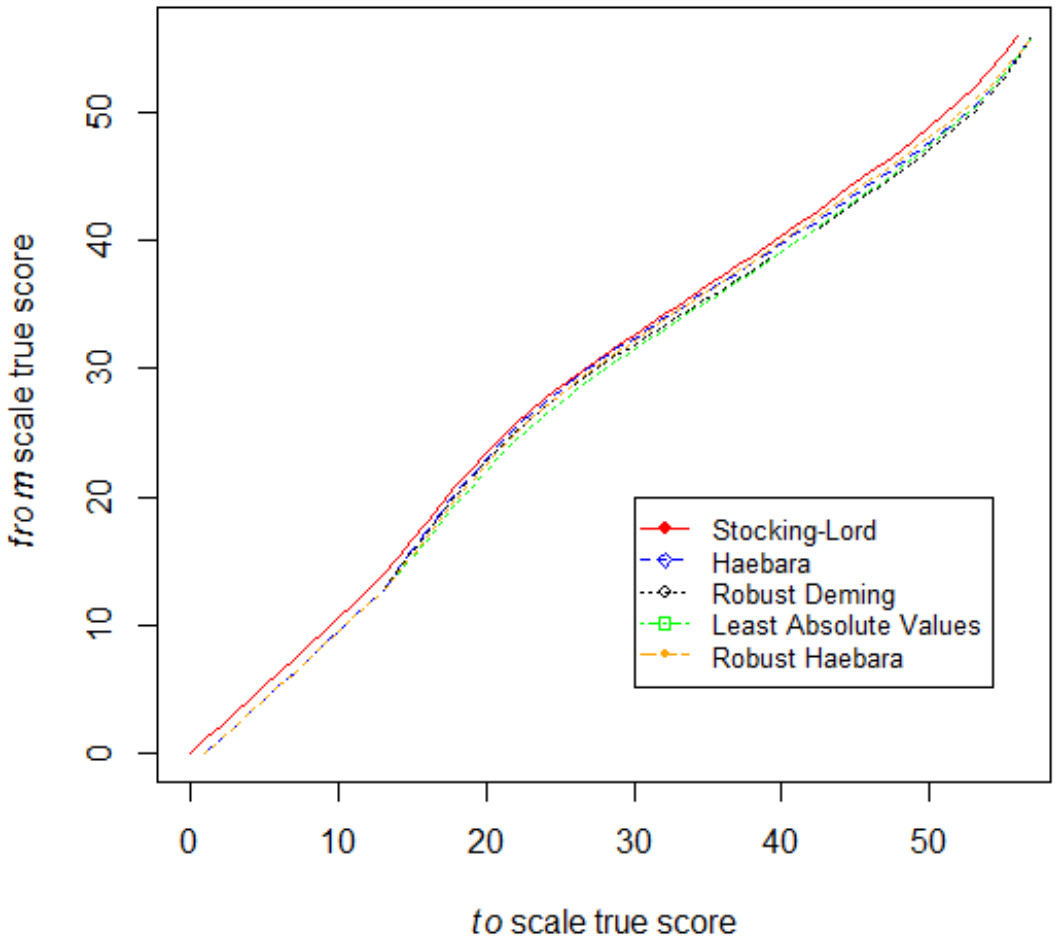


Figure 45. Estimated *to* scale true score equivalents of *from* scale true scores using IRT true score equating (Mathematics)



Similarly, the *b*- and *a*- parameters of the common items for the subject of Mathematics were plotted in Figures 43 and 44. The resulting equivalents are shown in Figure 45. The points of the *a*- and *b*- parameters of common items for the subject of mathematics were more scattered, and the lines obtained by the investigated methods are more separate. True scores obtained by using the Stocking-Lord method are generally higher than others at a given true score on the *to* scale, which is consistent with the results from the English test.

A problem of using real data is that the “true” scale transformation coefficients are unknown. As a consequence, there is no strongly supportive evidence for the comparison, and this study is inconclusive.

CHAPTER 5

DISCUSSION AND CONCLUSION

This chapter is organized with four sections. First, an overview of this study is provided. Second, major research findings are summarized. Third, discussions and future directions of the research are provided.

5.1 Overview of the Study

This research is aimed to address one of the most fundamental questions in the area of test equating: how to deal with outliers contained in the common item set when we conduct scale transformation and IRT equating? The importance of this question cannot be overemphasized because it is directly related to the accuracy of scoring a test.

The common-item nonequivalent groups design is commonly used in test equating. Under this design, the common item set plays a very important role in placing scores from different test forms onto a common scale. A rising issue was the inconsistency of item parameter estimates of common items from two test forms, which lead to distorted scale transformation and consequently inaccurate equating results (Cook and Eignor, 1991, Hu, Rogers, & Vukmirovic, 2008). Therefore, outlying common items should be carefully examined before equating is conducted.

Robust ways of dealing with outliers, based on the moment methods of scale transformation, have been proposed for decades, but they are not widely used. The widely used characteristic-curve based methods, i.e., Stocking-Lord and Haebara, are more robust to outliers than the abovementioned robust methods to some extent. However, the

impact of outliers still exists (He et al., in press), and it is explored more in this study by using complicated situations of outliers.

Existing methods have been focusing on detection and elimination of outlying common items, such as the delta-plot method (Angoff, 1972), the IRT parameter plots method (Kolen & Brennan, 2004), the displacement method (Murphy, Little, Fan, Lin, & Kirkpatrick, 2010), and the residual analysis method based upon ordinary least squares regression (He et al., in press). However, elimination of outlying common items might cause problems in IRT equating due to inadequate content representation of common items (Cook & Petersen, 1987).

To fill the gap in the literature on how to handle the outliers in the common item set, I introduced three new robust methods for scale transformation to reduce the influence of outlier on scale transformation and true score equating. The concept of M-estimators proposed by Huber (1977) is implemented in this study. To evaluate the performance of the new methods, I conducted a series of simulation studies. Finally, the methods are applied to the empirical data obtained from the *CBASE* test.

5.2 Summary of Findings

The findings for the series of simulation studies as well as applications for the *CBASE* data are summarized in the following sections.

5.2.1 Traditional Scale Transformation Methods with a Single Outlier

The simulation study was conducted to assess the traditional scale transformation methods in the presence of a single outlier in the common item set. As described in Sections 4.1.1, four scale transformation methods, i.e., Mean/Mean, Mean/Sigma, Stocking-Lord, and Haebara, were used to obtain scale transformation coefficients under nine simulated situations with or without a single outlier. In addition, three ability distributions of examinees were also compared. Further, indices of equating errors were compared among the investigated scale transformation methods.

First, the results with the simulation study suggest that the scale transformation methods based on the characteristic curves perform better than the moment methods when there is no outlier in the common item set. All the methods yield small and similar amount of bias, but the characteristic-curve methods produce less standard errors than the moment methods.

Second, a single outlier could result in larger equating errors. Particularly, the inclusion of a single outlier yields much larger bias than the condition of *No outlier*. In other words, the outlier dramatically enlarges the systematic errors in the test equating. In addition, the outlier also increases the random errors in test equating. The results are consistent with other studies (e.g., He et al., in press).

Third, it is evident that the scale transformation methods based on the characteristic curves perform better than the moment methods to some extent when a single outlier was simulated, as suggested by previous researchers (e.g., Stocking-Lord, 1983). The scale

transformation methods based on the characteristic curves have relatively smaller bias than the moment methods. The Stocking-Lord method performs better when the *a*-parameter increases, and the Haebara method performs better when the *a*-parameter decreases. However, the increments of equating errors are still remarkable even for the characteristic-curve based methods when $0.1 < |\Delta b| < 0.5$. Moreover, when the *b*-parameter moderately changed ($0.5 < |\Delta b| < 1$), the characteristic-curve methods also yield very large equating errors.

Having generally smaller equating errors, the Stocking-Lord method was chosen to represent the best traditional scale transformation method in the subsequent studies.

5.2.2 Performance of the Proposed Robust Methods of Scale Transformation

The study was conducted to investigate the performance of the proposed robust methods in scale transformation and score equating as compared to the Stocking-Lord method. As described in Sections 4.1.2, four scale transformation methods, Stocking-Lord and three proposed robust methods, namely, the robust Haebara method, Least Absolute Values or LAV method, and robust Deming method, were used to obtain scale transformation coefficients under nine simulated situations with or without a single outlier. Five ability distributions of examinees, representative of real situations, were considered in this study. Additional, indices of equating errors were compared among the investigated scale transformation methods.

The proposed methods are all based on the framework M-estimator (Huber, 1977), but each method has its specific design. The robust Haebara method is a modification of

the traditional Haebara method by introducing additional weights to individual common items. The weights are obtained by considering the area enclosed between the two item characteristic curves (Linn et al., 1981; Rudner, 1987). As a result, the new method is a modified weighted least squares method. The LAV method replaces the loss function of least squares in the traditional Haebara method by the loss function of least absolute deviations to reduce the influence of outliers on scale transformation. The robust Deming method is based on the framework of Deming regression where errors are included in the regression for both new and old test forms, but again with the M-estimates.

First, the results with the simulation study suggest that the robust Haebara method performs slightly worse than the Stocking Lord method under the condition of *No outlier* due to its enlarged equating error, particularly bias. However, the increase of bias is not large. Having smaller RMSE than the Stocking Lord method to some extent, the robust Haebara method produced the much smaller equating bias than the Stocking-Lord method when mild outlier ($0.1 < |\Delta b| < 0.5$) was simulated. In addition, the robust Haebara method produced much smaller RMSE, bias, and standard error than the Stocking-Lord method when moderate outlier ($0.5 < |\Delta b| < 1$) was simulated.

Second, despite that the LAV method has slightly larger RMSEs than the Stocking Lord method mainly due to its enlarged standard error, the LAV method has the least bias amongst the investigated methods under the condition of *No outlier*. Interestingly, the LAV often has the least standard errors in addition to decreased bias when a single outlier is simulated. In addition, the difference of equating between the LAV method and the robust Haebara method is rather small.

Third, the robust Deming method consistently yields larger RMSE, bias, and standard error of equating than the Stocking-Lord method in the absence of outlier. Despite having smaller bias than the Stocking Lord method, the robust Deming generally produces larger standard errors and consequently RMSE when in the presence of a single mild outlier ($0.1 < |\Delta b| < 0.5$).

In sum, the proposed robust Haebara method and the LAV method, although have different features, work well under different outlier conditions. The two methods are used in further investigations.

5.2.3 Comparisons between Robust Methods of Scale Transformation and Outlier Removal

The study was conducted to investigate the performance of the selected robust methods, the LAV method and the robust Haebara method, in scale transformation and score equating as compared to an existing outlier removal method, the displacement method (Murphy et al., 2010). To represent "perfect" identification and elimination, exclusion method is used to immediately eliminate the common item after simulation. As described in Sections 4.1.3, five scenarios of outlier handling, *Baseline* (neither outlier identification nor elimination by using the Stocking-Lord method), the robust Haebara method, the LAV method, the displacement method, the exclusion method, were used to obtain scale transformation coefficients under nine simulated situations with or without a single outlier. In this study, instead of using mild conditions of outlier, a single outlying common item under more severe conditions ($1 < |\Delta b| < 1.5$) is simulate. Five ability distributions of examinees were used in this study.

First, the results suggest that outlier elimination leads to remarkably smaller equating errors no matter which outlier elimination method is used. Specifically, the outlier elimination approaches substantially reduced the equating bias caused by the outlier. The displacement method generally has larger equating errors than the exclusion method particularly when severe *b*-parameter changes occur, whereas it produces fairly smaller equating errors than the exclusion method when *b*-parameter changes are moderate.

Second, both robust methods had slightly larger equating RMSE, bias and standard error than the Stocking Lord method under the condition of *No outlier*. However, the increments are almost unnoticeable. Similar to conditions of moderate *b*-parameter changes, both robust methods have significantly smaller RMSE than the *Baseline* method. The LAV method usually has the least RMSE amongst the investigated methods, regardless of the direction of *a*- and *b*- parameter changes and the magnitude of the *b*-parameter change. In addition, it is worth noting that the robust methods generally have the least equating bias among the five scenarios of outlier handling: the LAV is the least with severe *b*-parameter changes ($1 < |\Delta b| < 1.5$) and the robust Haebara method is the least with moderate *b*-parameter changes ($0.5 < |\Delta b| < 1$). The robust Haebara method, however, has considerably larger standard errors than the outlier elimination methods and the LAV method.

5.2.4 Numeric Illustration using Empirical Data

This study is an application of the proposed robust methods in empirical data of subjects English and Mathematics from the *CBASE*. The results indicate that scale transformation coefficients *A* and *B* obtained by the least absolute values method were

usually smaller than the traditional characteristic-curve methods, whereas the robust Haebara method yielded larger scale transformation coefficients *A* and *B* than the other methods. In addition, equated scores based on the robust method are generally smaller than the traditional Stocking-Lord and Haebara methods. However, it is inconclusive due to lack of true values of scale transformation.

5.2.5 Conclusions

Implementing the proposed robust methods to various simulated situations, one can conclude that:

1. The proposed robust Haebara method and LAV method work nearly as well as the traditional Stocking-Lord method when there is no simulated outlier.
2. More importantly, the proposed robust methods consistently outperform the traditional scale transformation methods in the presence of a single outlier.
3. The proposed robust methods provided smaller equated scores than the traditional methods when implemented in the *CBASE* English and Mathematics. However, there is no conclusive outcome from this study due to lack of true scale transformation coefficients.

5.3 Discussions and Future Directions

It is evident that the new proposed LAV method and robust Haebara method perform well in scale transformation. However, only a predetermined tuning constant of Huber function is implemented in this study. As for the robust Haebara method, weights are obtained based on the area enclosed by the two item characteristic curves and two

boundaries of a normal ability distribution. Similar idea was implemented in an early study (Linn et al., 1981). In the specific study, an item is considered as an outlier if its area is larger than 0.2. However, the criterion is not applied in this study. Instead, a relative criterion based on the "standardized" area between two corresponding item characteristic curves is employed. It is well-known that a tuning constant plays a very important role in determining the extent of down-weighting. As explicit in the Huber function, the M-estimator tends to be the least absolute deviation (L_1) estimator if the tuning consistent is too small. In this study, performances of the LAV method and the robust Haebara method seem to be similar – they have nearly identical equating errors. One may attempt to assume that the tuning constant here is too small, but it is contradicting the fact that differences between the two robust methods exist across the analyses. Nevertheless, it is worthwhile exploring more possible tuning constants in the future.

The robust Deming method is not continually investigated in the series of studies due to its inconsistent performance as compared to the other proposed methods. Conceptually, it is attractive. The method reflects the symmetric property of score equating, which is also considered in traditional Haebara method and the robust Haebara method. By introducing errors to both variables, the scale transformation procedure is fitted to a Deming regression. In addition, the proposed robust Deming method also takes into account of both *a*- and *b*-parameters. Although the idea and findings are somewhat encouraging, it should be acknowledged that the method has several strong assumptions and limitations. The errors of both *a*- and *b*-parameters are normally distributed. Further, the corresponding variances of *a*-parameters on different test forms are set to equal for

the convenience of computation. The assumption is also applied to the *b*-parameters on different test forms. Moreover, it is assumed that the transformations of *a*- and *b*-parameters are independent of each other to obtain the proposed joint likelihood function. However, the assumption of independence of transformations might not be held because the coefficient *A* is involved in transformations of both *a*-parameters and *b*-parameters. Additionally, distribution of errors obtained by the *a*- or *b*-parameters from their corresponding true values is unknown in practice. As a consequence, a generalized likelihood function with correlations between *a*- and *b*-parameters needs to be considered in the future. In addition, the small number of common items included in a test might have influence on the estimation of scale transformation coefficients.

One implication from this study is that eliminating outliers from the common item set before conducting the IRT true score equating generally reduces error level of score equating. Statistically, there is no question of this operation. However, in practice, we should consider content representation of the common items before eliminating an outlier. The critical point can never be overemphasized. Further investigation would study the situation of eliminating all corresponding common items as compared to the new robust approaches. However, sometimes it is not practical to eliminate all corresponding common items due to 1) limited amount of common items, and 2) particularly when multiple outliers based upon a common content area are encountered. In addition, in the studies only one outlier is simulated. In practice, it might not be the case. As a result, performances of the proposed robust methods in handling multiple outliers should be investigated.

Moreover, several limitations in the studies should be addressed in the future:

1. This study is limited to the unidimensional IRT true score equating. In the future, multidimensional IRT true score equating could be studied.
2. Empirical data from *CBASE* were used for this study. However, the existence of outlying common items for the specific data was not clear. To generalize the conclusion, data from other testing programs should be used in the future
3. The study does not consider situations with varying test length, number of common items. In the future, studies could be done to investigate these situations.
4. The equating procedure is limited to two test forms. In the future, multiple test forms could be tested.

REFERENCES

Andersen, R. (2008). *Modern Methods For Robust Regression*. Thousand Oaks: SAGE Publications.

Angoff, W. (1971). Scales, norms, and equivalent scores. In R.L Thorndike (ED.), *Educational Measurement* (2nd ed., pp.508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Paper Presented at the Annual Meeting of the American Psychological Association, Honolulu.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). Baltimore: Johns Hopkins University Press.

Angoff, W. H. (1996). *Scales, norms, and equivalent scores*. Princeton, NJ: ETS.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28, 147-162.

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques (2nd Ed.). New York: Dekker.

Bejar, I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (College Board Report No. 81-8). Princeton NJ: Educational Testing Service. (ETS No.81-35)

Bock, R., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25, 275-285.

Braun, H.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland and D.B. Rubin (Eds.), *Test Equating* (pp. 9-49). New York: Academic.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice*, 10, 37–45.

Cook, L. L., Eignor, D. R., & Hutton, L. R. (1979). *Considerations in the application of latent trait theory to objectives-based criterion-referenced tests*. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, 25, 31-45.

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

De Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Doran, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3,* 3-17.

Draper, N. R. and K. Smith (1998). *Applied Regression Analysis* (3rd Ed.). New York: Wiley.

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

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

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

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

Hanson, B. A., & B´eguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26, 3–24.

Hanson, B., & Zeng, L. (2004a). *ST: A computer program for IRT scale Transformation* (Revised by Cui, Z.) [Computer program].

Hanson, B., & Zeng, L. (2004b). *PIE: A computer program for IRT equating* (Revised by Cui, Z.) [Computer program].

Harris, D. J. (1991). *Equating with nonrepresentative common item sets and nonequivalent groups*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Hartley, R., & Zisserman, A. (2003). Multiple View Geometry in Computer Vision (2nd Ed.). Cambridge, UK: Cambridge University Press.

He, Y., Cui, Z., Fang, Y., & Chen, H. (in press). Using a Linear Regression Method to Detect Outliers in IRT Common Item Equating. *Applied Psychological Measurement*.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement*. 32, 311-333.

Huang, C. Y., & Shyu, C. Y. (2003). *The impact of item parameter drift on equating*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1, 799-821.

Huber, P.J. (1977). Robust Statistical Procedures, *Regional Conference Series in Applied Mathematics No. 27*, Society for Industrial and Applied Mathematics, Philadelphia.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley & Sons.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement*, 25, 39-52.

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

Kim, S., & Lee, W. C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, 43, 53-76.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, 22, 197-206.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, 18, 1–11.

Kolen, M. J. (1988). An NCME Instructional Module on Traditional Equating Methodology. *Educational Measurement: Issues and Practice*, 7, 29-36

Kolen, M. J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational measurement*, 19, 279-293.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin 55-23). Princeton, NJ: Educational Testing Service.

Linn, R. L., Levine, M. V., Hasting, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comparison. *Applied Psychological Measurement*, 5, 159-173.

Livingston, S.A. (1996). Book review of Test Equating. *Journal of Educational measurement*, 30, 369-373.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*, 73-95.

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

Lord, F. M. (1982). Item response theory and equating - A technical summary. In P. W. Holland & D. B. Rubin (Eds.), *Testing Equating*. New York: Academic.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equating. *Applied Psychological Measurement*, 8, 452-461.

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

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

Marco, G. L. (1977). Item characteristic curve solutions to the three intractable testing problems. *Journal of Educational Measurement, 16*, 139-160.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). Robust Statistics – Theory and Methods. Chichester, England : J. Wiley, 2006

Murphy, S., Little, I., Fan, M., Lin, C. H., & Kirkpatrick, R. (2010). *The Impact of Different Anchor Stability Methods on Equating Results and Student Performance*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Denver.

Ogasawara, H. (2001a). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement*, 25, 373–383.

Ogasawara, H. (2001b). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, 25, 53-67.

Osterlind, S. J. (2009). Modern Measurement: Theory, Principles, and Applications of Mental Appraisal (2nd Ed.). Boston, MA: Allyn & Bacon.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning (Vol. 161)*. Thousand Oaks, CA: Sage.

Petersen, N.S., Cook, L.L., & Stocking M.L. (1983). IRT versus conventional equating method IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137-156.

Petersen, N., Kolen, M. & Hoover, H. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (pp. 221–262). New York: American Council on Education and Macmillan.

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

Rousseeuw, R. J. & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Rudner, L. M. (1977). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the annual meeting of the American Educational Research Association. New York.

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

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

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York, NY: Springer.

Way, W. D., & Wang, K. L. (1991). *A comparison of four logistic model equating methods*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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.

Zimowski,M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). *BILOG-MG3* [Computer software]. St. Paul, MN: Assessment Systems Corporation.

VITA

Yong He was born January 14, 1973, in Jilin, China. He received the following degrees: B.S in Soil Sciences and Plant Nutrition with high honors from Shenyang Agricultural University (1995); M.S. in Plant Nutrition from South China Agricultural University (2002); M.S. in Agronomy from the University of Missouri (2008); M.A. in Statistics from the University of Missouri (2013); and Ph.D. in Educational Psychology from the University of Missouri (2013). He completed a summer internship with ACT Inc. in Iowa City, IA, in 2011.