

MACINNES, JOSHUA, Ph.D. A Simulation Study to Investigate Optimal Equating Anchor Set Construction Practices under the NEAT Design (2018)  
Directed by Dr. Richard M. Luecht. 367 pp.

This study examines anchor set construction techniques in observed score test equating under the non-equivalent with anchor-test design. It differs from other studies in that it seeks to understand the interaction between the examinee abilities, test specifications, and anchor set properties and develop a set of construction guidelines for practitioners.

This simulation study includes achievement and certification testing scenarios, each with 48 total difficulty and discrimination alignment conditions for the overall test and anchor set. Six candidate ability distributional conditions represent situations where the alternative test form candidates are either more able, more homogeneous, or more able and more homogenous compared to the base form group. In this study all 576 test, anchor, and ability conditions are equated by two linear methods, Tucker and Levine Observed Score, and two nonlinear methods, Frequency Estimation and Equipercntile.

The results of this study identified three interactions that are important to consider when equating and are not impacted by anchor set design: 1) well aligned test forms and similar examinee groups, 2) similar examinee groups and off-target test forms, and 3) well aligned test forms and dissimilar examinee groups. The results also suggest that three conditions are important and are impacted by anchor set design: 1) off-target test forms and dissimilar examinee groups, 2) well aligned test forms and more homogenous examinee groups, and 3) off-target test forms and more homogeneous and dissimilar examinee groups.

The second objective of the study is to develop a set of construction techniques for practitioners to use when dissimilar examinee ability groups are expected. The results suggest that the Equipercentile and Levine equating methods produce the smallest amount of systematic and total equating error when examinee groups differ in ability, are more homogeneous, or differ in ability and are more homogeneous. Two specific anchor set construction techniques are recommended for use with the Equipercentile method: 1) a midi anchor set with increased discrimination or 2) a midi anchor set with increased difficulty and discrimination. The results suggest that the Levine method is the most flexible, particularly if a more homogeneous sample is expected. Specifically, two types of anchor sets are recommended for use with the Levine method when a more able and more homogeneous sample is expected: 1) an anchor set with increased difficulty for low discrimination tests and 2) a traditional mini anchor set for high discrimination tests.

A SIMULATION STUDY TO INVESTIGATE OPTIMAL EQUATING  
ANCHOR SET CONSTRUCTION PRACTICES  
UNDER THE NEAT DESIGN

by

Joshua MacInnes

A Dissertation Submitted to  
the Faculty of The Graduate School at  
The University of North Carolina at Greensboro  
in Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy

Greensboro  
2018

Approved by

---

Committee Chair

For my family and MHS students.

APPROVAL PAGE

This dissertation written by Joshua MacInnes has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair \_\_\_\_\_

Committee Members \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_  
Date of Acceptance by Committee

\_\_\_\_\_  
Date of Final Oral Examination

## ACKNOWLEDGEMENTS

I want to thank the faculty of the ERM department and my family and friends for supporting me throughout this journey. Completing a PhD would not have been possible if it weren't for the support of so many.

To my committee members, thank you for being great teachers and mentors. From my initial inquiry about entering the program until the dissertation defense, Dr. Penfield was always quick to answer a question, explain a process, or ease my concerns. Thank you Dr. Sunnassee for mentoring me during the master's portion of the program and for assisting me during the initial stages of my equating research. Thank you Dr. Willse for the programming assistance and for helping me think more deeply about my research questions. Dr. Luecht, thank you for being supportive throughout the dissertation process and helping me expand a small idea and into a much bigger one. I've always appreciated the vision you have for solving practical psychometric problems.

To my family and friends, thank you for the support. To my wife, I could not have done this without you. Thank you for sacrificing so much so that I could go back to school. To my mom, thank you for paving the way, introducing me to the field, and for encouraging me throughout the process. I cannot possibly acknowledge all that have contributed to my success, but know that all of you made earning a PhD possible.

## TABLE OF CONTENTS

	Page
LIST OF TABLES .....	ix
LIST OF FIGURES .....	x
CHAPTER	
I. INTRODUCTION .....	1
Challenges for Test Developers and Psychometricians .....	2
Problem Statement and Research Questions .....	5
Study Significance .....	7
II. REVIEW OF THE LITERATURE .....	9
Purpose of Equating .....	9
Equating Designs .....	9
Random Groups (RG) .....	9
Single Groups (SG) .....	10
Non-equivalent Groups Anchor Test (NEAT) .....	11
Equating Methods .....	11
Identity Equating .....	12
Mean Equating .....	12
Linear Equating .....	13
Equipercentile Equating .....	13
NEAT Equating Methods .....	14
NEAT Linear Observed Score Equating Methods .....	15
NEAT Chained Linear Equating Methods .....	18
NEAT Frequency Estimation Method .....	18
NEAT Chained Equipercentile Method .....	19
Equating Error Estimation .....	19
General Equating Assumptions .....	21
NEAT Equating Assumptions .....	22
Content Requirement .....	23
Statistical Requirement .....	23
Correlation .....	24
Reliability .....	24
Difficulty .....	25
Spread of Item Difficulty .....	26
Factors Related to Difficulty .....	33

Summary of the Literature .....	35
III. METHODOLOGY .....	38
Item Generation .....	38
Exam Form Generation.....	39
Anchor Set Generation.....	41
Examinee Score Generation.....	44
Forms Generated.....	47
Equating Methods and Error Indices .....	50
IV. RESULTS .....	53
Evaluating Equating Results .....	54
Evaluating Achievement Test Forms .....	55
Achievement Test Form Difficulty .....	55
Achievement Anchor Difficulty Alignment .....	58
Achievement Anchor Difficulty Standard Deviation .....	59
Achievement Anchor Discrimination .....	59
Achievement Ability Conditions .....	59
Achievement Condition Interactions .....	65
Form Differences and Examinee Ability Differences .....	65
Anchor Differences and Examinee Ability Differences .....	72
Form Differences, Anchor Differences, and Examinee Ability Differences.....	83
Achievement Summary.....	91
Ability Mean Differences and Similar Form Difficulty .....	91
Ability Mean Differences and Dissimilar Form Difficulty .....	91
Ability Mean and Standard Deviation Differences and Similar Form Difficulty.....	92
Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty .....	92
Anchor Set Construction Techniques with Ability Mean Differences and Similar Form Difficulty .....	93
Anchor Set Construction Techniques with Ability Mean Differences and Dissimilar Form Difficulty .....	93
Anchor Set Construction Techniques with Ability Mean and Standard Deviation Differences and Similar Form Difficulty.....	93

Anchor Set Construction Techniques with Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty .....	94
Evaluating Certification Test Forms .....	95
Certification Test Form Difficulty .....	95
Certification Anchor Difficulty Alignment .....	99
Certification Anchor Difficulty Standard Deviation.....	99
Certification Anchor Discrimination .....	100
Certification Ability Conditions .....	100
Achievement Condition Interactions .....	106
Form Differences and Examinee Ability Differences .....	106
Anchor Differences and Examinee Ability Differences .....	112
Form Differences, Anchor Differences, and Examinee Ability Differences.....	123
Certification Summary.....	131
Ability Mean Differences and Similar Form Difficulty .....	131
Ability Mean Differences and Dissimilar Form Difficulty.....	131
Ability Mean and Standard Deviation Differences and Similar Form Difficulty.....	131
Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty .....	132
Anchor Set Construction Techniques with Ability Mean Differences and Similar Form Difficulty .....	132
Anchor Set Construction Techniques with Ability Mean Differences and Dissimilar Form Difficulty .....	133
Anchor Set Construction Techniques with Ability Mean and Standard Deviation Differences and Similar Form Difficulty.....	133
Anchor Set Construction Techniques with Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty .....	134
V. CONCLUSIONS AND DISCUSSION .....	136
Design Considerations .....	136
Research Question 1 .....	137
Research Question 2 .....	140
Ability Mean Differences and Similar Form Difficulty .....	143
Ability Mean Differences and Dissimilar Form Difficulty.....	144
Ability Mean and Standard Deviation Differences and Similar Form Difficulty.....	144
Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty .....	145

A Note about Some of the Results .....	146
Limitations and Future Research .....	148
Sample Size.....	148
Examinee Homogeneity .....	148
Presmoothing Methods .....	149
Levine Method .....	149
True Score Methods.....	149
Content.....	150
Internal Anchor Sets .....	150
REFERENCES .....	151
APPENDIX A. ACHIEVEMENT BIAS RESULTS.....	156
APPENDIX B. ACHIEVEMENT RMSE RESULTS .....	205
APPENDIX C. CERTIFICATION BIAS RESULTS .....	254
APPENDIX D. CERTIFICATION RMSE RESULTS .....	303
APPENDIX E. ACHIEVEMENT TOTAL AND ANCHOR SCORE CORRELATIONS .....	352
APPENDIX F. CERTIFICATION TOTAL AND ANCHOR SCORE CORRELATIONS.....	360

## LIST OF TABLES

	Page
Table 2.1. Summary of Sinharay & Holland (2006a).....	27
Table 2.2. Summary of Sinharay & Holland (2006b, 2007).....	28
Table 2.3. Summary of Liu, Sinharay, Holland, Feigenbaum, & Curley (2009, 2011) .....	30
Table 2.4. Summary of Liu, Sinharay, Holland, Curley, & Feigenbaum (2011) .....	30
Table 2.5. Summary of Fitzpatrick & Skorupski (2016) .....	32
Table 2.6. Summary of Trierweiler, Lewis, & Smith (2016) .....	34
Table 3.1. Test Form Conditions .....	39
Table 3.2. External Anchor Set Conditions .....	42
Table 3.3. Ability Conditions.....	46
Table 3.4. Achievement Base and Alternative Forms .....	47
Table 3.5. Achievement Equating Anchor Sets .....	48
Table 3.6. Certification and Licensure Base and Alternative Forms .....	49
Table 3.7. Certification and Licensure Equating Anchor Sets.....	50
Table 5.1. Guidelines for Practitioners .....	142

## LIST OF FIGURES

	Page
Figure 4.1. Achievement Tests: Bias Results for All Equating Methods when Form Difficulty Differences were 0.00 and 0.50 for both Discrimination Conditions .....	57
Figure 4.2. Achievement Tests: RMSE Results for All Equating Methods when Form Difficulty Differences were 0.00 and 0.50 for both Discrimination Conditions .....	58
Figure 4.3. Achievement Tests: Bias Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 0.60 .....	61
Figure 4.4. Achievement Tests: Bias Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 1.00 .....	62
Figure 4.5. Achievement Tests: RMSE Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 0.60 .....	63
Figure 4.6. Achievement Tests: RMSE Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 1.00 .....	64
Figure 4.7. Achievement Tests: Bias Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 0.60 .....	67
Figure 4.8. Achievement Tests: Bias Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 1.00 .....	68
Figure 4.9. Achievement Tests: RMSE Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 0.60 .....	70
Figure 4.10. Achievement Tests: RMSE Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 1.00 .....	71

Figure 4.11. Achievement Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 0.60.....	73
Figure 4.12. Achievement Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 1.00.....	74
Figure 4.13. Achievement Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60.....	75
Figure 4.14. Achievement Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00.....	76
Figure 4.15. Achievement Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 0.60.....	79
Figure 4.16. Achievement Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 1.00.....	80
Figure 4.17. Achievement Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60.....	81
Figure 4.18. Achievement Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00.....	82

Figure 4.19. Achievement Tests: Bias Results for All Anchor Conditions when Ability Differences were 0.50 for Linear Equating Methods when the Mean Item Discrimination was 0.60.....	84
Figure 4.20. Achievement Tests: Bias Results for All Anchor Conditions when Ability Differences were 0.50 for Linear Equating Methods when the Mean Item Discrimination was 1.00.....	85
Figure 4.21. Achievement Tests: Bias Results for All Anchor Conditions when Ability Differences were 0.50 for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60.....	86
Figure 4.22. Achievement Tests: Bias Results for All Anchor Conditions when Ability Differences were 0.50 for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00.....	87
Figure 4.23. Achievement Tests: RMSE Results for All Anchor Conditions when Homogeneous Ability Differences were 0.50 for All Equating Methods when the Mean Item Discrimination was 0.60.....	89
Figure 4.24. Achievement Tests: RMSE Results for All Anchor Conditions when Homogeneous Ability Differences were 0.50 for All Equating Methods when the Mean Item Discrimination was 1.00.....	90
Figure 4.25. Certification Tests: Bias Results for All Equating Methods when Form Difficulty Differences were 0.00 and 0.50 for both Discrimination Conditions.....	97
Figure 4.26. Certification Tests: RMSE Results for All Equating Methods when Form Difficulty Differences were 0.00 and 0.50 for both Discrimination Conditions.....	98
Figure 4.27. Certification Tests: Bias Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 0.60.....	102

Figure 4.28. Certification Tests: Bias Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 1.00.....	103
Figure 4.29. Certification Tests: RMSE Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 0.60.....	104
Figure 4.30. Certification Tests: RMSE Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 1.00.....	105
Figure 4.31. Certification Tests: Bias Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 0.60.....	107
Figure 4.32. Certification Tests: Bias Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 1.00.....	108
Figure 4.33. Certification Tests: RMSE Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 0.60.....	110
Figure 4.34. Certification Tests: RMSE Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 1.00.....	111
Figure 4.35. Certification Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 0.60.....	114
Figure 4.36. Certification Tests: Bias Results for All Anchor Conditions and Selected Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 1.00.....	115
Figure 4.37. Certification Tests: Bias Results for All Anchor Conditions and Selected Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60.....	116

Figure 4.38. Certification Tests: Bias Results for All Anchor Conditions and Selected Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00.....	117
Figure 4.39. Certification Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 0.60.....	118
Figure 4.40. Certification Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 1.00.....	119
Figure 4.41. Certification Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60.....	120
Figure 4.42. Certification Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00.....	121
Figure 4.43. Certification Tests: Bias Results for All Anchor Conditions for All Ability Differences for the Levine Equating Method when the Mean Item Discrimination was 0.60 .....	125
Figure 4.44. Certification Tests: Bias Results for All Anchor Conditions for All Ability Differences for the Levine Equating Method when the Mean Item Discrimination was 1.00 .....	126
Figure 4.45. Certification Tests: Bias Results for All Anchor Conditions for All Ability Differences for the Equipercentile Equating Method when the Mean Item Discrimination was 0.60.....	127
Figure 4.46. Certification Tests: Bias Results for All Anchor Conditions for All Ability Differences for Equipercentile Equating Method when the Mean Item Discrimination was 1.00.....	128

Figure 4.47. Certification Tests: RMSE Results for All Anchor Conditions When Homogeneous Ability Differences were 0.50 for the Levine and Equipercentile Methods when the Mean Item Discrimination was 0.60 .....	129
Figure 4.48. Certification Tests: RMSE Results for All Anchor Conditions When Homogeneous Ability Differences were 0.50 for the Levine and Equipercentile Methods when the Mean Item Discrimination was 1.00.....	130

## **CHAPTER I**

### **INTRODUCTION**

Testing programs regularly develop and administer new forms of an examination to improve test security. When developing additional, or alternative, forms of an examination it is important to build them so that examinees are not advantaged or disadvantaged based on the exam form administered. In other words, alternative forms should be constructed so that they have the same content representation and statistical properties as the original form. By constructing forms to be nearly identical, or parallel, from a content and statistical standpoint, testing programs are able to conduct statistical equating to establish cut scores for the new forms.

Although creating parallel forms seems straightforward, a number of real-world challenges can arise from the characteristics of the examinee population. Equating decisions are dependent, at least to some degree, on the quantity and consistency of the examinee ability distribution from one sample to the next. For instance, the equating design, number of items available for creating new forms, and accuracy of the results are limited as a result of the examinee population.

Testing programs have a number of choices when it comes to delivering and administering a test. Tests like the SAT® and ACT® administer paper and pencil forms during specified dates throughout the year, also referred to as testing windows. By testing

in windows, test forms can be assigned to examinees in a random way via a process called spiraling. Under the spiraling method, parallel test forms are alternated when passed out to examinees by exam administrators. Therefore, it is reasonable to assume the groups taking each test form are randomly equivalent, and any differences between the performance of the two groups is assumed to be attributable only to differences in difficulty between the forms.

However, many testing programs have moved to electronic testing, on demand at testing centers, which has made it much more difficult to create randomly equivalent samples when assigning examinees to forms. Although equating with non-equivalent groups is possible, it limits the equating methods available to psychometricians. The issue is compounded if candidate abilities vary throughout the year due to the academic calendar or as a result of policy changes within an organization or governing body. These operational testing constraints present a number of challenges for test developers and psychometricians, particularly when the anchoring benefits of item response theory (IRT) are not available, and observed score equating with classical, sample dependent, statistics becomes necessary.

### **Challenges for Test Developers and Psychometricians**

In order to equate under a non-equivalent with anchor-test (NEAT) design, each exam form must share a set of common items, also referred to as an anchor test or anchor set. Traditionally, the set of anchor items has been required to be parallel to the overall test from both a content and statistical standpoint (Kolen & Brennan, 2014). The

requirement that the anchor set be parallel to the overall test suggests that the common items must be a miniature, or mini, version of the overall examination.

The content stipulation is relatively straightforward and requires that the anchor set be representative of the content areas included in the test specifications. Content representation has been well studied (Cook & Peterson, 1987; Klein & Jarjoura, 1985; Peterson, Marco, & Stewart, 1982) and will not be investigated by this study.

However, the statistical stipulations are less straightforward and place a number of constraints on test developers. For instance, statistical properties may differ from one content area to another, which makes it difficult for test developers to construct anchor sets that are miniature versions of the overall tests. This mini constraint is even more restrictive when test developers have limited item banks, which is often the case for relatively new testing programs or when candidate samples are small.

The ability to relax the mini constraint placed on equating anchor sets would benefit testing programs by increasing the number of items available for use in anchor tests. Therefore, researchers have studied anchor test design by investigating mean item difficulty (Cronbach & Warrington, 1952; Gulliksen, 1945), spread of item difficulty (Lord, 1952; Richardson, 1936; Sinharay & Holland, 2006a), reliability (Budescu, 1985; Dorans & Holland, 2000; Fitzpatrick, 2008; Kolen & Brennan, 2014; Moses & Kim, 2007; Lord, 1980), correlation (Angoff, 1971; Budescu, 1985; Petersen, Kolen, & Hoover, 1989; Sinharay & Holland, 2006a), and anchor length (Budescu, 1985). Generally, the requirement that the anchor test be a mini version of the overall test has

been upheld, although Sinharay and Holland (2006a) made a case for re-examining the traditional requirement.

Sinharay and Holland (2006a) set out to investigate relaxing the spread of item difficulty requirement for anchor sets. They investigated anchor sets with the same content representation and mean item difficulty as the overall exam, but relaxed the spread of item difficulty requirement for the common items. In their anchor set design, the spread of item difficulty was allowed to be lower than the spread of item difficulty on the overall test. They referred to the anchor set as a “midi” or “semi-midi,” depending on the degree of item difficulty spread allowed. Sinharay and Holland (2006a) showed that the correlation between the anchor set and the overall test was higher when using midi anchors compared to mini anchors, which was an initial indication that the traditional spread of item difficulty requirement may not be necessary.

Follow-up studies have demonstrated that midi anchor sets produce similar linear and non-linear equating results when compared to using mini anchors (Sinharay & Holland, 2006b, 2007). Studies have shown that midi anchor sets perform well compared to mini anchor sets in the middle of the distribution of scores (Fitzpatrick & Skorupski, 2016; Sinharay, Holland, Curley, & Feigenbaum, 2011), when examinee samples are similar (Fitzpatrick & Skorupski, 2016; Holland, Feigenbaum, & Curley, 2009, 2011; Sinharay, Holland, Curley, & Feigenbaum, 2011), and when large ability differences exist between examinee samples (Holland, Feigenbaum, & Curley, 2009, 2011; Sinharay, Holland, Curley, & Feigenbaum, 2011). The results are encouraging for test developers

who would benefit from the ability to relax the spread of item difficulty requirement for anchor tests.

However, the results of a study by Trierweiler, Lewis, and Smith (2016) suggest that there are more contributing factors to consider, beyond simply reducing the spread of item difficulty for the anchor set. Such factors include the mean item difficulty and item discrimination of the test and anchor set. The authors suggest “additional studies should also explore and try to identify ‘optimal’ structures of anchor tests given specific test and population characteristics” (Trierweiler, Lewis, and Smith, 2016, pg 517).

### **Problem Statement and Research Questions**

The motivation for this study is twofold: 1) to build on the work of Sinharay and Holland (2006a) and others who have examined relaxing the established rule that requires anchor sets to have the same psychometric properties as the overall examination; and 2) to establish a set of rules that operational test developers in achievement testing and certification and licensure testing can use to build new exam forms under a variety of conditions. This study differs from previous studies in that it seeks to understand the interaction between the examinee abilities, test specifications, and anchor set properties when equating under the NEAT design. This study implements some of the suggestions of Trierweiler, Lewis, and Smith (2016) related to establishing ideal equating conditions for building alternative test forms using mini and midi anchor sets.

The first research question asks: how do examinee ability distributional characteristics, test development specifications, and anchor set properties interact to impact total equating error (root mean squared error) and systematic equating error (bias)

when equating with linear (Tucker and Levine Observed Score) and nonlinear (Frequency Estimation and Equipercentile) equating methods under the NEAT design?

The purpose of the first research question is to examine how midi and mini anchor sets function in different conditions, such as in certification and licensure testing and achievement testing. In certification and licensure testing, the mean item difficulty is usually lower than the mean examinee ability, the exam is built with a cut score in mind, and the examinee ability distribution is somewhat homogeneous due to the educational and professional eligibility requirements of the credential. In achievement testing, the mean item difficulty and the mean examinee ability are usually well aligned, the tests are not designed to have a cut score, and the examinee ability distribution represents a broad cross-section of the general population at a specified academic level.

Specifically, the first research question examines how equating with linear and nonlinear methods function as: 1) the examinee distributional properties change between samples with respect to the ability mean and standard deviation; 2) the examination purpose, specifications, and alignment are altered with respect to mean item difficulty and mean item discrimination; and 3) the alignment between the test specifications and the anchor set specifications are varied with respect to mean item difficulty, spread of item difficulty, and mean item discrimination.

The second research question asks: with respect to the test purpose and specifications, can anchor set assembly rules be established for linear (Tucker and Levine Observed Score) and nonlinear (Frequency Estimation and Equipercentile) equating methods when differences in group characteristics are expected? The second research

question is a synthesis of the results and is intended to provide useful findings for practitioners. In operational psychometrics, group differences can be anticipated, at least to a degree. For instance, examinee samples might be cyclical and follow a traditional academic calendar where more able examinees test in the spring and less able examinees test in the fall. Also, policy changes by a governing body, institution, or testing program may result in fewer examinees from a specific demographic subgroup sitting for an examination. Establishing anchor construction rules for specific situations will benefit practitioners and build on the research of Trierweiler, Lewis, and Smith (2016) to establish ideal anchor construction practices.

### **Study Significance**

The motivation for this study is to build on the work of Sinharay and Holland (2006a) and others who have examined relaxing the established rules that require an equating anchor set to be a mini version of the overall test. The ability to build non-mini anchor tests could benefit test developers, particularly those who have relatively small item banks. Equating studies that have examined reducing the spread of item difficulty of anchor sets have shown promising results in the middle of the distribution of scores under a number of examinee ability distributional conditions. However, the body of research on midi anchor sets has not established rules for choosing between midi and mini anchors.

This research adds to the body of literature on equating anchor set construction by identifying ideal anchor set construction practices for use in specific equating scenarios under the NEAT equating design. Specifically, this study examines a number of anchor set conditions relative to difficulty, discrimination, and their alignment with the overall

test conditions while also varying the exam specifications and examinee ability distributions. The goal of the study is to help answer the call of Trierweiler, Lewis, and Smith (2016) to develop a better understanding of how the statistical properties of the chosen anchor test impact the equating results under a variety of testing conditions.

## **CHAPTER II**

### **REVIEW OF THE LITERATURE**

#### **Purpose of Equating**

It is common for testing programs to assemble more than one form of an examination, and although tests are typically developed to be of similar difficulty, one form may be more or less difficult than another form. As a result, testing programs need a way to ensure that scores are comparable from one form to the next. The statistical process used to convert scores on a new, or alternative, form of an examination to be equivalent to scores on the original, also called the base or reference form, of the same examination is known as equating (Angoff, 1971; Kolen & Brennan, 2014).

#### **Equating Designs**

There are three primary equating designs, with the name of each design referring to how examinees are assigned to a respective test form. The three commonly used equating designs include: random groups (RG), single group (SG), and non-equivalent groups with an anchor test (NEAT). This section provides a brief description of each equating design.

#### **Random Groups (RG)**

Under the RG design, which is also referred to as an equivalent groups design, two groups are randomly administered two different forms of a test. Since the groups are

assigned to forms randomly they are assumed to be equivalent. Therefore, any differences between the scores on the forms are assumed to be directly attributed to differences in difficulty of the forms.

The RG design works well when all examinees are tested during a single session. However, the RG design is not always practical. For instance, testing programs administer exams on different dates and in different locations, which makes it difficult to ensure random assignment of examinees to forms.

### **Single Groups (SG)**

There are two ways to administer exams in the SG design: 1) administer both forms of the exam to the same examinees, or 2) split the single group into two groups and use counterbalancing to administer the two exams. The first SG option, however, is typically not feasible due to potential issues with the order with which the exams are administered, particularly when examinee fatigue becomes an issue. Therefore, this section focuses on the SG with counterbalancing design.

In the SG with counterbalancing design the group is first split into two groups. One group is administered Form X first, and the other group is administered Form Y first. When the second form of the exam is administered, each group takes the other form. The SG with counterbalancing design is feasible when administering the same form of an exam is operationally possible, when administration order effects are not likely, and when the sample size is not large enough for the random groups design (Kolen & Brennan, 2014).

### **Non-equivalent Groups Anchor Test (NEAT)**

The final equating design is the NEAT design, also referred to as the common-item nonequivalent groups design. Given the limitations of the other methods, the NEAT design is a commonly used in practice. Under the NEAT design the groups are assumed to be non-equivalent. To address the issue of non-equivalent groups, each form includes a set of items that are common to both test forms.

The common items, also referred to as anchor items or equating items, can be internal or external to the test (Kolen & Brennan, 2014). If anchor items are internal, the score on those items counts towards an examinee's exam score. External items do not count towards an examinee's score, but rather are used only for equating purposes.

### **Equating Methods**

A number of methods are available when equating alternative forms of an examination. Observed score equating and item response theory (IRT) equating comprise the two overarching categories, and each category includes linear and non-linear equating methods.

Observed score equating methods are the focus of this study, and therefore are the focus of this section. The most common observed score equating methods are mean, linear, and Equipercntile equating. Each general method will be outlined in this section, along with identity equating which is often included in equating studies for comparative purposes.

## **Identity Equating**

The simplest form of equating is identity equating, which assumes that the two test forms are identical. Therefore, identity equating treats scores on each form as equivalent and no statistical equating takes place. Since other methods will have some amount of error associated with the results, identity equating is often included as a benchmark with which to compare the results of other equating methods.

## **Mean Equating**

Mean equating assumes that two test forms differ in difficulty by a constant value across the entire score scale. When equating scores using mean equating, the average score difference is calculated between the reference form and the new form, and that difference is applied across the entire scale. The equation used to perform mean equating is

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

where  $\mu(Y)$  and  $\mu(X)$  represent the mean of the reference form and the alternative form respectively. Solved for the reference form,  $y$ ,  $m_Y(x)$ , represents a score,  $x$ , on form X, transformed to the scale of form Y (Kolen & Brennan, 2014).

Mean equating is limited since the difficulty difference between the two forms is assumed to be the same for low, average, and high performers. Although the assumption may not be realistic, mean equating can be successfully applied in many situations with small samples, since only the mean of each score distribution is needed and tends to be a relative stable statistic.

## **Linear Equating**

Linear equating allows for differences in difficulty between test forms to vary across the score scale by using the mean and standard deviation of the scores on the test forms. The equation for linear equating is

$$l_Y(x) = y = \sigma(Y) \left[ \frac{x - \mu(X)}{\sigma(X)} \right] + \mu(Y),$$

where  $\mu(Y)$  and  $\mu(X)$  represent the mean of the reference form and new form respectively,  $\sigma(Y)$  and  $\sigma(X)$  represent the respective test score standard deviations, and  $l_Y(x)$  represents a score,  $x$ , on form X, transformed to the scale of form Y (Kolen & Brennan, 2014).

Linear equating is more flexible than mean equating since it allows for differences to vary across the scale. Linear equating simplifies to mean equating if the score standard deviations are equivalent on both forms. Although linear equating is considered to be superior to mean equating, it requires a larger sample due to the need to calculate both the mean and standard deviation of each score distribution.

## **Equipercentile Equating**

Equipercentile equating is a more general form of equating and uses observed score distributional properties to place the new form on the same scale as the reference form. Since Equipercentile equating is non-linear, it allows for one form to be more, or less, difficult across the entire score scale, which is an advantage of the Equipercentile method compared to mean and linear equating. The Equipercentile equating function,  $e_Y(x)$ , as described by Braun and Holland (1982) in Kolen and Brennan (2014) is

$$e_Y(x) = G^{-1}[F(x)],$$

where  $e_Y(x)$  is used to convert scores from form X to the scale of form Y,  $G$  is the cumulative distribution function of Y, and  $F$  is the cumulative distribution function of X.

The function  $e_Y(x)$  refers each test's score distributions within the same population of examinees. Since the functions are symmetric, the equation

$$e_X(y) = F^{-1}[G(y)]$$

is also true for converting the distribution of scores on form Y to the scale of form X.

Equipercentile methods are more general, and thus more flexible, compared with mean and linear methods. Equipercentile methods are appropriate when the difficulty relationship between the two exams goes beyond the first two moments. However, Equipercentile methods generally require larger samples compared with mean and linear methods, due to the need to eliminate as many gaps as possible in the score distribution for each test form.

### **NEAT Equating Methods**

The purpose of this section to describe specific forms of mean, linear, and Equipercentile equating methods used under the NEAT equating design, which is the focus of this study. When equating with nonequivalent groups it is necessary to include a set of common items, represented in notational form as V. For consistency, the notation of the reference form of an examination will be Y and the notation for the new form will be X when describing each equating method. Likewise, where subscripts are included the

number 1 represents the examinee population associated with form X and the population associated with form Y is represented with the number 2.

### **NEAT Linear Observed Score Equating Methods**

The Tucker and Levine linear observed score equating methods are similar approaches to linear equating. The Tucker and Levine methods both make use of the concept of the “synthetic” population introduced by Braun and Holland (1982), even though the concept of the synthetic population was developed after the Tucker (Gulliksen, 1950) and Levine (1955) methods. The synthetic population makes use of the concept that the two populations taking form X and Y can be pooled to form a single population, providing as much information as possible. Each group is weighted by its size, which provides a proportional amount of information to the linear equating equation.

Although the Tucker and Levine methods are similar, they have noticeable differences with regard to the classical test theory idea of the true score. The Tucker method considers only observed scores, and includes the following assumptions: 1) the same linear regression function is assumed for X on V, and Y on V, for both populations; and 2) the same conditional variance is assumed for X given V, and for Y given V, for both populations. The two assumptions allow for the calculation of the regression slopes ( $\gamma$ ) for each examinee population using the equations

$$\gamma_1 = \frac{\sigma_1(X, V)}{\sigma_1^2(V)}$$

and

$$\gamma_2 = \frac{\sigma_2(Y, V)}{\sigma_2^2(V)},$$

which are used to calculate the synthetic population means,  $\mu_s$ , for each population using

$$\mu_s(X) = \mu_1(X) - w_2\gamma_1[\mu_1(V) - \mu_2(V)]$$

and

$$\mu_s(Y) = \mu_2(Y) + w_1\gamma_2[\mu_1(V) - \mu_2(V)].$$

The final linear equation relating the new form and reference form under the Tucker method is

$$l_{Y_s}(x) = \left[ \frac{\sigma_s(Y)}{\sigma_s(X)} \right] [x - \mu_s(X)] + \mu_s(Y).$$

Like the Tucker method, the Levine observed score method makes a similar assumption about the regression of X and Y on V, but related to true scores,  $T$ . The Levine assumptions include: 1) X, Y, and V are assumed to all measure the same thing; 2) the same linear regression function is assumed for  $T_X$  on  $T_V$ , and  $T_Y$  on  $T_V$ , for both populations; and 3) the same measurement error variance is assumed for X and Y for

both populations. The assumptions allow for the calculation of the regression slopes for each examinee population using the equations

$$\gamma_1 = \frac{\sigma_1(X)\sqrt{\rho_1(X, X')}}{\sigma_1(V)\sqrt{\rho_1(V, V')}}$$

and

$$\gamma_2 = \frac{\sigma_2(V)\sqrt{\rho_2(Y, Y')}}{\sigma_2(V)\sqrt{\rho_2(V, V')}}$$

without assuming a classical congeneric model. Under a classical congeneric model the regression slopes for each population are calculated using the equations

$$\gamma_1 = \frac{\sigma_1^2(X) + \sigma_1(X, V)}{\sigma_1^2(V) + \sigma_1(X, V)}$$

and

$$\gamma_2 = \frac{\sigma_2^2(Y) + \sigma_2(Y, V)}{\sigma_2^2(V) + \sigma_2(Y, V)}$$

The final linear equation relating the forms under the Levine method is

$$l_{Y_s}(x) = \left[ \frac{\sigma_s(Y)}{\sigma_s(X)} \right] [x - \mu_s(X)] + \mu_s(Y).$$

### **NEAT Chained Linear Equating Methods**

Another NEAT linear equating method is the chained method. The chained method makes the transitive assumption that “if X is related to V, and V is related to Y, then X is related to Y” (Kolen & Brennan, 2014, pp. 121). The chained method does not use synthetic weighting, and therefore effectively ignores the population weights used in the Tucker and Levine equations.

### **NEAT Frequency Estimation Method**

Frequency Estimation is a form of Equipercntile equating that uses the distribution of scores for X, Y, and V to calculate percentile ranks from the conditional distributions of X, and Y, with V (Kolen & Brennan, 2014). The concept of the synthetic population is fundamental to the Frequency Estimation method since it is a weighted combination of the combined population distributions of the reference and alternative forms. The synthetic populations are estimated by the equations

$$f_s(x) = w_1f_1(x) + w_2f_2(x)$$

and

$$g_s(y) = w_1g_1(y) + w_2g_2(y),$$

where  $f_s(x)$  and  $g_s(y)$  are the population distributions for the form X and Y scores.

The Frequency Estimation method also assumes that the conditional distribution is the same in both populations for the total score given. The conditional distributions are described by the equations

$$f_1(x|v) = f_2(x|v)$$

and

$$g_1(y|v) = g_2(y|v).$$

Because of the assumption it is recommended that the Frequency Estimation method only be used if the two populations are similar (Kolen & Brennan, 2014).

### **NEAT Chained Equipercentile Method**

Chained Equipercentile (CE) equating follows the same overarching steps as the chained linear equating method. The steps include first converting the Equipercentile function from form X to V using the form X population. Then, the Equipercentile relationship between V and Y is established using the form Y population. Finally, the two functions previously presented allow scores on X to be converted to scores on Y.

### **Equating Error Estimation**

One way to compare the accuracy of equating methods in an equating study is to estimate the error associated with equating over multiple replications. One form of equating error is random error, which is associated with issues of sampling (Kolen & Brennan, 2014). Random error exists due to the inability to include the entire population in an equating study, and therefore the sample of examinees does not exactly represent the population.

The standard error of equating (SEE) is a way of estimating random error in an equating study. The SEE is the square root of the error variance, and is calculated using the equation

$$SEE = \sqrt{\left[\frac{1}{R}\right] \sum var[\widehat{eq}_Y(x_i)]},$$

where  $\widehat{eq}_Y(x_i)$  is the sampling equating function at score  $x$  and replication  $i$  and  $R$  is the number of replications (Kolen & Brennan, 2014).

Another source of equating error is consider to be systematic error. Systematic error can be introduced in a variety of ways, and is associated with choices made related to the equating methodology (Kolen & Brennan, 2014). A common measure of systematic error is bias, which is calculated using the equation:

$$Bias = \left[\frac{1}{R}\right] \sum \widehat{eq}_Y(x_i) - eq_Y(x_i),$$

where  $\widehat{eq}_Y(x_i)$  is the sampling equating function at score  $x$  at replication  $i$ ,  $eq_Y(x_i)$  is the criterion equating function at score  $x$  at replication  $i$ , and  $R$  is the number of replications.

Root mean squared error (RMSE) attempts to capture both random and systematic error, and is calculated as

$$RMSE = \sqrt{Bias^2 + SEE^2}.$$

RMSE is calculated by squaring and summing the SEE and bias indices at each score point and then taking the square root of the result.

## **General Equating Assumptions**

In order for a process to be considered equating the test forms must have specific properties. Several researchers have presented guidelines for test equating (Angoff, 1971; Dorans & Holland, 2000; Kolen, 1988; Kolen & Brennan, 2014; Livingston, 2014; Lord, 1980; Peterson, Kolen, & Hoover, 1989) with most providing similar guidance. The equating guidelines presented by Kolen and Brennan (2014) and Dorans and Holland (2000) are summaries of commonly accepted equating requirements.

Kolen and Brennan (2014) describe five properties that are required for a process to be considered equating: the equating transformation must be symmetric, the two tests must have the same content and statistical specifications, the test form presented to an examinee should be a matter of indifference, the observed score distributions of each exam should be approximately the same, and the equating relationship between the two forms should be group invariant. Likewise, Dorans and Holland (2000) present five similar requirements for equating: the test forms should measure the same construct, the reliability of each form should be approximately the same, the equating transformation must be an inverse relationship, it should be a matter of indifference which form an examinee receives, and the equating relationship between the two forms should be population invariant.

The same construct guidelines presented by Kolen and Brennan (2014) and Dorans and Holland (2000) imply that the two test forms must be built to the same content and statistical standards. This is an important distinction between equating and linking, which do not have the same requirements regarding construct. Likewise, the two

forms must demonstrate similar statistical and distributional properties, which ensures that the two tests have similar observed reliabilities and distributional properties.

The symmetry and equity properties described by Kolen and Brennan (2014) and Dorans and Holland (2000) were originally presented by Lord (1980). The symmetry guideline requires the equating relationship between the two forms to be the same. The equity property states that it should be a matter of indifference to the examinee which form he or she is presented. These properties ensure that scores have the same meaning for examinees and addresses issues of fairness.

The invariance property presented by Kolen and Brennan (2014) and Dorans and Holland (2000) also addresses issues of fairness, from a group and subgroup standpoint. The property requires the equating relationship to be the same, regardless of the group or population used to determine the relationship. For the equating results to be generalizable to other populations for which the tests have been designed, the group invariance property should hold.

### **NEAT Equating Assumptions**

As previously mentioned, under the NEAT equating design both test forms include a set of anchor items to account for group differences. Traditionally, the group of anchor items included on the exams have been assumed to be a miniature, or mini, version of the overall exams from both a content and statistical standpoint (Kolen & Brennan, 2014). The purpose of this section is to address literature regarding these two requirements of the anchor item set.

## **Content Requirement**

The requirement that the content of the common items be consistent with the overall test is justifiable from a content validity standpoint, as differences in content may indicate a different construct is being tested. Specifically, content requirements are important for criterion-referenced examinations, where the interpretation of scores is related to an established criterion and not based on the normative performance of a group (American Educational Research Association et al., 2014). The anchor set content requirement in equating and linking is also a consideration included in *The Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014).

Equating research has supported the validity assumptions regarding content representation. Studies by Cook and Peterson (1987), Klein and Jarjoura (1985), and Peterson, Marco, and Stewart (1982) also concluded that anchor tests should be representative of the overall test from a content standpoint, especially if the groups differ in ability by more than a trivial amount.

## **Statistical Requirement**

The statistical requirement for the anchor items is more complicated, and less clear, than the content requirement. Traditional thinking has provided vague guidance about the anchor items from a statistical perspective, with most suggestions simply stating that the anchors should be statistically representative of the total test (Angoff, 1971; Dorans & Holland, 2000; Kolen, 1988; Kolen & Brennan, 2014; Livingston, 2014; Lord, 1980; Peterson, Kolen, & Hoover, 1989). Even *The Standards for Educational and*

*Psychological Testing* (American Educational Research Association et al., 2014)

provides little guidance, suggesting only that the anchor items should reflect the range of difficulty of the overall test.

Although the statistical requirement for anchor items has existed for many years, little evidence exists to support such a requirement (Sinharay & Holland, 2006a, 2006b, 2007). The evidence that does exist, however, largely focuses on considerations related to correlation, reliability, and item difficulty.

### **Correlation**

The higher the correlation between the anchor test and the forms being equated the better the anchor test will be for equating (Angoff, 1971; Budescu, 1985; Petersen, Kolen, & Hoover, 1989; Sinharay & Holland, 2006a). However, the requirement of the anchor set to be highly correlated with the test as a whole does not depend on it being a mini version of the overall test, as Sinharay and Holland (2006a) demonstrated.

### **Reliability**

To equate two test forms, the reference and alternative forms are expected to be parallel, or at a minimum tau equivalent, suggesting that the observed reliabilities of the test forms should be nearly identical (Budescu, 1985; Dorans & Holland, 2000; Kolen & Brennan, 2014; Moses & Kim, 2007; Lord, 1980). Therefore, it is implied that the anchor item set should also need to have a similar observed reliability compared to the total test. However, since anchor sets have fewer items it is understandable that an anchor set would have a lower observed reliability compared with the overall test.

Budescu (1985) used a mathematical proof to show that the magnitude of the correlation between the anchor set and overall test is dependent upon the reliability and length of the anchor set. Although the ideal anchor length is when the proportion of anchor items included on the test is 50%, a proportion that large may not be practical operationally. Budescu (1985) suggested that to produce the most efficiently reliable test, an anchor set should make up approximately 20-25% of the total items on the test.

Fitzpatrick's (2008) presidential address at the National Council on Measurement in Education acknowledged the same need for reliable anchor sets, suggesting that at least 15 anchor items should be included when equating test forms. Much like the need for a high correlation between the anchor set and the total test, the need for similar reliabilities for the anchor set and the overall test does not require the anchor set to be a mini version of the test.

### **Difficulty**

Item difficulty is well understood in terms of the overall test. Gulliksen (1945) provided a mathematical proof showing that the inter-item correlations are maximized when the average item difficulty is near 0.50 and the variance of item difficulties is small. Cronbach and Warrington (1952) found that a test with constant item difficulties would be ideal from a validity standpoint for setting cut scores and Lord (1952) found that minimizing the variability of item difficulty maximizes reliability and curvilinear correlation.

The full-test item difficulty concepts have also been applied to research on anchor sets. Based largely on the aforementioned research regarding difficulty for the whole test,

Sinharay and Holland (2006a) provided a mathematical proof showing that the correlation between the anchor set and the total test increases as the spread of item difficulty is reduced. The idea of reducing the spread of item difficulty on a shorter test to increase the correlation between a smaller test and a criterion is not new (Richardson, 1936). However, the idea contradicts the traditional requirement that an anchor set must be a mini version of the overall test from a spread of item difficulty standpoint.

### **Spread of Item Difficulty**

After providing the proof, Sinharay and Holland (2006a) conducted four simulation studies to examine the correlation between the total test and the anchor test. The authors explained the potential benefits of using an anchor item set with a smaller spread of item difficulty, called a “semi-midi” or “midi” anchor, compared to the total test. Midi anchors are more of a theoretical ideal where all of the anchor items have almost exactly the same difficulty as the overall test mean. However, semi-midi anchors have a spread of item difficulty half as large as the spread of item difficulty for the overall test. The results of the Sinharay and Holland (2006a) study are summarized in Table 2.1 and provide support for using anchor sets which have a smaller spread of item difficulty than the overall test to improve the correlation between the overall test and the anchor set.

Table 2.1. Summary of Sinharay & Holland (2006a)

Conditions Examined	Results
Study #1: Midi, semi-midi, and mini anchor sets	Study#1: The observed correlations from highest to lowest were: midi, semi-midi, mini.
Study #2: Content representative tests	Study #2: The observed correlations from highest to lowest were: midi, semi-midi, mini.
Study #3: Targeted and off-targeted anchor sets	Study #3: Semi-midi and midi anchor sets had higher correlations with the overall test than mini. Targeted mini anchor sets always had less average correlation than the off-targeted midi or semi-midi anchor sets.
Study #4: Selected items from operational test to construct semi-midi and mini anchor sets	Study #4: Higher correlations between the total test and the anchor set were found using the semi-midi anchor set compared to the mini anchor set.

In the first two studies Sinharay and Holland (2006a) found that midi anchor sets had the highest average correlations with the total test under all conditions included in the study. The results of the third study provided additional support, as midi anchor sets that were not well targeted had higher correlations with the total test than well targeted mini anchor sets. For the final study, which used operational test results, the semi-midi anchor set had a higher correlation with the total test compared to the mini anchor set. The results of the four studies by Sinharay and Holland (2006a) provided evidence that reducing the spread of item difficulty of the anchor set could potentially improve equating results.

Building off the work of Sinharay and Holland (2006a), a follow-up study was conducted by Sinharay and Holland (2006b, 2007) to investigate other equating

properties using midi, semi-midi, and mini anchor sets. The results of Sinharay and Holland (2006b, 2007) are summarized in Table 2.2.

Consistent with their initial study, Sinharay and Holland (2006b, 2007) found that midi and semi-midi anchor sets produced larger correlations with the total test compared to mini anchors. The first simulation study examined samples of 100, 500, and 5000; test lengths of 45, 60, and 78; and post-stratification Equipercentile (PSE) and CE equating methods. The study indicated that equating results using semi-midi and midi anchors were less biased, had smaller standard deviation (SD) with small samples, and smaller RMSE compared with equating results using mini anchors. The second and third parts of the study, which used simulated data and pseudo-real data, found mixed results and concluded that equating with semi-midi and mini anchor sets would yield similar results.

Table 2.2. Summary of Sinharay & Holland (2006b, 2007)

Conditions Examined	Results
<u>Study #1</u> : Examined sample size (100, 500, 5000), test length (45, 60, 78), and PSE and CE methods	<u>Study#1</u> : Equating with midi anchors yielded lower bias, SD, and RMSE compared with mini anchors. Also, error increased as the test length increased.
<u>Study #2</u> : Simulation from operational testing data with multiple content areas	<u>Study #2</u> : Mixed results, but the semi-midi and mini anchor set performance was similar.
<u>Study #3</u> : Pseudo-operational data with 120 item test with four content areas	<u>Study #3</u> : Little differences observed between the semi-midi and mini anchor set equating results with respect to bias and SEE. Mini anchors were slightly better for PSE equating and semi-midi anchors were slightly better for CE equating.

A number of studies have focused on replication of the results of Sinharay and Holland (2006b, 2007), with most studies using more realistic semi-midi anchor sets instead of theoretical midi anchor sets. As a result, recent research on midi anchor sets simplifies the semi-midi and midi anchor set distinction made by Sinharay and Holland (2006a, 2006b, 2007) to a single term, midi, which refers to an anchor set with half the spread of item difficulty as the overall examination.

Studies by Liu, Sinharay, Holland, Feigenbaum, and Curley (2009, 2011) and Liu, Sinharay, Holland, Curley, and Feigenbaum (2011) examined midi anchor sets on an operational SAT I exam. The exam included 78 operational items with 35 intact anchors. From the original form, two 20 item external anchor sets were constructed, one as a midi and the other as a mini, for comparison. Two chained equating methods, linear and Equipercentile, were used in the study as well as two poststratification equating methods, Tucker and Frequency Estimation. Like the earlier studies, bias, SEE, and RMSE were examined. However, in the studies examinee samples were divided into very similar, moderately similar, moderately dissimilar, and very dissimilar to assess the performance of midi and mini anchor sets under different equating conditions. Under most conditions the midi outperformed the mini for equating, and as the groups became less similar the midi anchor was preferred. The findings of Liu, Sinharay, Holland, Feigenbaum, and Curley (2009, 2011) are summarized in Table 2.3 and Liu, Sinharay, Holland, Curley, and Feigenbaum (2011) are summarized in Table 2.4.

Table 2.3. Summary of Liu, Sinharay, Holland, Feigenbaum, & Curley (2009, 2011)

Conditions Examined	Results
1) Groups that were very similar in ability	<u>Bias:</u> Equating with midi anchors tended to have less bias compared to mini anchors under most conditions. Both anchor types performed similarly when the groups were similar in ability, but as the groups were less similar the midi anchors produced less equating bias.
2) Groups that were moderately similar in ability	
3) Groups that were moderately dissimilar in ability	
4) Groups that were very dissimilar in ability	<u>SEE:</u> Anchor type had little impact on SEE.  <u>RMSE:</u> The results followed a similar pattern to the bias results.

Specifically, Sinharay, Holland, Feigenbaum, and Curley (2009, 2011) found that equating with similar ability groups produced similar results, with respect to bias, regardless of anchor type. However, when the groups were less similar, using midi anchor sets tended to perform better than mini anchors with respect to bias. The bias results were consistent across the score scale for all equating methods included in the study.

Although differences in the SEE results were negligible, the RMSE results followed a similar pattern to the bias results. When the two groups were of similar abilities, both anchor types produced similar RMSE results. As groups became less similar, equating with midi anchor sets reduced RMSE across the score scale for all equating methods.

Table 2.4. Summary of Liu, Sinharay, Holland, Curley, & Feigenbaum (2011)

Conditions Examined	Results
Group Differences	<u>Poststratification Equating:</u> When groups were similar in ability the midi and mini anchors performed similarly with respect to bias. When groups were less similar the midi anchor performed better than the mini anchors, especially near the middle of the distribution.
Bias and Weighted Absolute Bias	
Poststratification, Chained, and IRT equating methods	<u>Chained Equating:</u> When groups were similar in ability the midi and mini anchors performed similarly with respect to bias. When groups were less similar the midi and mini anchors perform better at different locations along the scale.  <u>IRT:</u> Mixed results were observed when comparing equating with midi and mini anchors with respect to bias.  <u>Weighted Absolute Bias:</u> Midi anchors performed better than mini anchors in most cases, although some mixed results were observed.

Liu, Sinharay, Holland, Curley, & Feigenbaum (2011) examined equating bias, group differences, and included PSE, chained, and IRT equating methods. With respect to group differences, the study supported the results of Sinharay, Holland, Feigenbaum, and Curley (2009, 2011) and found that the bias results supported using midi anchor sets under most conditions as group ability differences increased. In the discussion, the authors acknowledged that on middle difficulty items less similar groups should perform noticeably different, thus supporting the use of midi anchor sets over mini anchor sets.

The findings of Liu, Sinharay, Holland, Feigenbaum, & Curley (2011) were confirmed using IRT observed score equating methods of mean/mean, Stocking and

Lord, and concurrent calibration by Fitzpatrick and Skorupski (2016). A synopsis of the results of the Fitzpatrick and Skorupski (2016) study is provided in Table 2.5.

Table 2.5. Summary of Fitzpatrick & Skorupski (2016)

Conditions Examined	Results
Mean/Mean Equating	In the middle of the distribution of scores equating with midi and mini anchor sets produced similar results. In some cases
Stocking and Lord Equating	midi anchors performed better in the middle of the distribution of scores.
Concurrent Calibration Equating	In the extremes of the distribution of scores mini anchors produced better equating results than equating with midi anchors.
	Mini anchors were less informative because of the larger difficulty spread compared with the midi anchors.
	The midi anchors were more robust to large ability differences compared to the mini anchors.

Fitzpatrick and Skorupski (2016) found that anchor type had a larger impact on equating when group ability differences were more substantial, and equating results using midi anchor sets were more robust to group differences than using mini anchor sets. The study also found the anchor types produced small equating differences in the middle of the distribution of scores and larger differences at the extremes of the score scale. Although neither anchor type was clearly favored, the results supported relaxing the spread of item difficulty requirement when group differences are larger. The Fitzpatrick and Skorupski (2016) study also concluded that midi anchors may be better suited when tests are poorly targeted for the examinee population.

As a result of the encouraging findings of the studies examining midi anchor sets, Sinharay, Haberman, Holland, and Lewis (2012) presented a supplemental proof to provide theoretical evidence for constructing midi anchor sets instead of mini anchor sets in two specific situations: 1) when tests are of medium difficulty or 2) when the cut score of the examinations is near the center of the exam score scale.

### **Factors Related to Difficulty**

Although the prospect of improving equating results by including more anchor items close to the mean difficulty of the overall exam is encouraging, Trierweiler, Lewis, and Smith (2016) investigated other factors. Trierweiler, Lewis, and Smith (2016) began their study by replicating the conditions and correlational results of Sinharay and Holland (2006a). Trierweiler, Lewis, and Smith (2016) then turned their attention to manipulating two specific conditions: 1) the anchor item discrimination and 2) the difficulty location. The authors showed via a simulation study that using items with larger discriminations while varying the item difficulty ranges for the overall and anchor tests can yield correlational results which contrast the findings of Sinharay and Holland (2006a). A synopsis of the results of Trierweiler, Lewis, and Smith (2016) are provided in Table 2.6.

Specifically, Trierweiler, Lewis, and Smith (2016) extended the Sinharay and Holland (2006a) replication by increasing the discrimination values to fall between 0.9 and 2.4, which caused the reliability of each anchor set to change. With the shift in discrimination, the midi anchor set had the lowest correlation with the total test compared to the other anchor sets included in the study. By increasing item discrimination, the

anchor set with the highest correlation was an anchor set midway between the mini and semi-midi anchor sets in terms of spread of item difficulty.

Table 2.6. Summary of Trierweiler, Lewis, & Smith (2016)

Conditions Examined	Results
Replication of Sinharay and Holland (2006a)	The authors successfully replicated the results of Sinharay and Holland (2006a)
Increasing item discrimination	When item discrimination was increased, midi anchors produced the lowest correlation between the anchor set and the overall test compared to other anchor sets in the study.
Increasing item discrimination and manipulating difficulty location	For all discrimination and difficulty conditions examined, the highest correlation between the total test and anchor test was observed with a mini anchor set.  The two most important factors for increasing anchor and total test correlation are: 1) the anchor and total test true score correlation and 2) the anchor test reliability.  Authors suggest finding an ideal combination for anchor sets with respect to the test specifications and examinee population.

Trierweiler, Lewis, and Smith (2016) also altered the difficulty location for the total test and anchor set, which created an easier test. The second change caused the mini anchor set to have the highest correlation with the total test.

Additional discrimination and difficulty conditions were added to examine discrimination and difficulty of the total test and anchor set. In 21 of the 36 conditions the mini anchor set produced higher correlations with the total test compared to the midi anchor set. Therefore, reducing the variance of the item difficulty in the anchor set did not automatically increase the correlation with the total test.

Trierweiler, Lewis, and Smith (2016) caution against blindly using midi anchors over mini anchors when equating, and suggest that other factors must be considered. The authors showed that to improve the anchor to total test correlation the two most important factors are: 1) the anchor and total test true score correlation and 2) the anchor test reliability. Although Trierweiler, Lewis, and Smith (2016) recommend continuing the practice of using mini anchor sets to assemble most anchor sets, they acknowledge the need to identify optimal anchor test structures for use with specific characteristics of the test and examinee population.

### **Summary of the Literature**

When equating under a NEAT design, traditional thinking requires the common set of items to be a mini version of the overall test from both a content and statistical standpoint (Kolen & Brennan, 2014). The content requirement for the anchor test has been supported in *The Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) and in studies by Cook and Peterson (1987), Klein and Jarjoura (1985), and Peterson, Marco, and Stewart (1982).

However, Sinharay and Holland (2006a) argued that little evidence exists to support the statistical requirement, which they used to investigate relaxing the statistical rules for anchor sets. The authors were specifically interested in the ability to reduce the spread of item difficulty for anchor sets by using a midi anchor test. The ability to use more items located close to the mean item difficulty of the overall exam would be beneficial to test developers, particularly those with relatively small item banks.

A number of authors (Fitzpatrick & Skorupski, 2016; Sinharay & Holland, 2006a, 2006b, 2007; Sinharay, Holland, Curley, & Feigenbaum, 2011; Sinharay, Holland, Feigenbaum, & Curley, 2009, 2011) have compared equating results using midi anchor sets and mini anchor sets. Synthesizing the results, it appears that equating with midi anchor sets yields similar results to equating with mini anchors, particularly in the middle of the distribution of scores and when little differences in ability exists between the two examinee samples. When examinee groups are less similar in ability, midi anchors have been shown to improve equating results compared to using mini anchor sets.

A study by Trierweiler, Lewis, and Smith (2016) revealed that simply reducing the spread of item difficulty for the anchor set is not the only factor to consider. In their study the mean difficulty, of the overall test and anchor set, and mean item discrimination, were examined to provide evidence that other factors should be considered when building a case for, or against, specific requirements for an equating anchor set. The authors suggested that “additional studies should also explore and try to identify ‘optimal’ structures of anchor tests given specific test and population characteristics” (Trierweiler, Lewis, and Smith, 2016, pg 517).

The suggestion by Trierweiler, Lewis, and Smith (2016) is important. The majority of research that has compared equating results using midi and mini anchor sets has focused on achievement testing situations. In achievement testing, the mean item difficulties are typically aligned closely with the mean examinee abilities. However, in certification and licensure testing the mean item difficulty of the tests are often lower than the ability mean for the examinees. Likewise, the examinee ability distribution is

often more homogeneous in certification and licensure testing, since credentials require that examinees complete specific educational and professional requirements in order to gain eligibility to sit for an examination. Test developers in the certification and licensure testing industry may benefit the most from the ability to relax the traditional anchor set rules, particularly those with smaller item banks. This study seeks to examine the conditions under which anchor set construction rules for NEAT equating could be relaxed, in both achievement and certification and licensure testing.

## CHAPTER III

### METHODOLOGY

This chapter describes the methodology used to assess how statistical properties of an equating anchor set impact observed score equating results under the non-equivalent groups with anchor test (NEAT) design when characteristics of the examinee ability distributions and test specifications are varied. This chapter describes the simulation design, which includes variations of the following conditions: examinee ability differences, test purposes, test psychometric properties, anchor set statistical characteristics, and equating methodologies. The item, form, anchor, and examinee generation processes and statistical indices used to assess equating error are also outlined.

#### Item Generation

Item statistical properties were generated in R version 3.3.3 (R Core Team, 2017) using the catR package (Magis, 2011). The catR package has the ability to employ the item response theory (IRT) three parameter logistic model (3PL),

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta - b_i)]},$$

where  $\theta$  is the examinee's latent ability,  $D$  is a scaling constant,  $a_i$  is the item discrimination parameter,  $b_i$  is the item difficulty parameter, and  $c_i$  is the lower asymptote which has also been described as a pseudo-guessing parameter. The item

difficulty and discrimination parameters were varied during the study based on the test form and associated anchor set properties generated, while the pseudo-guessing parameter was constrained to fall between 0.00 and 0.25 for all conditions. The specific item difficulty and discrimination conditions are discussed when the form and anchor set conditions are presented.

### Exam Form Generation

This section presents the exam form conditions generated in the study. An important aspect of this study was to simulate conditions often seen in two types of testing situations: 1) achievement testing and 2) certification and licensure testing. An overview of the conditions that were generated in catR for each test type is provided in Table 3.1.

Table 3.1. Test Form Conditions

Condition	Achievement	Certification
Baseline Form Difficulty ( $\delta_b, \sigma_b$ )	$\sim N(0.00; 1.00)$	$\sim N(-0.64; 1.00)$
Number of Items	60	150
Test Mean Discrimination ( $\delta_a, \sigma_a$ )	(0.60, 1.00; 0.10)	(0.60, 1.00; 0.10)
Alternative Form Difficulty Differences ( $\Delta\delta_b = \mu_b(X) - \mu_b(Y)$ )	(0.00, 0.25, 0.50)	(0.00, 0.25, 0.50)

The two types of tests, which are described in Table 3.1 as achievement and certification, were designed to have respective mean IRT item difficulties of 0.00 for the achievement testing conditions and -0.64 for the certification testing conditions. The mean difficulty conditions for the base forms were chosen because of the typical alignment between mean item difficulty and mean examinee ability for each test type. Although discussed in detail later, the target mean IRT ability distribution was simulated to be 0.00 for examinees taking both test types. In achievement testing, item difficulty and examinee ability are usually closely aligned in well constructed tests, therefore a well targeted exam would include a mean IRT item difficulty of 0.00 to align with the mean ability of the target examinee population. Certification tests have cut scores which are often set where the pass rate is approximately 90% for the target examinee population. Therefore, a mean IRT item difficulty of -0.64 was chosen for the certification condition to align with a typical pass rate for the target population mean ability. Items with extreme IRT difficulty parameters, outside of a -3.00 to 3.00 range, were replaced through resampling so that all test forms included items with reasonable difficulty parameters that might be observed on operational test forms.

For each test type, the length of the tests were held constant at 60 items for achievement testing conditions and 150 items for certification testing conditions. The test lengths were chosen to represent a typical number of items included on operational tests of each type.

Item discrimination of the overall test impacts equating and should be considered when choosing items to include in an anchor set (Trierweiler, Lewis, & Smith, 2016).

Therefore, two item discrimination conditions were included for each test type: 1) tests with a relatively low mean IRT item discrimination of 0.60; and 2) tests with a relatively high mean IRT item discrimination of 1.00. For both discrimination conditions the standard deviation was held constant at 0.10. Any items generated with extreme IRT discrimination parameters, outside the range of 0.30 and 1.60, were replaced through resampling as those items would likely not be included on an operational assessment. The chosen range was similar to the low, 0.30 to 0.80, and medium, 0.60 to 1.60, discrimination ranges included in the Trierweiler, Lewis, and Smith (2016) study.

Testing programs have varying standards for building new test forms, because writing new items is expensive and the ability to field test new items depends upon the quantity of tests administered and the number of experimental item slots available on the test. Since examinee volumes and program budgets limit the degree to which new exam forms can be parallel to previous test forms, off-target test forms were included as a condition in the study. Base forms of the achievement and certification tests included mean IRT item difficulties of 0.00 and -0.64, respectively, and alternative test forms were generated to have mean IRT item difficulty differences of 0.00, 0.25, and 0.50 to assess equating with tests that are well-targeted, minimally off-targeted, and moderately off-targeted for each test's intended purpose. For all conditions, the IRT item difficulty standard deviation was held constant at 1.00.

### **Anchor Set Generation**

A primary motivation for this study was to investigate the traditional requirement that an equating anchor set must be a miniature, or mini, version of an overall test.

Specifically, the statistical properties of the anchor set were investigated by this study. For an anchor set to be considered a mini version of the overall test, the item difficulty and discrimination statistics of the anchor set must be representative of the overall test. In order to assess the mini requirement the mean item difficulty, item difficulty standard deviation, and mean item discrimination of the anchor test were manipulated to create mini and non-mini, or off-target, anchor sets with respect to the overall test. Each mini and off-target condition is presented in Table 3.2.

Table 3.2. External Anchor Set Conditions

Condition	
Mean Difficulty Alignment ( $\mu_b$ )	( $\delta_b, \delta_b+0.25$ )
Difficulty Standard Deviation ( $\sigma_b$ )	(0.50, 1.00)
Discrimination Alignment ( $\sigma_{Anchor}/\sigma_{Total}$ )	(1.00, 1.20)
Proportion of Items	25%

Mean item difficulty alignment is an important aspect of an anchor set. Two mean item difficulty conditions were included in the study: 1) an anchor set well aligned with the overall test form, with a mean IRT item difficulty difference of 0.00; and 2) an off-target anchor set not well aligned with the overall test form, with a mean IRT item difficulty difference of 0.25. The off-target condition was chosen as a practical difference that might be observed in operational testing, as it's unlikely that the difference between the mean IRT item difficulty of the overall test and the anchor set would be much larger

than 0.25. Differences of 0.25 and 0.50 were included in the Sinharay and Holland (2006a) study, although they acknowledged that 0.50 would be a substantial difference in difficulty.

The spread of item difficulty has been shown to impact equating results in midi anchor set research, which was first introduced by Sinharay and Holland (2006a). To be consistent with the reviewed literature, the spread of item difficulty conditions included: 1) an anchor set with a spread of IRT item difficulty identical to the overall test, with a standard deviation of 1.00; and 2) an anchor test with a spread of IRT item difficulty equal to half that of the overall test, or a standard deviation of 0.50.

In addition to item difficulty, item discrimination has been shown to be an important consideration when comparing midi and mini anchor item sets (Trierweiler, Lewis, & Smith, 2016). Two item discrimination conditions were included: 1) an anchor test with a mean item discrimination well-targeted for the overall test; and 2) an anchor test with a more discriminating set of anchor items compared with the overall test.

For the more discriminating anchor item condition, the ratio of the mean IRT item discrimination of the anchor set to the mean IRT item discrimination of the overall test was generated to be 1.20. The 1.20 ratio was chosen for two specific reasons: 1) it represents a sizeable increase of 20% from the overall test; and 2) it still stays within a reasonable range of item discrimination values that could be observed in operational testing. The study by Trierweiler, Lewis, and Smith (2016) used ranges instead of means for the discrimination parameters, and although their low, 0.30 to 0.80, and medium, 0.60 to 1.60, conditions were reasonable, their highest range, 0.90 to 2.40, would likely not be

observed in practice. Out of range anchor items were resampled to ensure that items with extreme IRT item discrimination parameters, those outside of 0.30 and 1.60, were replaced.

Finally, the proportion of external anchor items, 25%, was chosen to represent a practical number of anchor items that might be included on an operational test form. The seminal study on midi anchor sets by Sinharay and Holland (2006a) included 50% anchor items. However, including a large number of anchor items is neither necessary nor practical, as an anchor set consisting of 20-25% of the total items on the test has been shown to be the most efficient (Budesu, 1985). Follow-up studies by Sinharay and Holland (2006b, 2007); Liu, Sinharay, Holland, Feigenbaum, and Curley (2009, 2011); Sinharay, Holland, Curley, and Feigenbaum (2011) included fewer anchor items than Sinharay and Holland (2006a), which is further justification to use a more practical proportion of items for this study. Therefore, 60-item achievement tests generated in the study included 15 external anchor items and 150-item certification tests included 38 external anchor items.

### **Examinee Score Generation**

Differences between the examinee abilities from one sample to the next are important practical considerations when equating in operational testing, particularly when the changes might be anticipated due to the academic calendar, or a policy change within the testing organization or by a governing or accrediting body. The ability means and standard deviations of the examinee samples were manipulated to represent what might be observed in practice.

A major difference between achievement testing and certification testing are the ability distributional characteristics. In certification testing, examinee abilities are usually more homogeneous compared to examinee abilities in achievement testing. The homogeneity in certification testing results from educational and professional milestones that must be completed by candidates in order to be eligible to sit for the examination. In an achievement testing context, by contrast, examinees are often at the same point in their educational careers, but they represent a much broader spectrum of the general population compared to the specialized backgrounds of candidates seeking a certification or license to practice.

Therefore the baseline, or target, population for each test type was simulated to reflect what might be observed in practice. For achievement testing, the targeted examinee distribution was generated to have a mean IRT ability of 0.00 with a standard deviation of 1.00. For certification and licensure testing, the baseline examinee distribution was generated to have a mean IRT ability of 0.00 and a standard deviation of 0.50. Due to differences in standard deviation between the two testing scenarios, direct comparison of the results was not possible. However, the intention of the study was not to compare equating results in achievement testing to equating results in certification and licensure testing, but rather to examine each scenario independently.

To investigate changes in the ability distributions from one administration to the next, GENEQUATE 4.0 Software (Luecht, 2014) was used to generate dichotomous examinee response data for a number of ability distribution conditions. The ability distribution conditions are provided in Table 3.3. The IRT 3PL model item parameters

generated by the catR package were provided to GENEQUATE 4.0, which generated 2000 examinee responses for each ability condition and replication. The generated conditions included theoretical population IRT ability distributions with mean differences of 0.00, 0.25, and 0.50 from the baseline group.

Table 3.3. Ability Conditions

Condition	Achievement	Certification
IRT Ability ( $\mu_{\theta}; \sigma_{\theta}$ )	$\sim N(0.00; 1.00)$	$\sim N(0.00; 0.50)$
Ability Mean Differences ( $\Delta\mu_{\theta} = \mu_{\theta X} - \sigma_{\theta Y}$ )		(0, 0.25, 0.5)
Ability SD Ratio ( $\sigma_{\theta X} / \sigma_{\theta Y}$ )		(1.00, 0.50)
Sample Sizes		2,000

The ability standard deviation ratio between the baseline group of examinees and the examinees taking the alternative test forms was also manipulated. Specifically, a scenario of interest was equating with a more homogeneous examinee population from one administration to the next, such as what could result from the cyclical nature of the testing calendar or policy change that limits the population that is eligible to take an examination. The ability distribution conditions were similar to those used in the Fitzpatrick and Skorupski (2016) and Sinharay and Holland (2006b, 2007) studies. A total of 30 baseline group and alternative test form group pairs will be generated for each test and anchor combination to replicate each equating study 30 times.

## Forms Generated

Information detailing selected statistics for the forms generated in the study are provided in Tables 3.4 through 3.7. Table 3.4 provides information about the base and alternative forms generated under the achievement conditions included in the study and Table 3.5 includes the corresponding achievement equating anchor set information. Table 3.6 details information about the base and alternative forms generated under the certification and licensure conditions included in the study and Table 3.7 includes the corresponding certification and licensure equating anchor set information

Table 3.4. Achievement Base and Alternative Forms

	<u>Base Forms</u>		<u>Alternative Forms (30 Form Sets)</u>					
	Y1	Y2	X1	X2	X3	X4	X5	X6
	$\delta_a=.6$	$\delta_a=1.0$	$\Delta\delta_b=0$ $\delta_a=.6$	$\Delta\delta_b=0.25$ $\delta_a=.6$	$\Delta\delta_b=0.50$ $\delta_a=.6$	$\Delta\delta_b=0$ $\delta_a=1.0$	$\Delta\delta_b=0.25$ $\delta_a=1.0$	$\Delta\delta_b=0.50$ $\delta_a=1.0$
Mean a	0.61	1.00	0.60	0.60	0.60	1.00	1.00	1.00
SD a	0.09	0.09	0.10	0.10	0.10	0.10	0.10	0.10
Min a	0.36	0.82	0.31	0.31	0.31	0.69	0.69	0.70
Max a	0.78	1.23	0.90	0.92	0.92	1.31	1.35	1.35
Mean b	0.01	-0.01	0.00	0.25	0.50	0.00	0.25	0.50
SD b	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
Min b	-2.24	-2.10	-2.85	-2.56	-2.89	-2.80	-2.78	-2.62
Max b	2.89	2.87	2.98	2.82	2.97	2.96	2.97	3.00
Mean c	0.12	0.13	0.12	0.13	0.12	0.12	0.13	0.13
SD c	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Min c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max c	0.25	0.24	0.25	0.25	0.25	0.25	0.25	0.25

Table 3.5. Achievement Equating Anchor Sets

	EQ 1	EQ 2	EQ 3	EQ 4	EQ 5	EQ 6	EQ 7	EQ 8
Mean a	0.60	0.72	0.60	0.73	0.61	0.73	0.60	0.73
SD a	0.09	0.08	0.12	0.10	0.09	0.11	0.12	0.10
Min a	0.42	0.59	0.39	0.55	0.50	0.46	0.39	0.55
Max a	0.75	0.88	0.86	0.93	0.76	0.92	0.86	0.93
Mean b	-0.01	0.01	-0.01	0.00	0.26	0.26	0.24	0.25
SD b	1.00	0.99	0.49	0.50	1.01	0.99	0.49	0.50
Min b	-1.91	-2.23	-0.87	-0.95	-1.92	-1.32	-0.62	-0.70
Max b	1.62	1.83	0.92	0.84	1.74	2.23	1.17	1.09
Mean c	0.15	0.12	0.09	0.10	0.11	0.13	0.09	0.10
SD c	0.08	0.08	0.08	0.05	0.08	0.08	0.08	0.05
Min c	0.01	0.01	0.00	0.01	0.02	0.03	0.00	0.01
Max c	0.24	0.22	0.22	0.19	0.24	0.24	0.22	0.19
	EQ 9	EQ 10	EQ 11	EQ 12	EQ 13	EQ 14	EQ 15	EQ 16
Mean a	1.00	1.20	1.01	1.21	1.01	1.21	1.01	1.21
SD a	0.08	0.08	0.10	0.10	0.11	0.11	0.10	0.10
Min a	0.87	1.07	0.83	1.03	0.74	0.94	0.83	1.03
Max a	1.16	1.36	1.21	1.41	1.20	1.40	1.21	1.41
Mean b	0.01	0.01	0.00	0.00	0.26	0.26	0.25	0.25
SD b	0.99	0.99	0.50	0.50	0.99	0.99	0.50	0.50
Min b	-2.23	-2.23	-0.95	-0.95	-1.32	-1.32	-0.70	-0.70
Max b	1.83	1.83	0.84	0.84	2.23	2.23	1.09	1.09
Mean c	0.12	0.12	0.10	0.10	0.13	0.13	0.10	0.10
SD c	0.08	0.08	0.05	0.05	0.08	0.08	0.05	0.05
Min c	0.01	0.01	0.01	0.01	0.03	0.03	0.01	0.01
Max c	0.22	0.22	0.19	0.19	0.24	0.24	0.19	0.19

Table 3.6. Certification and Licensure Base and Alternative Forms

	<u>Base Forms</u>		<u>Alternative Forms (30 Form Sets)</u>					
	Y1 $\delta_a = .6$	Y2 $\delta_a = 1.0$	X1 $\Delta\delta_b = 0$ $\delta_a = .6$	X2 $\Delta\delta_b = 0.25$ $\delta_a = .6$	X3 $\Delta\delta_b = 0.50$ $\delta_a = .6$	X4 $\Delta\delta_b = 0$ $\delta_a = 1.0$	X5 $\Delta\delta_b = 0.25$ $\delta_a = 1.0$	X6 $\Delta\delta_b = 0.50$ $\delta_a = 1.0$
Mean a	0.61	1.00	0.60	0.60	0.60	1.00	1.00	1.00
SD a	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Min a	0.30	0.72	0.30	0.31	0.31	0.66	0.65	0.68
Max a	0.86	1.25	0.96	0.94	0.97	1.34	1.35	1.40
Mean b	-0.63	-0.63	-0.64	-0.39	-0.14	-0.64	-0.39	-0.14
SD b	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Min b	-2.56	-2.92	-3.00	-2.98	-2.99	-2.99	-2.99	-2.97
Max b	2.03	1.99	2.81	2.90	2.96	2.72	2.99	2.98
Mean c	0.14	0.12	0.12	0.12	0.13	0.12	0.12	0.13
SD c	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Min c	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max c	0.24	0.25	0.25	0.25	0.25	0.25	0.25	0.25

Table 3.7. Certification and Licensure Equating Anchor Sets

	EQ 1	EQ 2	EQ 3	EQ 4	EQ 5	EQ 6	EQ 7	EQ 8
Mean a	0.60	0.73	0.60	0.72	0.60	0.72	0.60	0.72
SD a	0.10	0.09	0.07	0.08	0.09	0.10	0.07	0.08
Min a	0.38	0.54	0.39	0.53	0.42	0.47	0.39	0.53
Max a	0.77	0.90	0.72	0.88	0.80	0.99	0.72	0.88
Mean b	-0.64	-0.64	-0.63	-0.63	-0.39	-0.40	-0.38	-0.38
SD b	0.99	1.01	0.51	0.51	1.00	0.99	0.51	0.51
Min b	-2.74	-2.61	-1.90	-1.84	-2.37	-2.10	-1.65	-1.59
Max b	1.08	1.46	0.45	0.45	1.80	2.09	0.70	0.70
Mean c	0.11	0.14	0.11	0.13	0.13	0.11	0.11	0.13
SD c	0.08	0.08	0.07	0.08	0.07	0.07	0.07	0.08
Min c	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Max c	0.25	0.25	0.25	0.25	0.25	0.24	0.25	0.25
	EQ 9	EQ 10	EQ 11	EQ 12	EQ 13	EQ 14	EQ 15	EQ 16
Mean a	1.01	1.21	1.00	1.20	1.00	1.20	1.00	1.20
SD a	0.09	0.09	0.08	0.08	0.10	0.10	0.08	0.08
Min a	0.82	1.02	0.81	1.01	0.75	0.95	0.81	1.01
Max a	1.18	1.38	1.16	1.36	1.27	1.47	1.16	1.36
Mean b	-0.64	-0.64	-0.63	-0.63	-0.40	-0.40	-0.38	-0.38
SD b	1.01	1.01	0.51	0.51	0.99	0.99	0.51	0.51
Min b	-2.61	-2.61	-1.84	-1.84	-2.10	-2.10	-1.59	-1.59
Max b	1.46	1.46	0.45	0.45	2.09	2.09	0.70	0.70
Mean c	0.14	0.14	0.13	0.13	0.11	0.11	0.13	0.13
SD c	0.08	0.08	0.08	0.08	0.07	0.07	0.08	0.08
Min c	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Max c	0.25	0.25	0.25	0.25	0.24	0.24	0.25	0.25

### Equating Methods and Error Indices

The equate R package version 2.0.6 (Albano, 2017) was used to perform the equating methodologies in the study. Specifically, the Tucker Observed Score, Levine Observed Score under a classical congeneric model, Frequency Estimation, and Equipercetile equating methods were calculated by the equate R package for the 30 replications included in the study. For the Frequency Estimation and Equipercetile

methods, loglinear presmoothing methods were implemented for the observed test-anchor score distribution. An attempt was made to use the same smoothing techniques as Sinharay and Holland (2006b, 2007), which preserved the first five univariate moments and a covariance moment. Although it was possible for the achievement tests, the covariance moment was not able to be preserved for smoothing the distributions of the certification tests. IRT methods were not included in the study, as the specific interest of the study was to examine equating using observed score methods and build on the research of Sinharay and Holland (2006a, 2006b, 2007) and Trierweiler, Lewis, and Smith (2016).

After the equating was completed, bias and root mean squared error (RMSE) were calculated. The methods for calculating bias and RMSE differed slightly from the bias and root mean squared error calculations presented in the last chapter, because the criterion was not an equating function, but rather output generated by GENEQUATE 4.0.

An advantage to using the GENEQUATE 4.0 software is that it provided an observed score for the alternative exam form taken by the simulated examinees, and a true score for the base form that the simulated examinees did not take. Therefore, the simulated true scores for the alternative form examinees on the base form were used as “truth” with which to compare the equating results. The methods used to calculate bias and total error with respect to “truth” were similar to the frame work described by Luecht and Ackerman (2018), and are outlined in the next few paragraphs.

Systematic error was estimated by calculating bias for each equated alternative form score,  $\widehat{e}q_Y(x_i)$ , on the base form scale,  $y_i^*$ . To calculate bias, each equated result

from the equate R package was compared to the true score output from GENEQUATE 4.0 using the equations

$$E_i = y_i^* - T_{Y_i},$$

and

$$Bias = \left[ \frac{1}{R} \right] \sum E_i | \theta,$$

where  $E_i$  is the difference between the equated base form score for each examinee,  $y_i^*$ , at replication  $i$ ,  $T_{Y_i}$  is the base form true score for each examinee at replication  $i$ ,  $R$  is the number of replications, and the bias calculation is conditional on the true theta (Kolen & Brennan, 2014). By calculating bias this way, the results were conditional on the true theta values and comparable across the distribution.

Total error was estimated by calculating RMSE using the equation:

$$RMSE = \sqrt{\left[ \frac{1}{R} \right] \sum E_i^2 | \theta},$$

where  $E_i$  is deviation of the equated base form score for each examinee,  $y_i^*$ , at replication  $i$ , consistent with the bias calculation, and conditional on the true theta. By calculating RMSE this way, the results are comparable to the bias results across the ability distribution.

## **CHAPTER IV**

### **RESULTS**

This chapter presents the results of this study and uses data visualization through the use of graphs to display equating error for the bias and root mean squared error (RMSE) results. Selected results are provided throughout the chapter and results for all conditions are provided in Appendices A through D. Appendices E and F include the mean correlation between the scored items and the external anchor items for each condition included in the study.

This study included two overarching testing situations, achievement testing and certification and licensure testing, which are presented separately in the chapter. Within each scenario, the following conditions were manipulated: alternative form mean difficulty, external anchor set construction, and ability mean and standard deviation. All conditions were manipulated within the context of tests with mean item discriminations of 0.60 and 1.00.

First, the bias and RMSE results of each condition are described in context of the two discrimination conditions, while holding all other conditions constant. By presenting them this way, each condition can be thought of as a main effect. Although relevant, interpretation of each main effect should be done with caution, as the interactions of each condition are the primary focus of the first research question: how do examinee ability distributional characteristics, test development specifications, and

anchor set properties interact to impact total equating error (RMSE) and systematic equating error (bias) when equating with linear (Tucker and Levine Observed Score) and nonlinear (Frequency Estimation and Equipercentile) equating methods under the NEAT design? The interactions are presented after the main effects to answer the first research question.

After the main effects and interactions are presented, trends are summarized for the purpose of answering the second research question, which reads: with respect to the test purpose and specifications, can anchor set assembly rules be established for linear (Tucker and Levine Observed Score) and nonlinear (Frequency Estimation and Equipercentile) equating methods when differences in group characteristics are expected?

### **Evaluating Equating Results**

As the results are presented it's important to consider practical differences. One way of evaluating equating error is to consider the concept of score difference that matters (DTM) (Dorans & Feigenbaum, 1994; Holland & Dorans, 2006). It's unlikely that two equating methods will produce identical results while a series of conditions are manipulated. However, the concept of DTM is to only consider differences in equated results that would typically lead to a different rounded integer score for examinees, and to overlook very small differences that would not lead to different rounded scores. For instance, DTM is often defined as an absolute value of 0.50, because a difference of 0.50 would usually lead to a different equated score. Therefore, DTM was defined as such for the purposes of reporting the results of this study.

## **Evaluating Achievement Test Forms**

Achievement tests, such as the SAT<sup>®</sup> and ACT<sup>®</sup>, are not designed with a cut score in mind. Therefore, when evaluating equating error results it's important to consider the entire score scale. The following sections review the equating error results for each condition included in the study for achievement tests. This section focuses on differences as they appear within each decile of the alternative form ability distribution.

### **Achievement Test Form Difficulty**

The base form achievement tests were generated to have a mean item response theory (IRT) difficulty of 0.00 and a standard deviation of 1.00. The alternative form conditions included difficulty differences of 0.00, 0.25, and 0.50 under the two overarching mean discrimination conditions 0.60 and 1.00. This section presents equating error results under the aforementioned conditions. Bias results are presented in Figure 4.1 and RMSE results in Figure 4.2.

When the two forms were identical in difficulty and the mean item discrimination was 0.60, all equating methods produced approximately the same magnitude of bias across the ability distribution. However, as the mean difficulty difference between the forms increased, the linear and non-linear methods produced two distinctly different patterns. The nonlinear equating methods produced a consistently positive bias across the ability distribution, with only substantial deviations at the extremes of the distribution. However, the linear equating methods produced a curve, with large positive bias in the tails of the ability distribution and slightly negative bias in the middle of the distribution.

Differences between the bias results as the mean difficulty differences increased were larger than the DTM threshold for the linear equating methods.

The results were similar under conditions where the discrimination was 1.00, although the nonlinear and linear methods produced slightly different results even when form difficulties were the same. When form differences increased, the nonlinear and linear equating methods followed the same pattern as when the discrimination was 0.60. However, the curve was more extreme for the linear methods, far surpassing the DTM threshold.

Similar RMSE results were observed for all four equating methods when the mean difficulty was the same for both forms and the mean discrimination was 0.60. However, differences between equating methods were observed as the alternative form became more difficult. The linear methods produced less RMSE at the low end of the ability distribution and the nonlinear methods produced lower RMSE at the upper end of the ability distribution. In the middle of the distribution, all equating methods produced similar RMSE.

When the mean discrimination was 1.00, similar results were found. However, differences between linear and nonlinear methods were slightly more pronounced in the tails of the ability distribution when form differences were large.

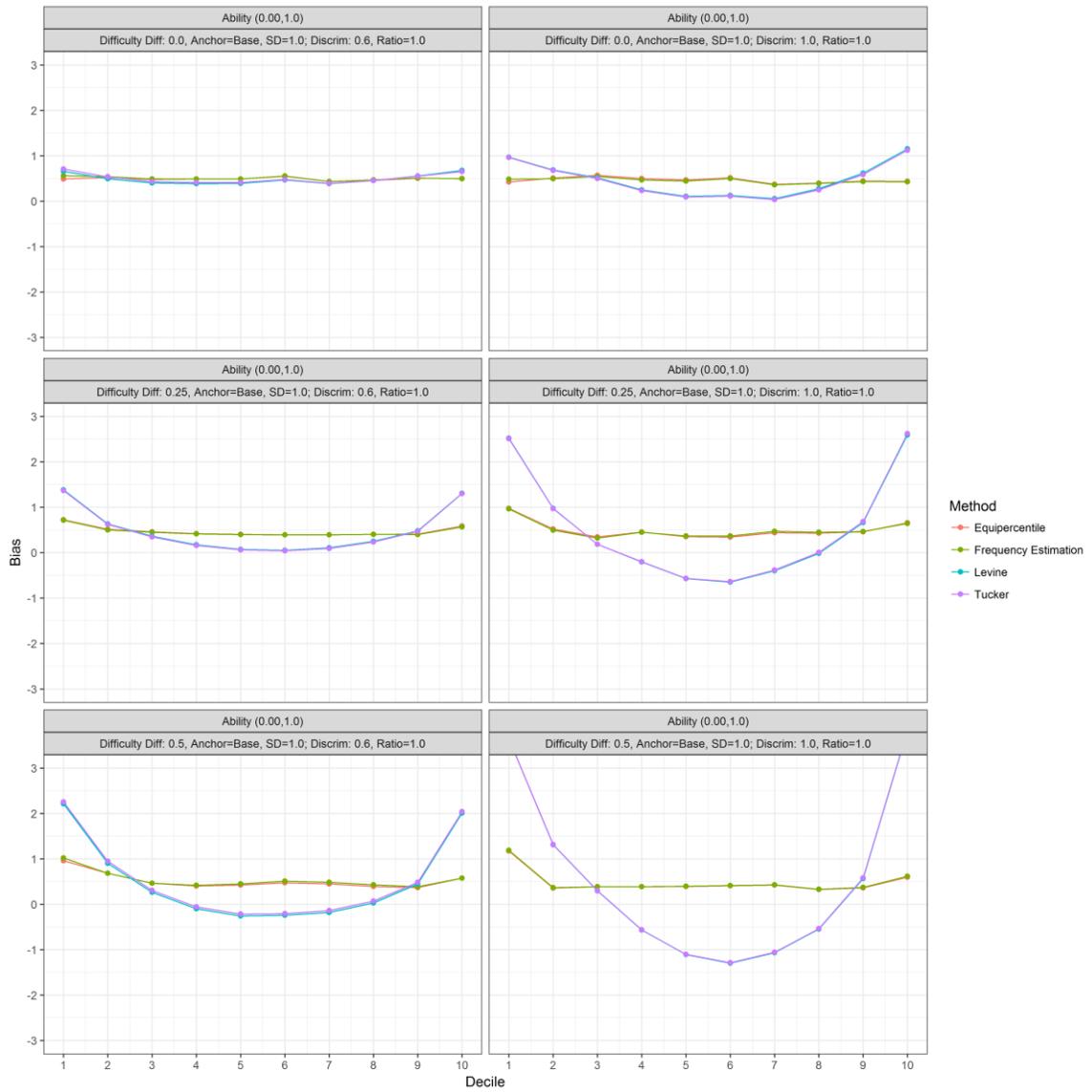


Figure 4.1. Achievement Tests: Bias Results for All Equating Methods when Form Difficulty Differences were 0.00 and 0.50 for both Discrimination Conditions

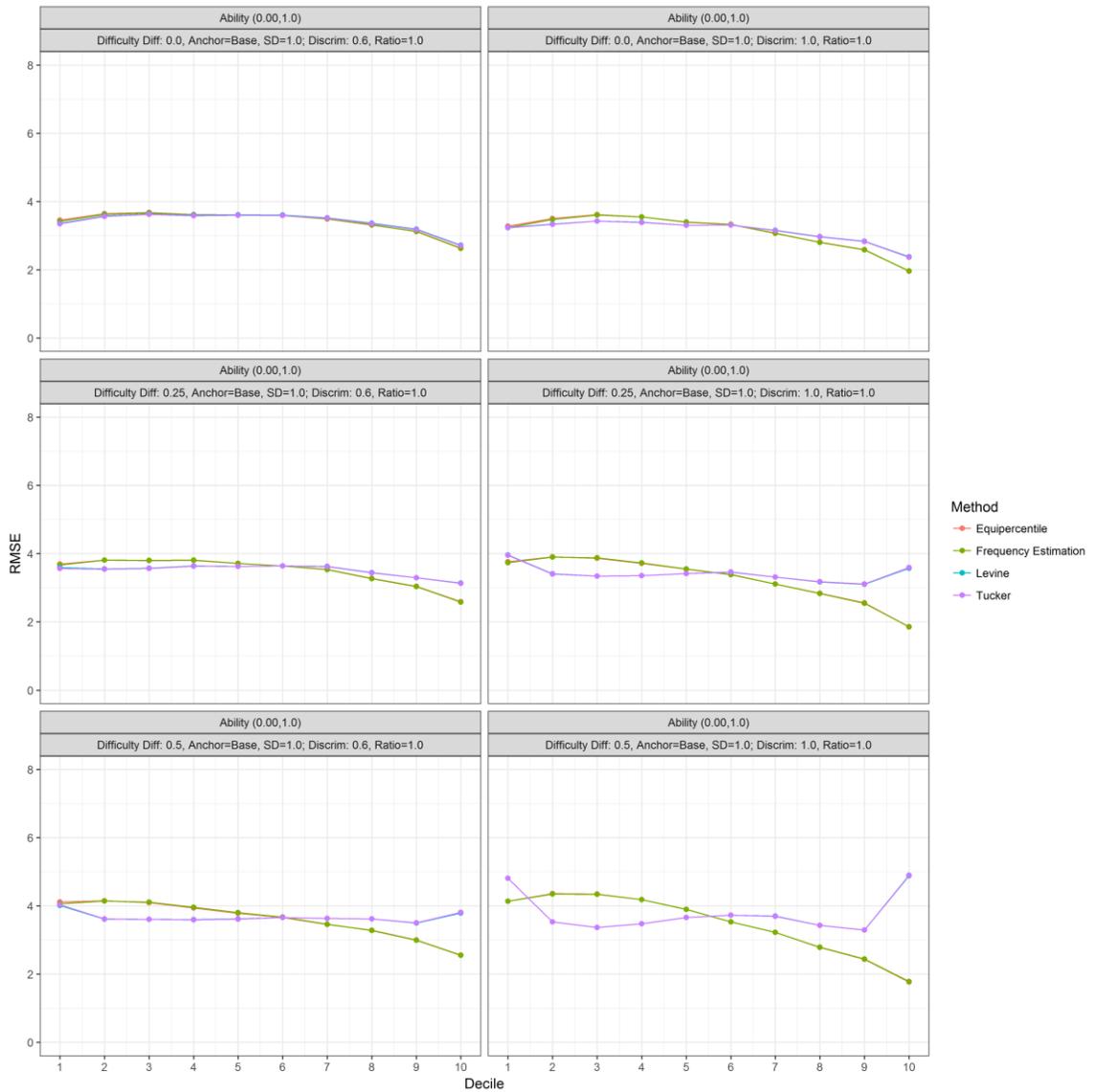


Figure 4.2. Achievement Tests: RMSE Results for All Equating Methods when Form Difficulty Differences were 0.00 and 0.50 for both Discrimination Conditions

### Achievement Anchor Difficulty Alignment

The mean difficulty of the anchor set was generated to have a mean IRT difficulty that was either: 1) in alignment with the base form, or 2) shifted to have a mean 0.25

greater than the base form. Changing the difficulty of the anchor set alone did not have a practical impact on equating error, when all other conditions were held constant.

### **Achievement Anchor Difficulty Standard Deviation**

Two standard deviation conditions were included for the difficulty of the anchor set: 1) a difficulty standard deviation of 1.00, and 2) a standard deviation of 0.50, referred to as a midi anchor set. With respect to bias and RMSE, while all other conditions were held constant, the results were similar regardless of the anchor set standard deviation.

### **Achievement Anchor Discrimination**

Two discrimination conditions were included in the study for anchor sets: 1) an anchor set with the same mean discrimination as the overall test, and 2) an anchor with increased discrimination compared to overall test, by 20%. Overall, the results were similar for both discrimination conditions for the achievement testing scenario, when other conditions were held constant.

### **Achievement Ability Conditions**

A specific interest of the study was to examine conditions where the alternative form ability group was both more able and more homogeneous. Therefore, the study included mean ability differences of 0.00, 0.25, and 0.50 and ability standard deviations of 1.00 and 0.50. Figures 4.3 and 4.4 include bias results for conditions where only the ability of the base and alternative form groups were manipulated. The figures present results for both mean discrimination conditions, 0.60 and 1.00, respectively.

When the base and alternative form groups had the same ability mean and standard deviation, all four equating methods produced essentially the same bias for the 0.60 discrimination condition. A few trends were observed when the mean ability of the alternative form was increased, which are displayed along the left column of Figure 4.3. For instance, the Levine equating method produced approximately the same bias regardless of the ability differences between the groups, the Tucker and Frequency Estimation methods produced negative bias as group differences increased, and the Equipercentile method bias results appeared to shift closer to zero as the groups became more different. It should be noted that the observed differences exceeded the DTM threshold for the Tucker and Frequency Estimation methods as group differences increased.

Similar results were observed when the mean discrimination was 1.00, although the linear methods produced a curve. The curve was more positive in the tails of the ability distribution and less positive, or negative, near the middle of the distribution. The curve was observed even when there were no ability differences between the groups.

When the standard deviation of the alternative form group was 0.50 the bias results were quite different, and are presented in the right columns of Figures 4.3 and 4.4. In all cases the Levine method produced the smallest, and most consistent bias results across the ability distribution and the Equipercentile method produced the second most stable bias results. The Tucker and Frequency Estimation methods produced extremely negative bias at the lower end of the ability distribution and large positive bias at the upper end of the distribution. All methods produced similar bias just above the center of

the distribution of abilities, which was shifted farther right as the group differences became larger.

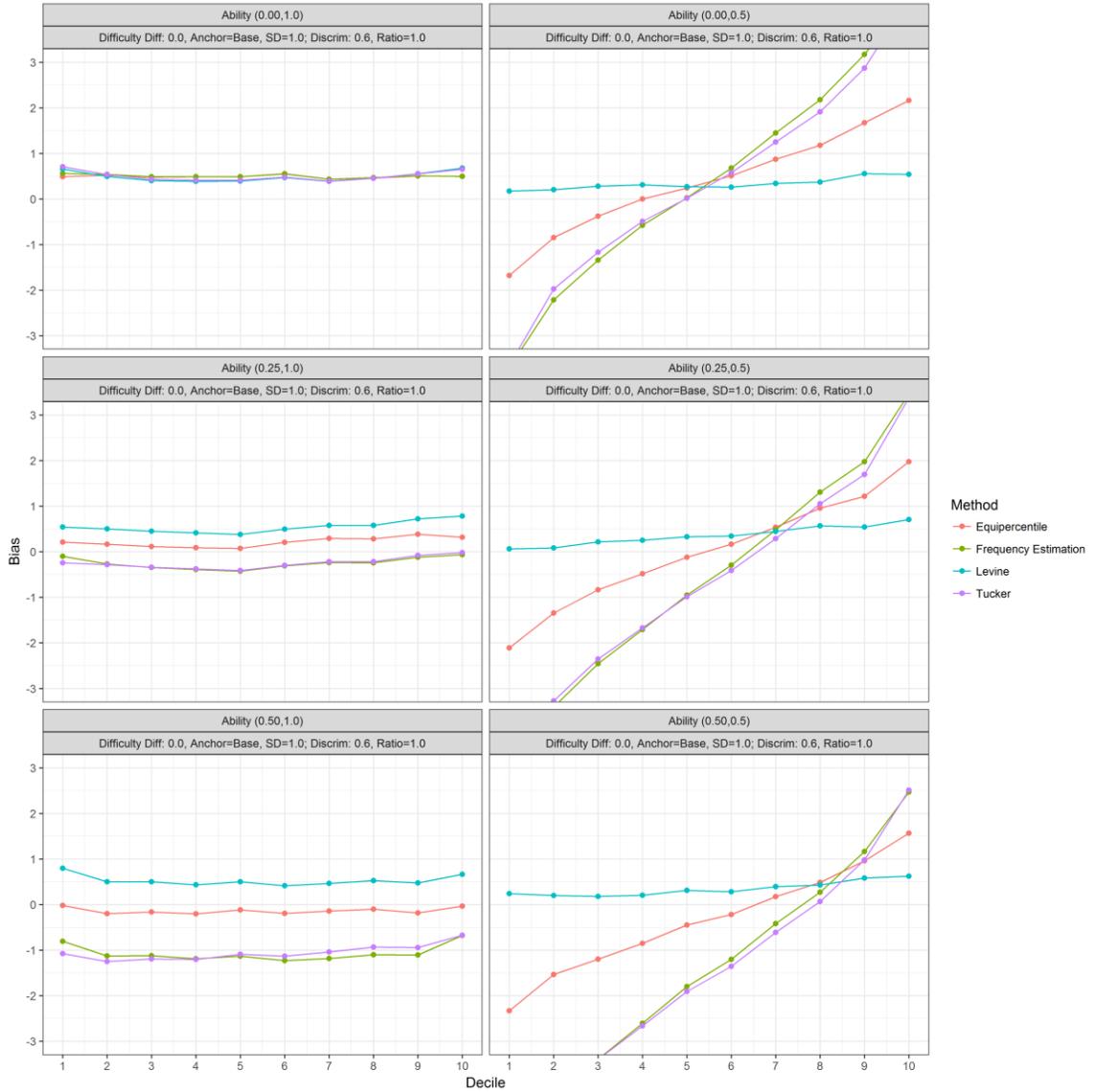


Figure 4.3. Achievement Tests: Bias Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 0.60

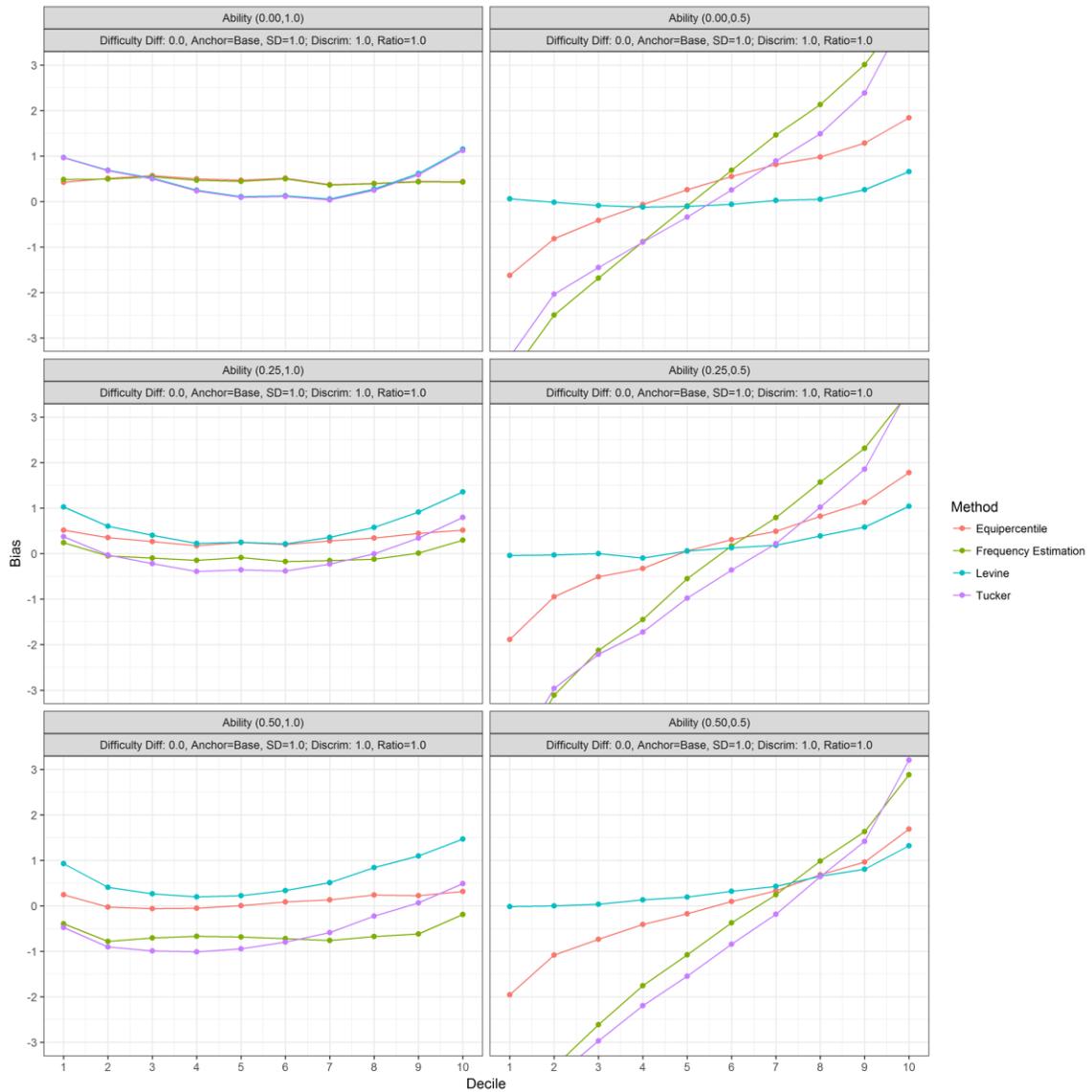


Figure 4.4. Achievement Tests: Bias Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 1.00

RMSE results are presented in Figures 4.5 and 4.6 for conditions where the mean item discrimination was 0.60 and 1.00, respectively. Generally, the RMSE results were less dramatic compared to the bias results. Minimal RMSE differences were observed as

the group differences increased and the ability standard deviation was held constant at 1.00, which are displayed along the left column of Figures 4.5 and 4.6.

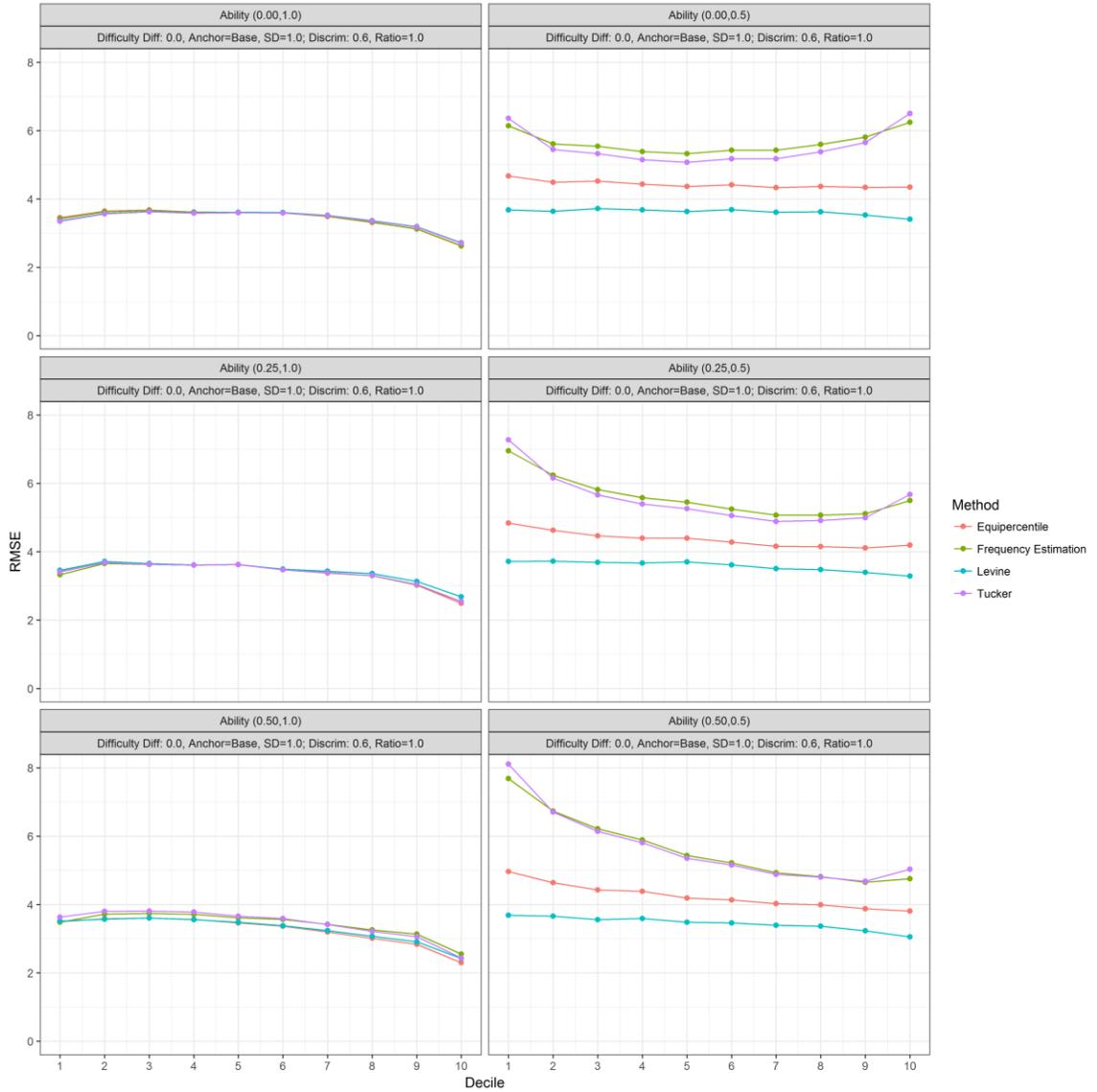


Figure 4.5. Achievement Tests: RMSE Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 0.60

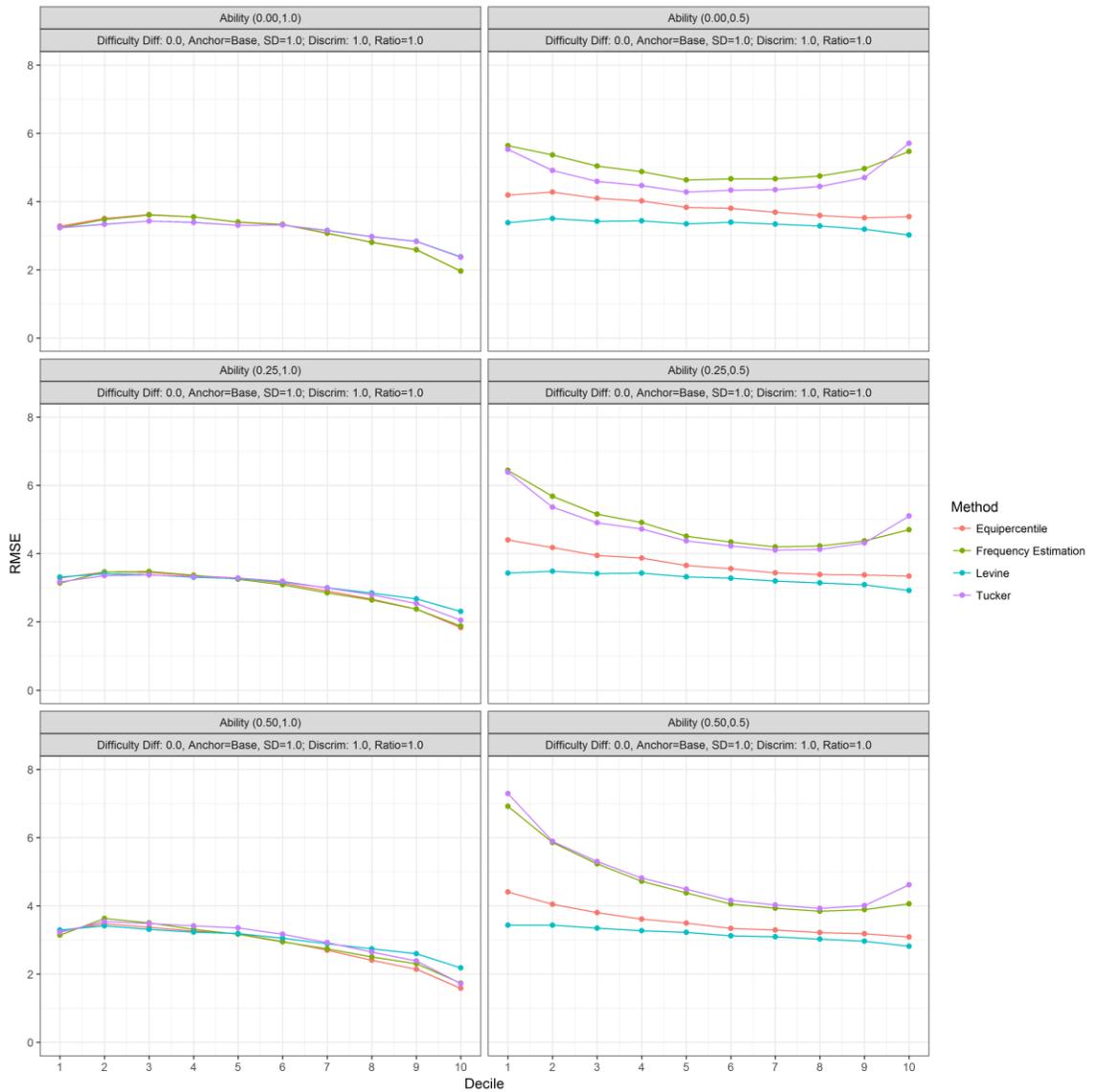


Figure 4.6. Achievement Tests: RMSE Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 1.00

When the ability standard deviation decreased for the alternative form group, the Levine and Equipercntile methods tended to produce the lowest, and most consistent, RMSE results across the ability distribution. A slightly larger difference between the two equating methods was observed when the mean discrimination was 0.60 compared to

1.00. The Tucker and Frequency Estimation methods produced similar RMSE trends, although much larger than the Levine and Equipercentile methods.

### **Achievement Condition Interactions**

There were a number of interactions within the achievement testing scenario which are discussed in this section. First, base and alternative form difficulty differences are discussed within the context of group ability differences. Then, results from varying the anchor set conditions are discussed with respect to group ability differences. Finally, results from manipulating form difficulty, anchor set specifications, and ability differences are presented.

#### **Form Differences and Examinee Ability Differences**

The interaction between off target exams and ability distributional differences was an important motivation for this study. Three alternative form mean difficulty difference conditions, 0.00, 0.25, and 0.50, and three mean ability differences, 0.00, 0.25, and 0.50, were included. However, trends for the middle difficulty and ability conditions, 0.25, are not presented, as they were similar to the results when differences were 0.50, only smaller in magnitude.

Figures 4.7 and 4.8 present bias results for conditions where test form mean difficulty differences were 0.00 and 0.50 and mean ability differences were 0.00 and 0.50. Figure 4.7 presents bias results when the mean item discrimination was 0.60 and Figure 4.8 presents bias results when the mean item discrimination was 1.00. In both figures, the top two rows present results when the alternative form ability standard

deviation was 1.00 and the bottom two rows include results when the standard deviation was 0.50.

For conditions where the alternative form ability standard deviation was 1.00, differences in form difficulty appeared to influence the bias results for the linear equating methods the most, and larger differences were observed when the discrimination was 1.00. Although bias differences were noticeable for the Tucker and Levine methods as the abilities changed, the magnitude was much larger as the form differences increased, surpassing the DTM threshold. It is important to note that the Tucker method experienced much larger changes in bias than the Levine method, which remained relatively stable as the abilities changed.

The nonlinear methods were more influenced by changes in the examinee ability distribution, but remained relatively stable as form differences increased, with respect to bias. The Frequency Estimation method was influenced the most by ability differences, with a shift in bias larger than the DTM threshold. The Equipercentile method was relatively stable to shifts in form difficulty and examinee ability when the standard deviation of the alternative form ability was 1.00.

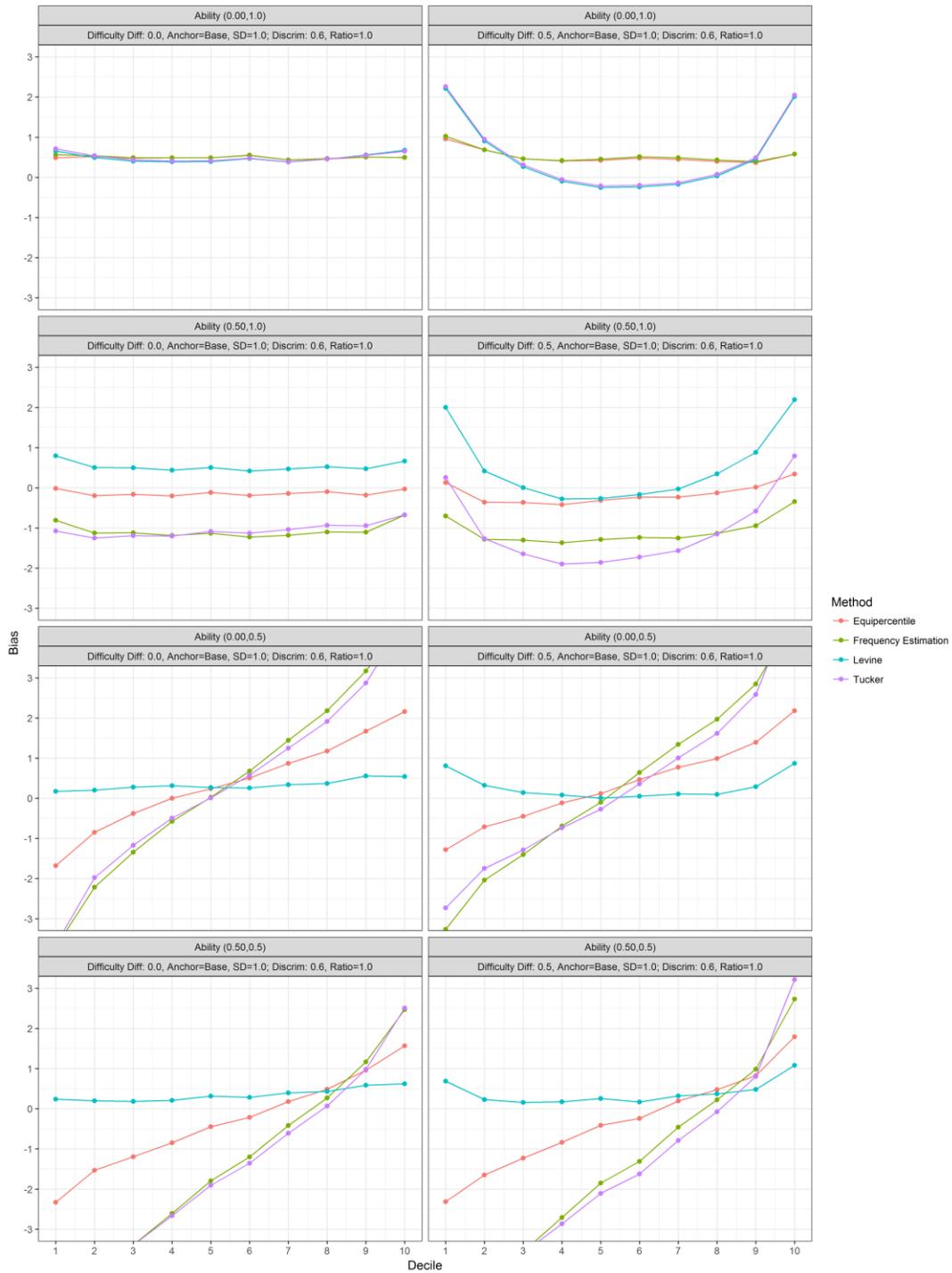


Figure 4.7. Achievement Tests: Bias Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 0.60

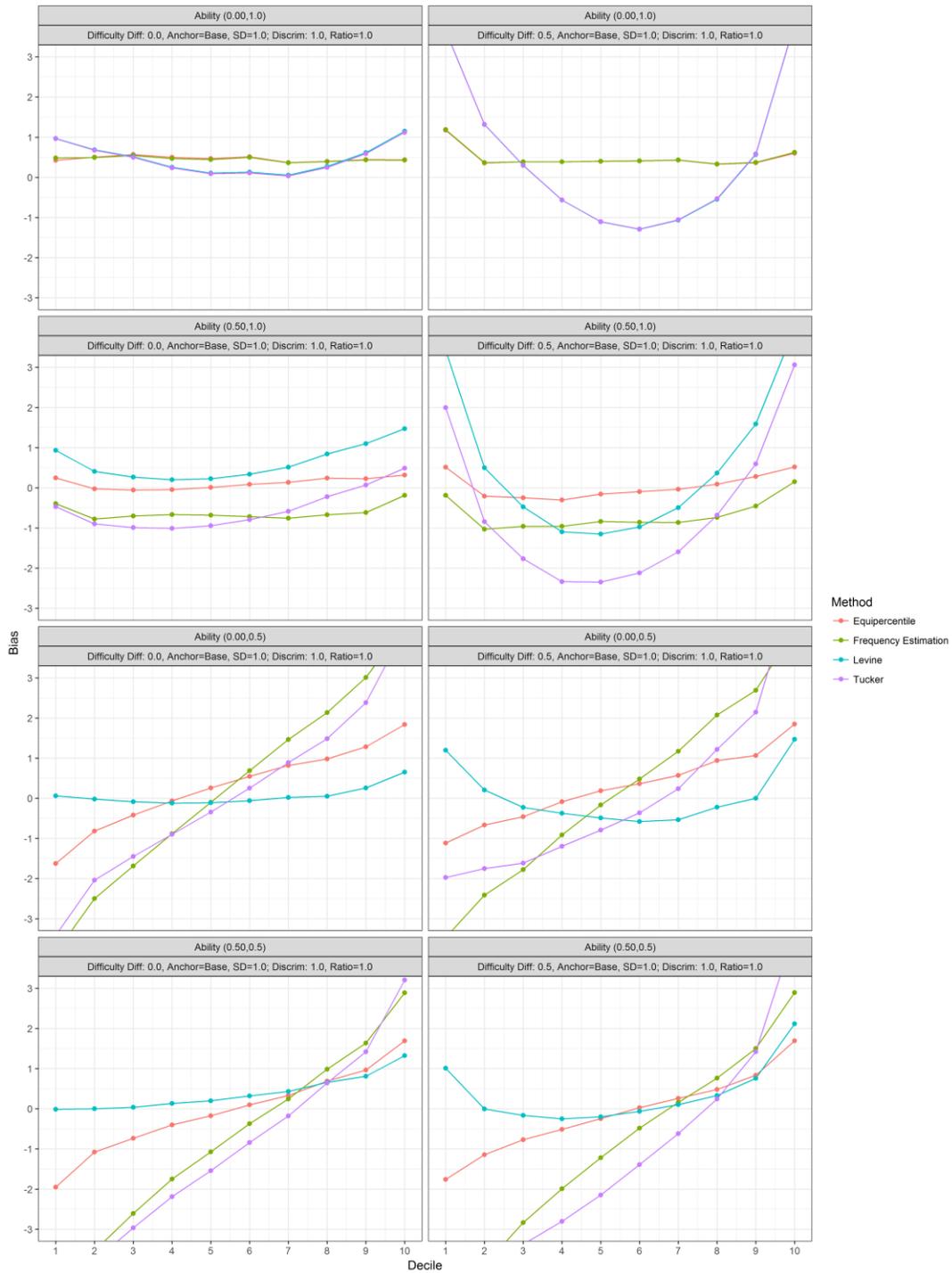


Figure 4.8. Achievement Tests: Bias Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 1.00

Under conditions where the ability standard deviation of the alternative form group was 0.50, the bias results were much different. The Levine method was the most stable under all form difficulty and ability conditions, with large bias differences only noticeable in the tails of the distribution of abilities when the discrimination was 1.00 and the forms differences were 0.50. The other three equating methods were largely influenced by the reduced ability standard deviations, with bias results that mirrored what was described when the discrimination was 0.60.

Figures 4.9 and 4.10 present RMSE results for the interaction between alternative form difficulty and ability differences. Figures 4.9 and 4.10 presents RMSE results when the mean item discrimination conditions were 0.60 and 1.00, respectively. In both figures, the top and bottom rows present results when the alternative form ability standard deviations were 1.00 and 0.50, respectively.

When the ability standard deviations were held constant at 1.00, the RMSE results were much more consistent than the bias results. The least influenced methods were the nonlinear methods, which were only influenced in the extremes of the ability distribution when the discrimination was 1.00. Generally, as differences in form difficulty increased, the linear methods produced slightly less RMSE at the low end of the distribution of abilities while the nonlinear methods produced less RMSE at the upper end of the distribution.

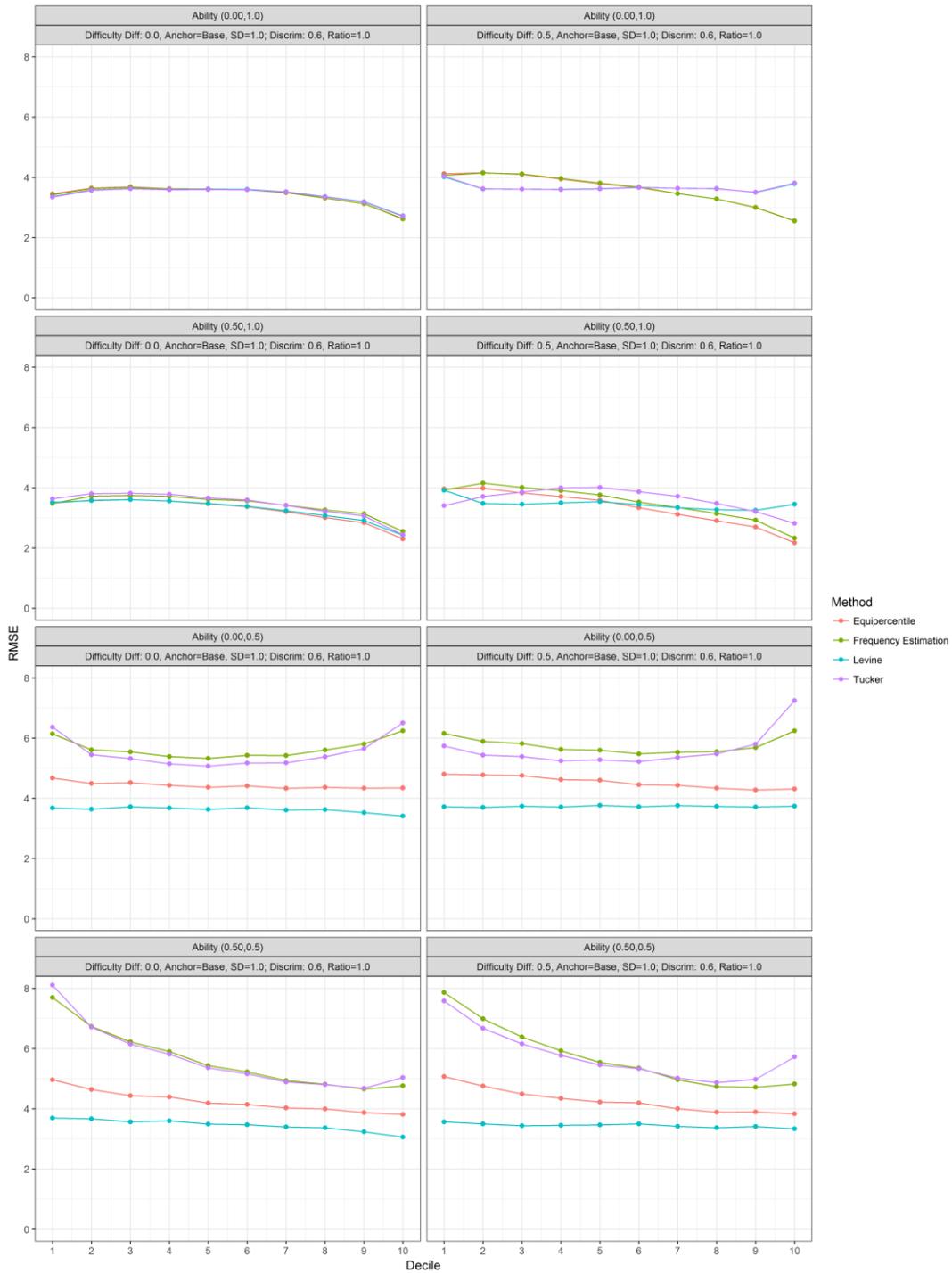


Figure 4.9. Achievement Tests: RMSE Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 0.60

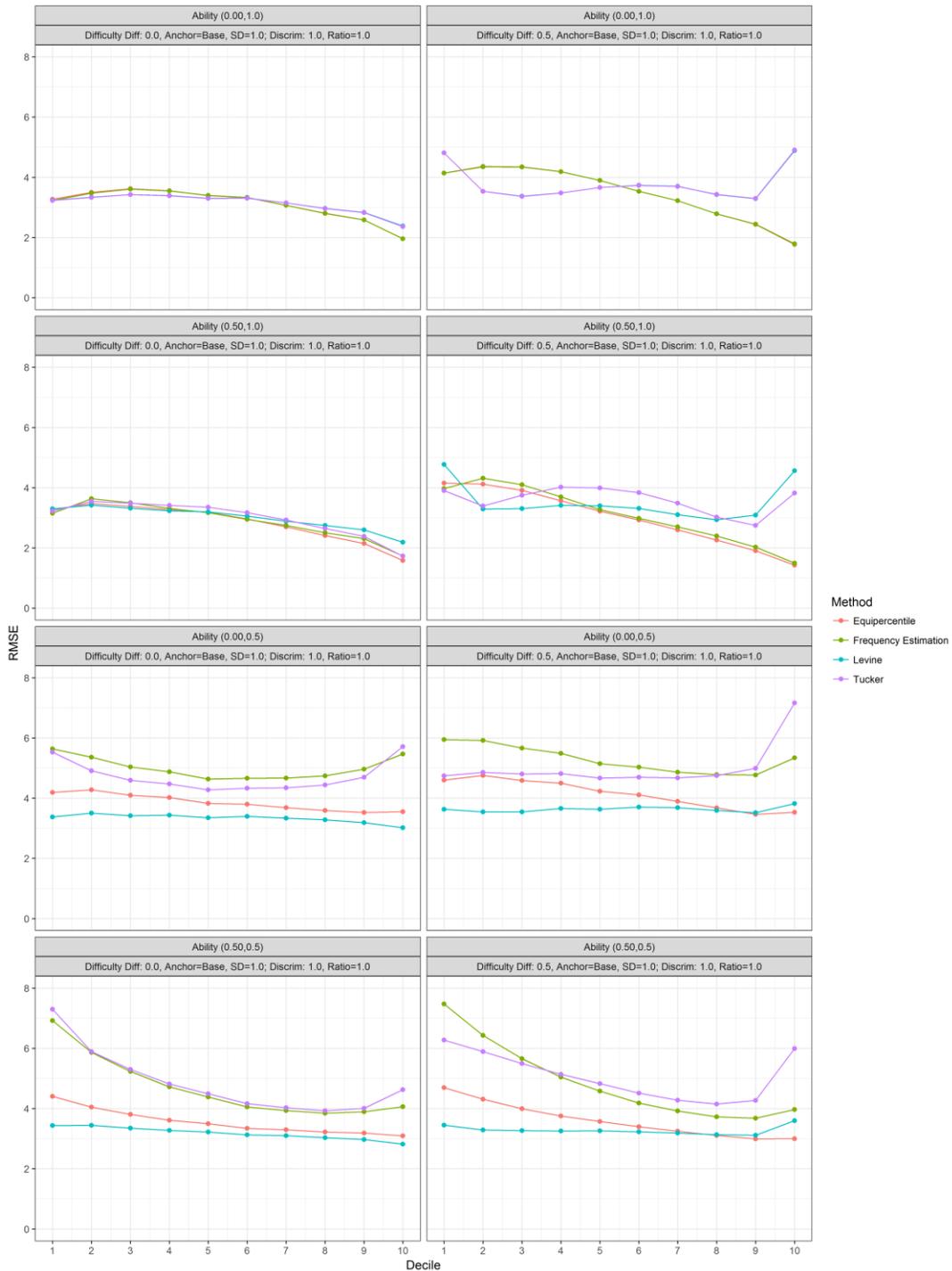


Figure 4.10. Achievement Tests: RMSE Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 1.00

The Levine method produced the lowest RMSE when the alternative form ability distribution was more homogeneous for all conditions, followed by the Equipercentile method. When the discrimination increased from 0.60 to 1.00, the RMSE results produced by the Equipercentile method began to resemble the results from the Levine method. As form differences became larger, the Equipercentile method produced slightly better RMSE results compared to the Levine method for higher performing examinees. The Tucker and Frequency Estimation methods produced much larger RMSE results than the Levine and Equipercentile methods.

### **Anchor Differences and Examinee Ability Differences**

Another specific interest of this study was to examine the relationship between anchor set construction conditions when differences in group abilities are expected. Therefore, this section summarizes bias and RMSE results from manipulating the aforementioned conditions.

The results revealed that neither anchor set condition had a substantial impact on bias when the test forms had the same mean difficulty and the group abilities had the same standard deviation. However, when the alternative form abilities were more homogeneous, equating differences between the anchor set conditions were observed. Figures 4.11 and 4.12 present the bias results for the linear equating methods when the mean item discriminations were 0.60 and 1.00, respectively, under different anchor set construction conditions. Similarly, Figures 4.13 and 4.14 display the bias results for the nonlinear methods under the same conditions.

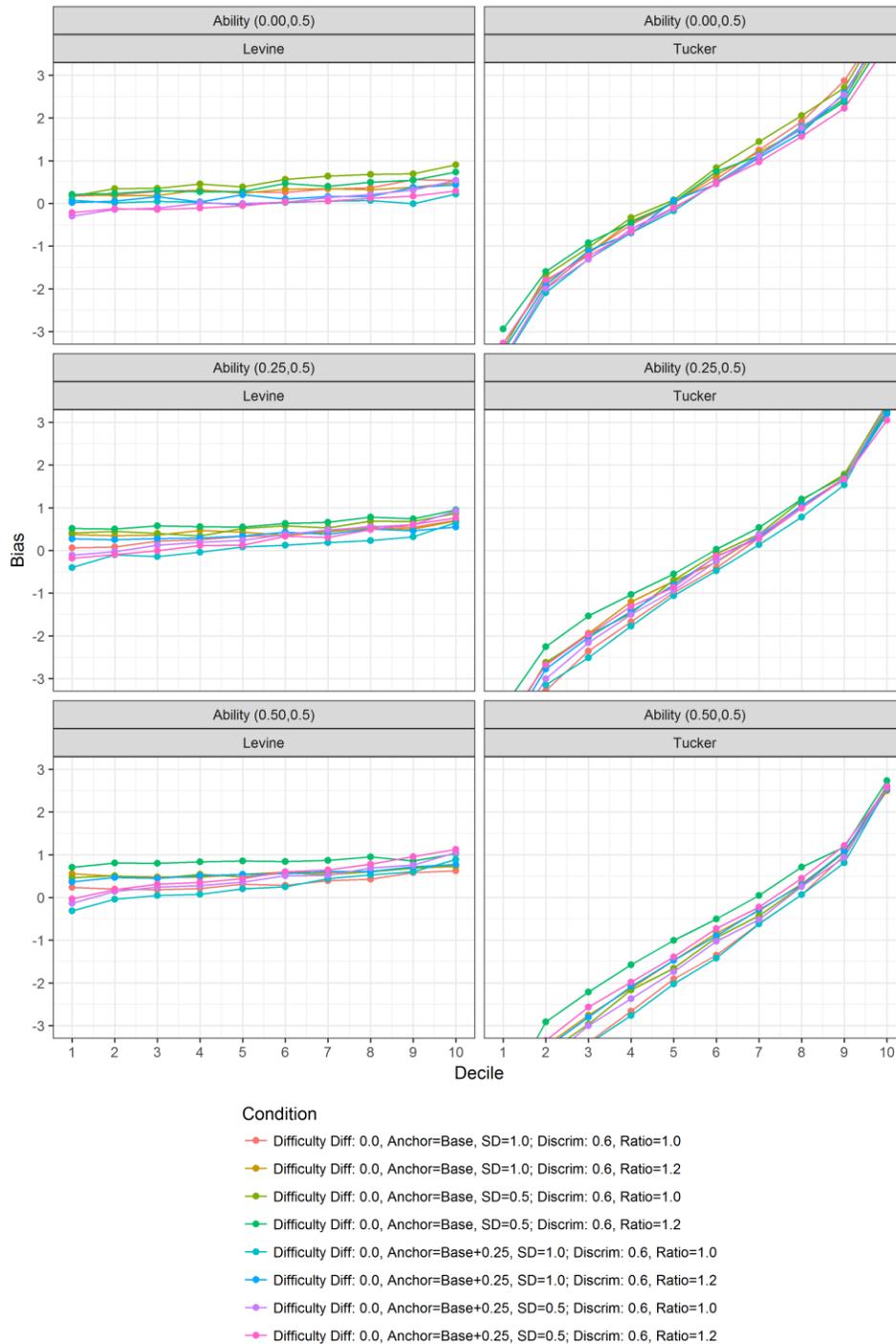


Figure 4.11. Achievement Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 0.60

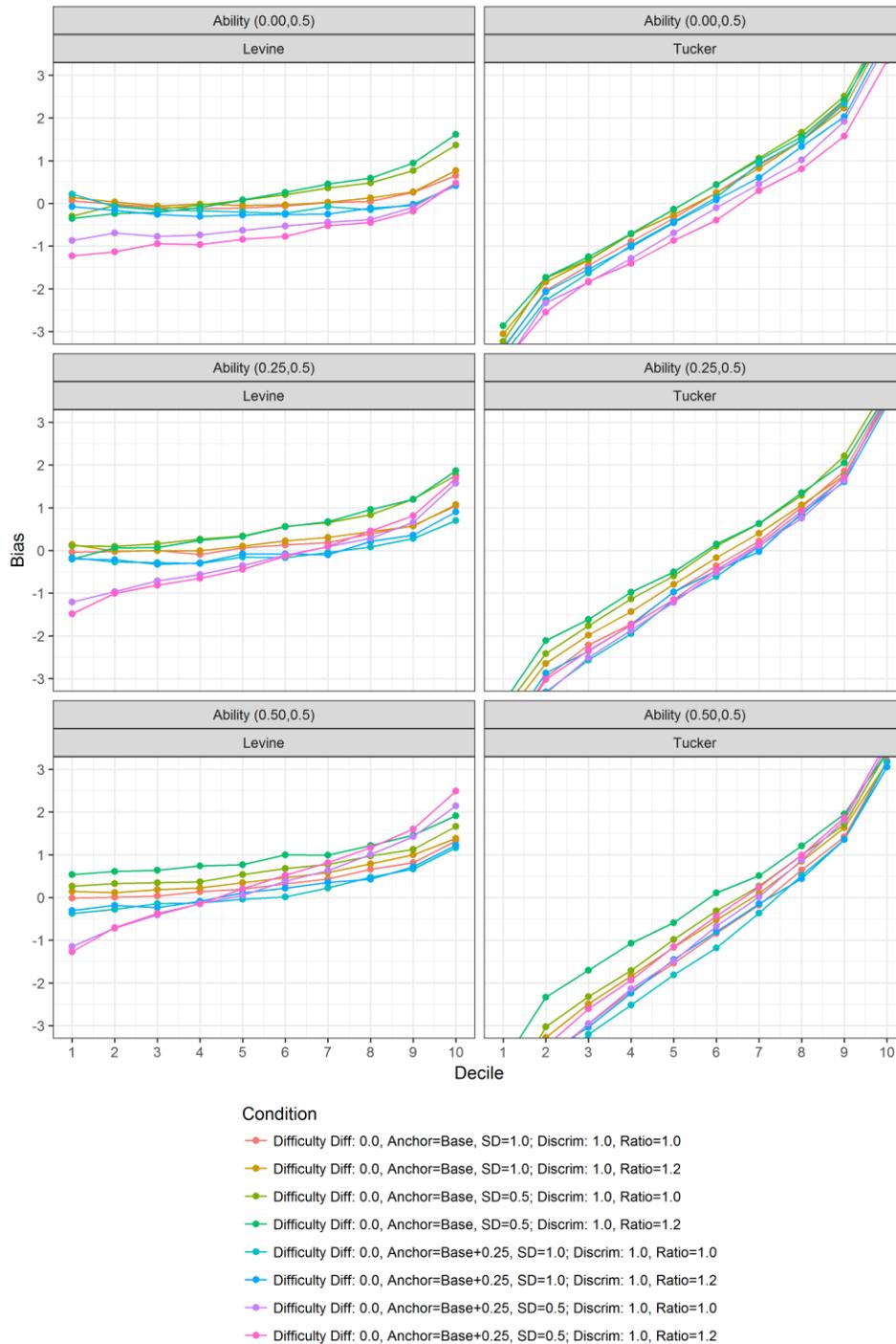


Figure 4.12. Achievement Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 1.00

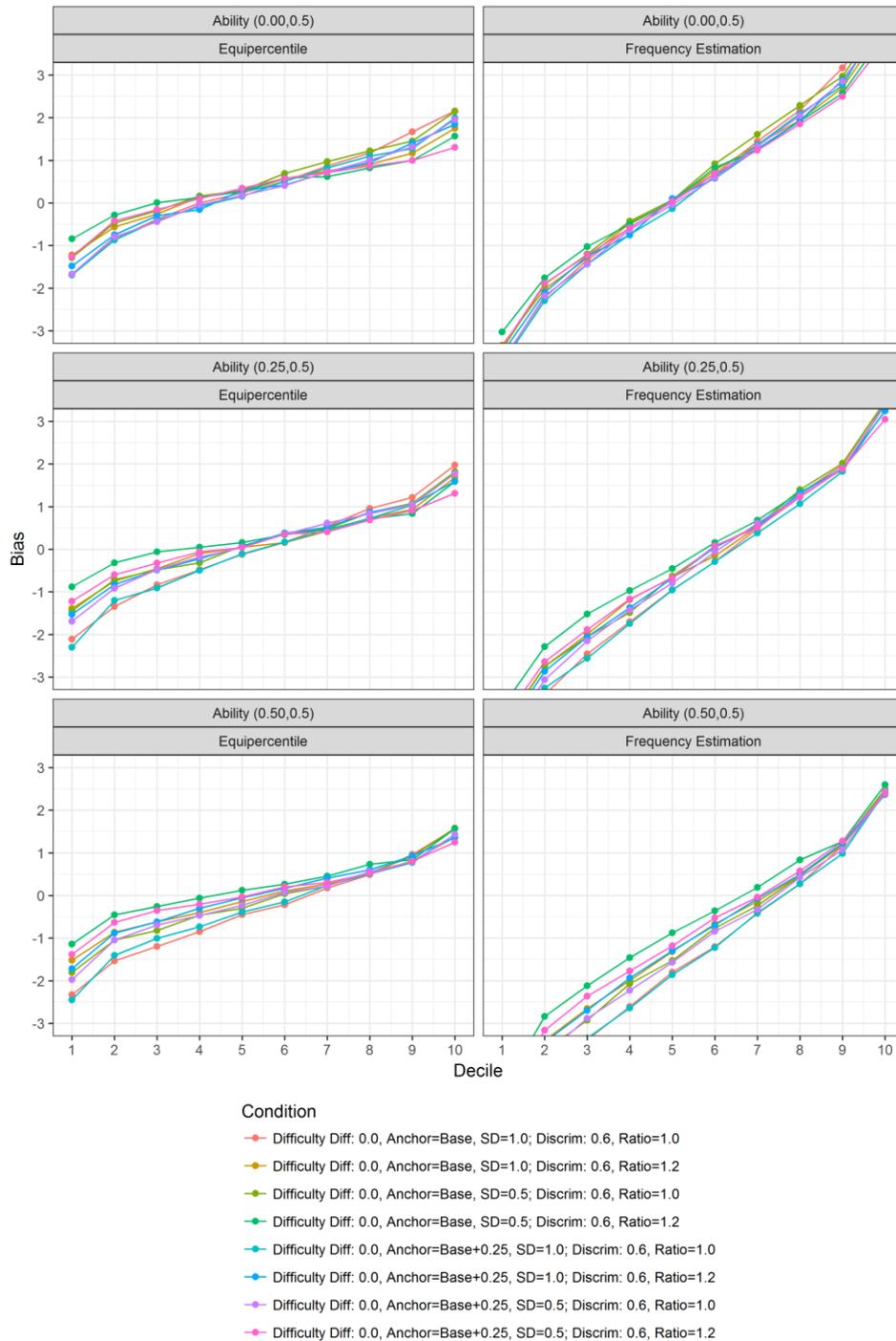


Figure 4.13. Achievement Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60

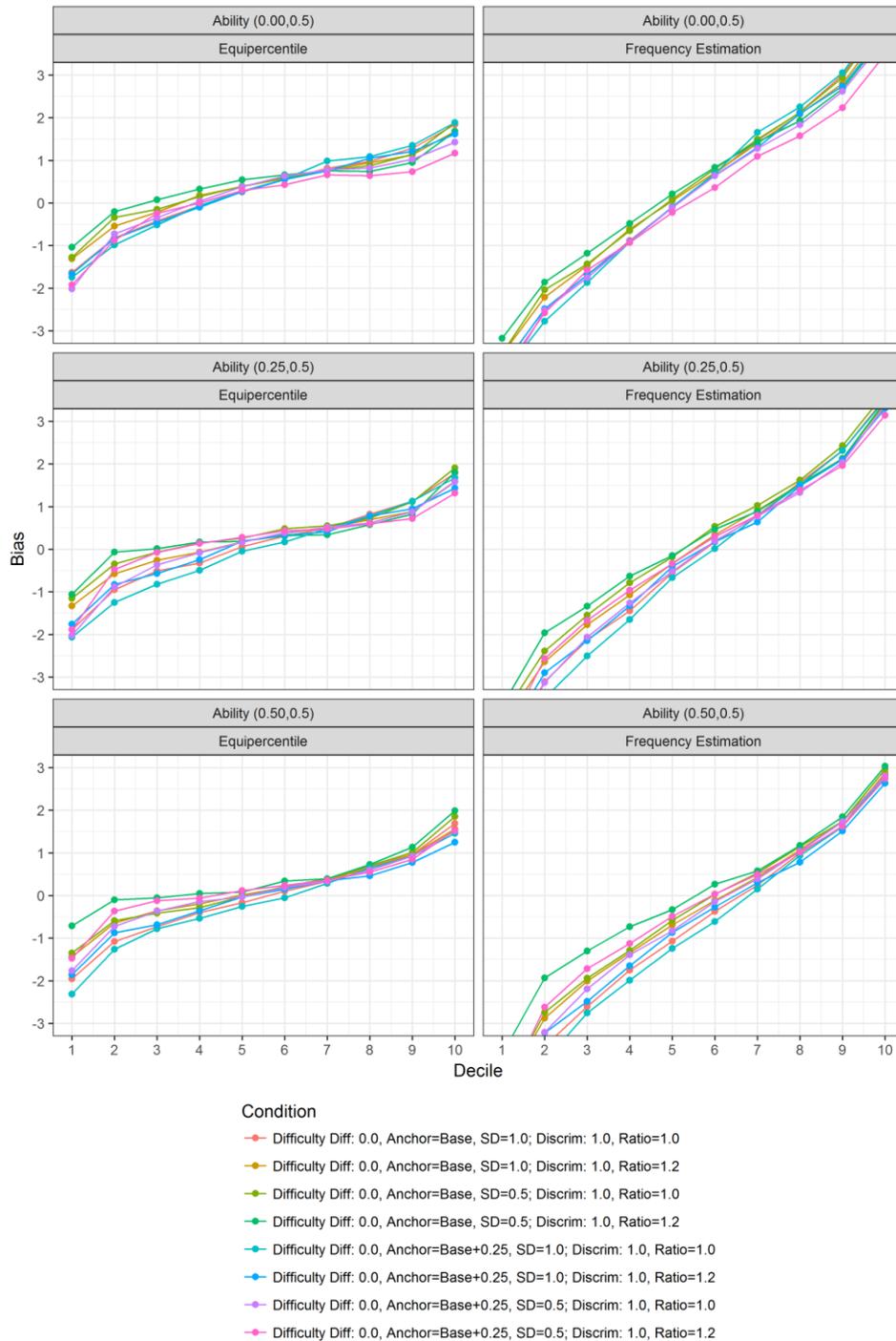


Figure 4.14. Achievement Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00

For tests with lower mean item discrimination, the bias results for the Levine method favored shifting the mean of the anchor set difficulty by 0.25 to produce the least biased results across most of the ability distribution when the alternative form groups were more able. However, bias differences compared to the mini anchor set were rarely greater than the DTM threshold. When the mean item discrimination for the test was 1.00, the bias results for the Levine method indicated that four anchor sets produced nearly equivalent results across the ability distribution: 1) a traditional mini anchor set, 2) an anchor set with increased mean discrimination, 2) an anchor set with increased difficulty, and 3) an anchor set with both increased difficulty and discrimination.

The Tucker and Frequency Estimation methods produced large bias results outside of the center of the ability distribution, regardless of the anchor or mean ability condition, when the alternative form group was more homogeneous. None of the equating anchor set results reduced the bias results enough to meaningfully improve the Tucker and Frequency Estimation methods for use with a more homogeneous alternative form group.

For the Equipercentile method, the anchor sets which produced the least biased results across the ability distribution for both discrimination conditions included: 1) a midi anchor set with increased discrimination, and 2) a midi anchor set with increased difficulty and discrimination. In middle of the distribution of examinee abilities all anchor sets produced similar bias, but in the tails of the ability distribution differences between the highest performing anchor sets and the traditional mini anchor set were greater than the DTM threshold.

The RMSE results suggested that anchor set conditions had little impact on the equating results when the forms had the same mean difficulty and the groups had the same ability standard deviation. However, with more homogeneous groups, greater RMSE differences were observed when the anchor conditions were altered. Figures 4.15 and 4.16 present the RSME results for the linear equating methods when the mean item discriminations were 0.60 and 1.00, respectively, and Figures 4.17 and 4.18 present RMSE results for the nonlinear methods when the ability standard deviation was 0.50.

The Levine method produced similar RMSE results for all anchor set conditions and ability differences when the mean item discrimination was 0.60. When the mean item discrimination was 1.00, a number of anchor sets produced similarly low RMSE. The two anchor sets that tended to produce the largest RMSE when the mean discrimination was increased were: 1) a midi anchor set with increased mean difficulty, and 2) a midi anchor set with increased mean item difficulty and increased discrimination.

When the mean item discrimination was 0.60, the Tucker and Frequency Estimation methods produced the lowest RMSE when using two types of anchors: 1) a midi anchor with increased discrimination or 2) a midi anchor with increased mean difficulty and increased item discrimination. Similar to the large bias results produced by the Tucker and Frequency Estimation methods when the alternative form group was more homogeneous, the RMSE was also larger than the other equating methods under the same conditions.

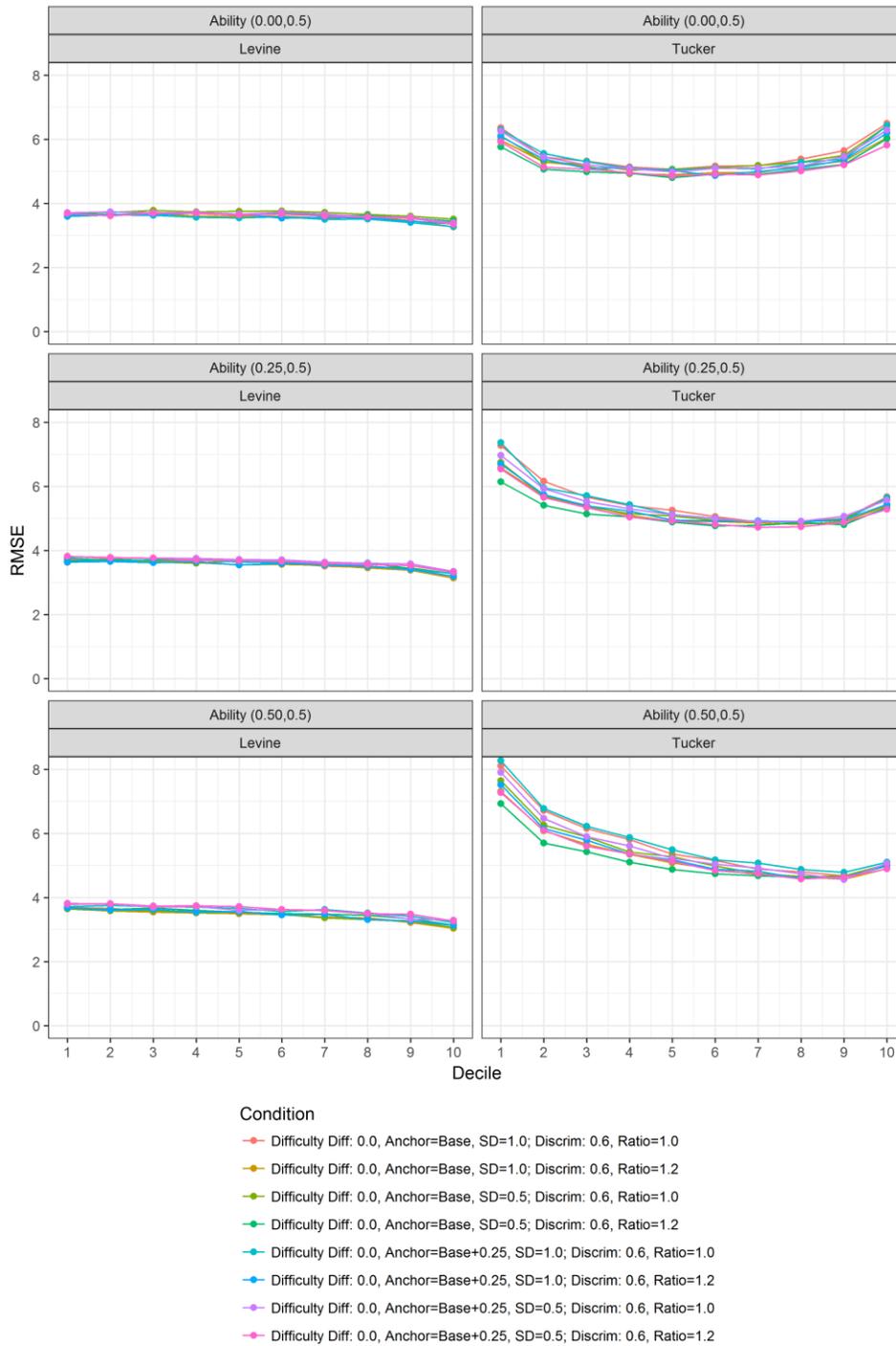


Figure 4.15. Achievement Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 0.60

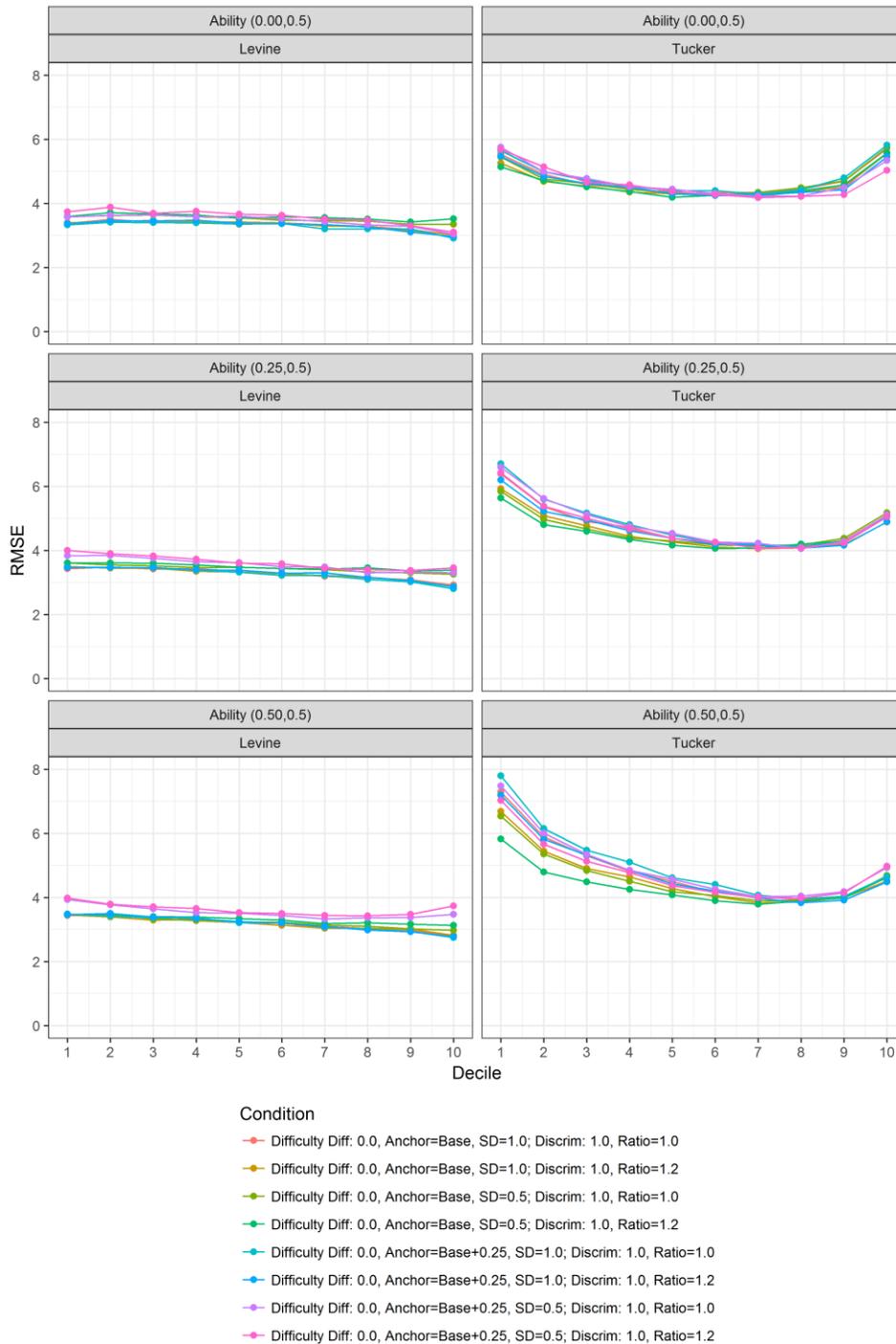


Figure 4.16. Achievement Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 1.00

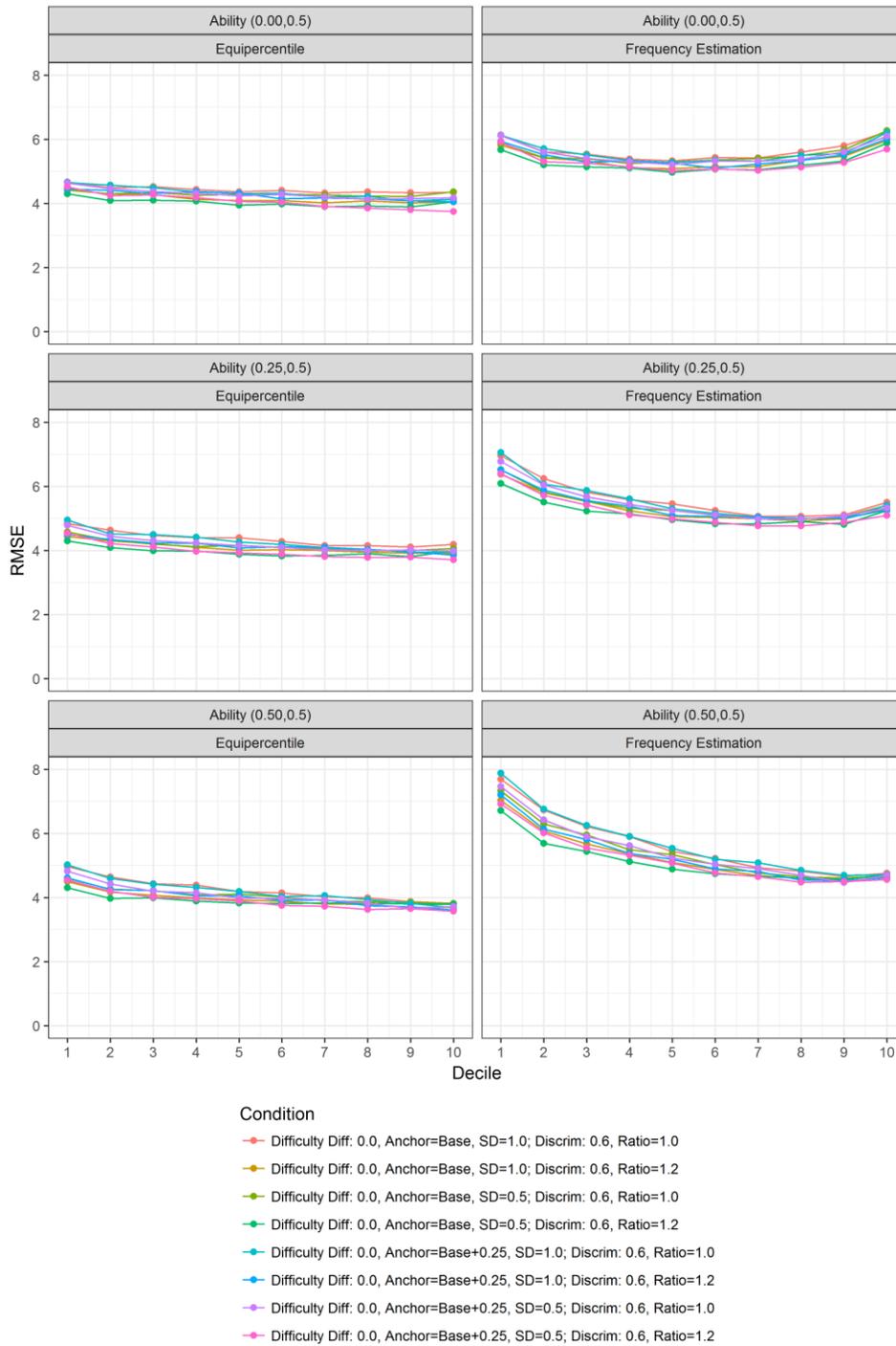


Figure 4.17. Achievement Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60

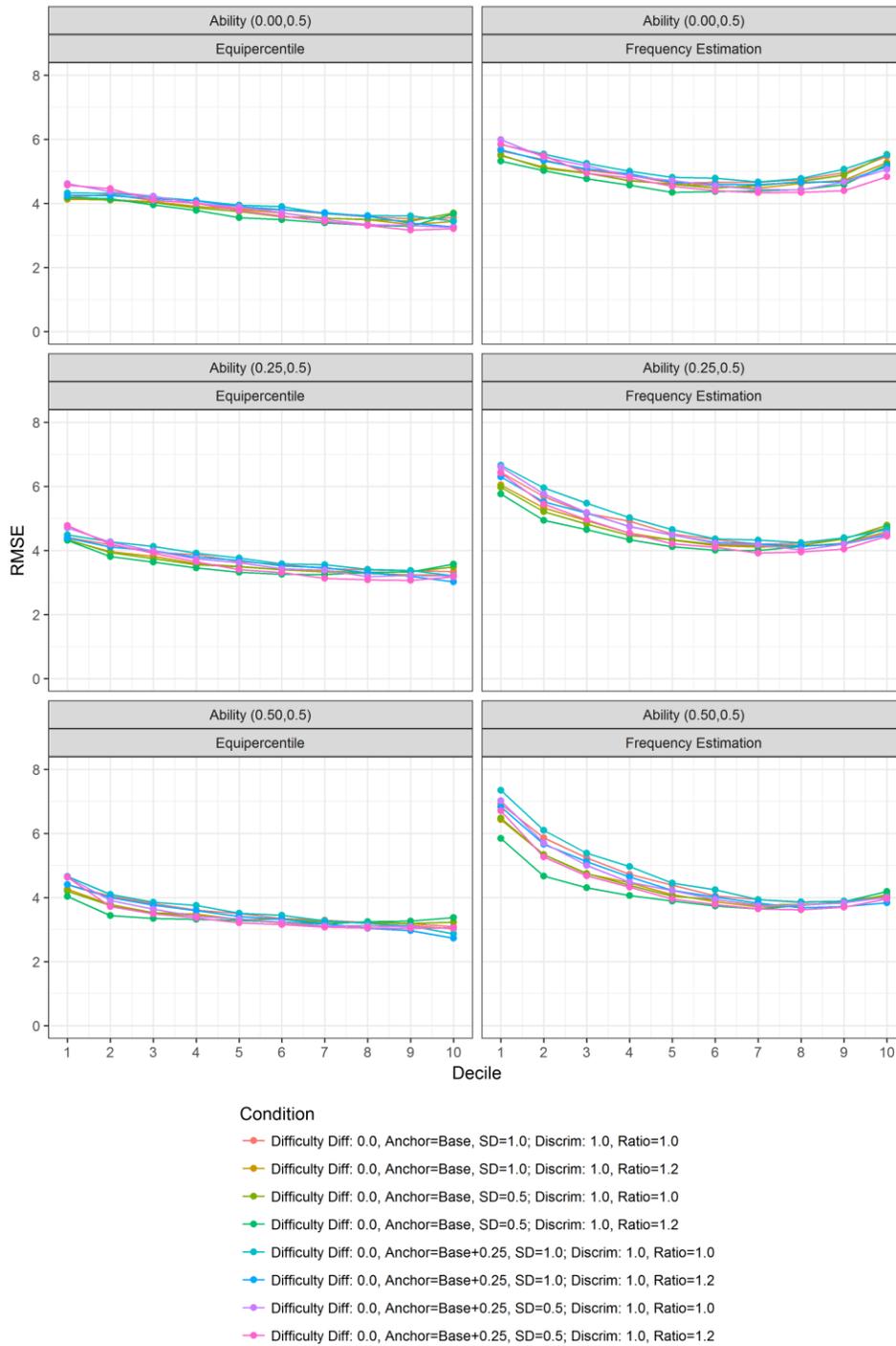


Figure 4.18. Achievement Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00

The Equipercentile method produced the lowest RMSE across the ability distribution for both discrimination conditions with two types of anchor sets: 1) a midi with increased item discrimination or 2) a midi with increased item difficulty and increased discrimination. The observed trend was true for all conditions where the alternative form group was more homogeneous.

### **Form Differences, Anchor Differences, and Examinee Ability Differences**

This section summarizes the results for situations where the alternative form was more difficult than the base form, the anchor set characteristics were manipulated, and examinee ability differences were more able or more able and more homogeneous. Figures 4.19 and 4.20 present the bias results for the linear equating methods when the mean item discriminations were 0.60 and 1.00, respectively, and Figures 4.21 and 4.22 display the bias results for the nonlinear methods. Each figure presents conditions where the test form mean difficulty difference was 0.50 and the mean ability difference of the alternative form group was 0.50. Results for both ability standard deviations are included in each figure.

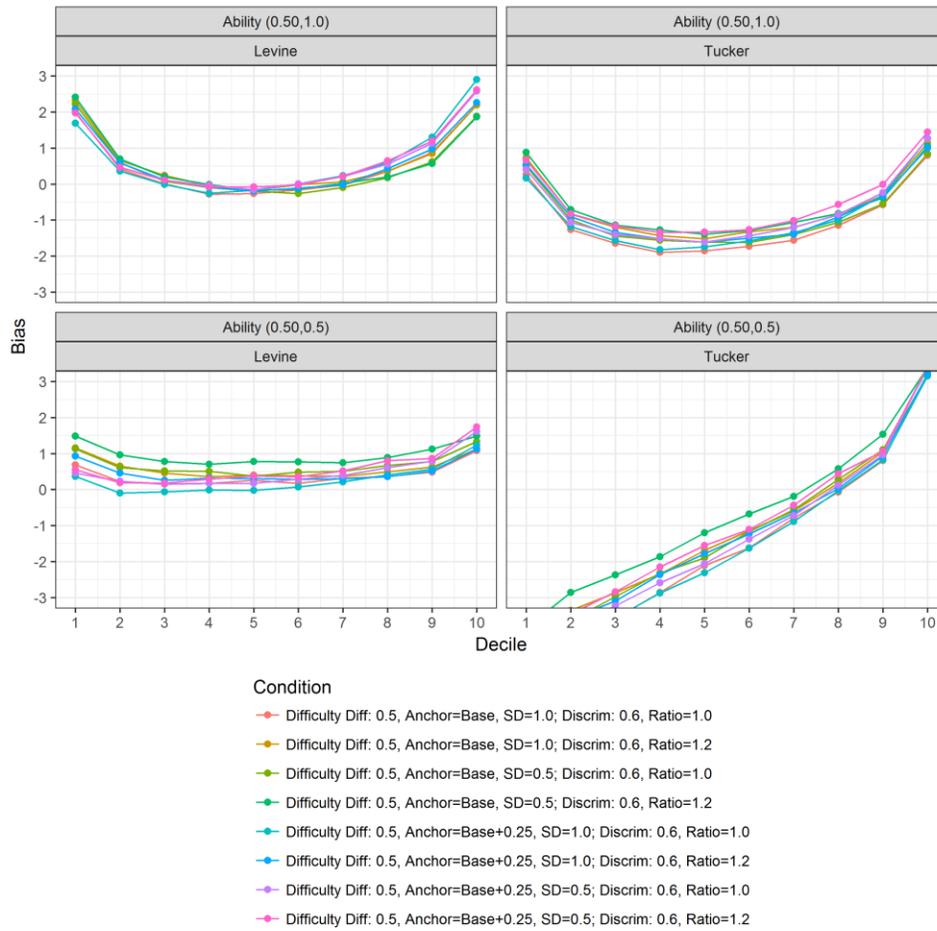


Figure 4.19. Achievement Tests: Bias Results for All Anchor Conditions when Ability Differences were 0.50 for Linear Equating Methods when the Mean Item Discrimination was 0.60

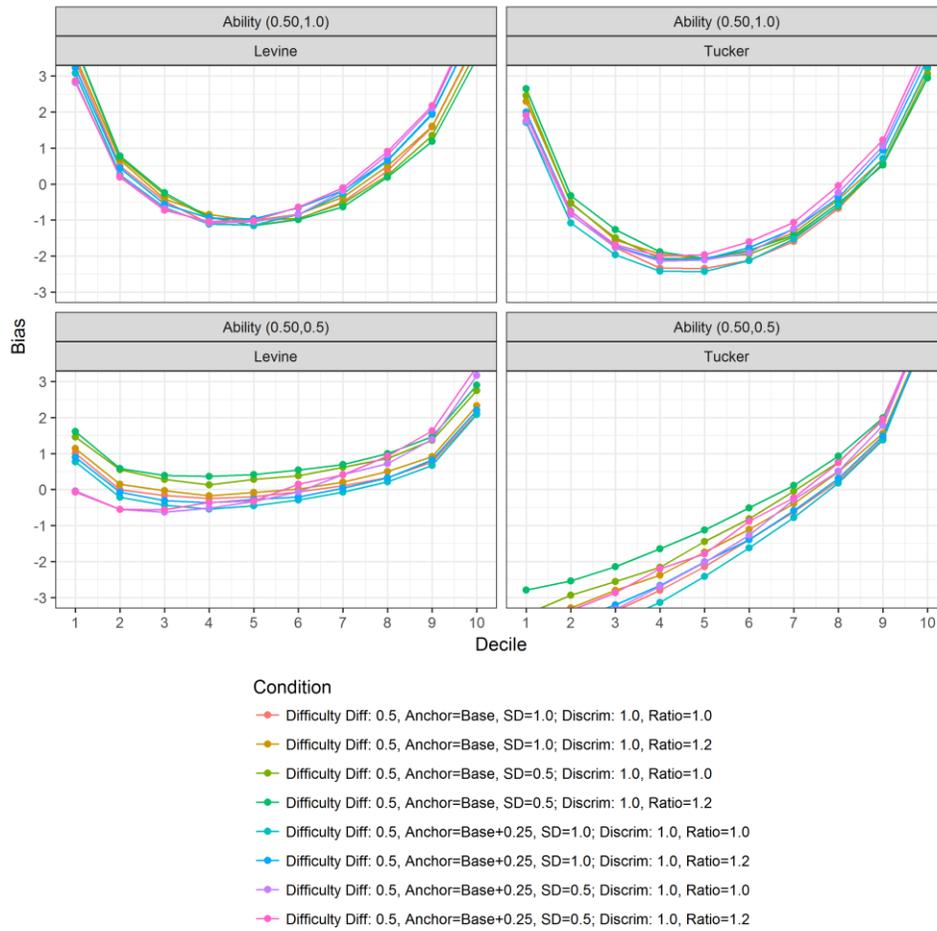


Figure 4.20. Achievement Tests: Bias Results for All Anchor Conditions when Ability Differences were 0.50 for Linear Equating Methods when the Mean Item Discrimination was 1.00

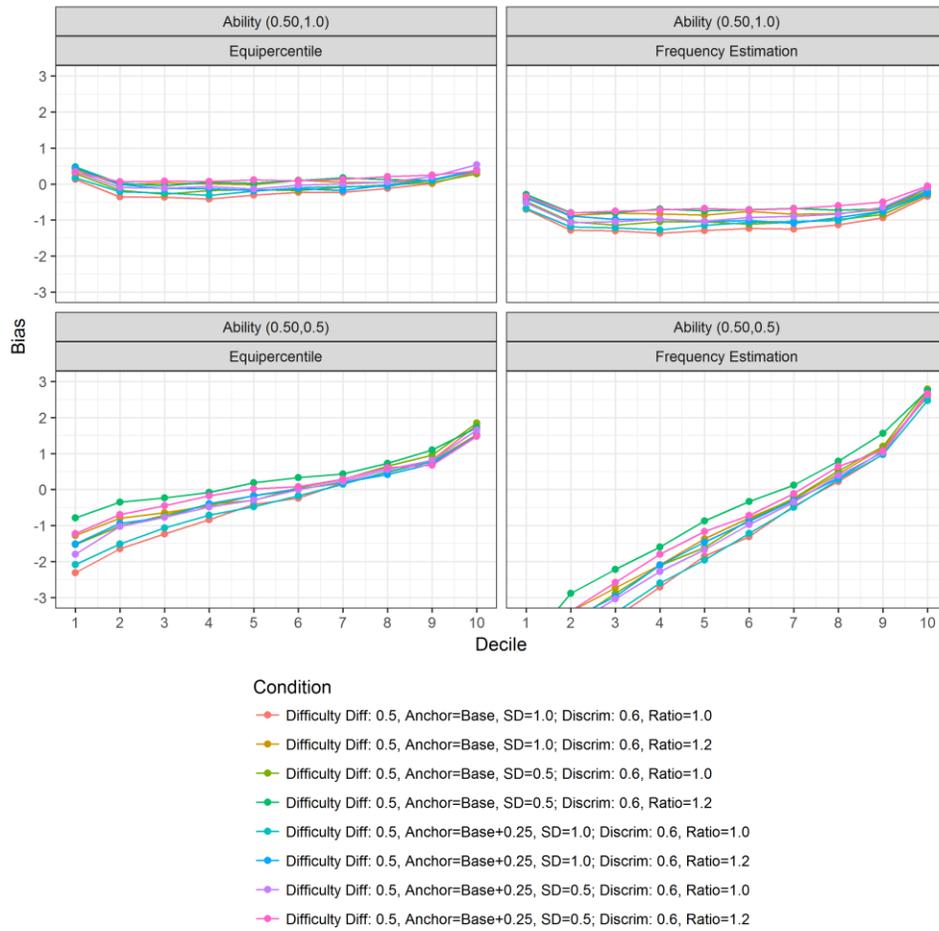


Figure 4.21. Achievement Tests: Bias Results for All Anchor Conditions when Ability Differences were 0.50 for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60

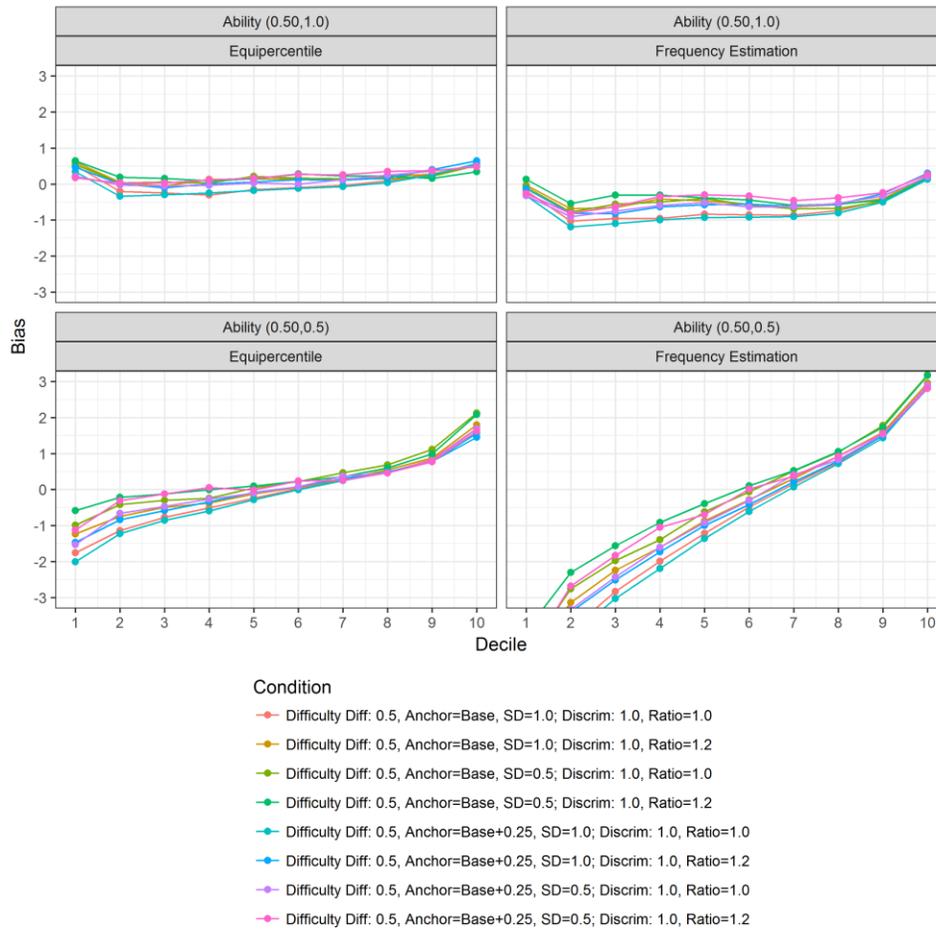


Figure 4.22. Achievement Tests: Bias Results for All Anchor Conditions when Ability Differences were 0.50 for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00

The Levine anchor set conditions produced similar bias results when the standard deviation of the alternative form ability group was 1.00, regardless of discrimination condition. Bias differences between the best performing anchor conditions and the traditional mini anchor set were typically less than the DTM threshold when the ability standard deviation was 1.00. However, there were two distinct results when the alternative form group was more homogeneous. Under the lower discrimination

condition, using a more difficult anchor set produced the least amount of bias. On the other hand, when the discrimination was increased, increasing the mean item discrimination within the anchor set produced the least amount of bias across most of the ability distribution.

The bias results for the Tucker and Frequency Estimation methods revealed that using either 1) a midi anchor set with increased discrimination or 2) a midi anchor with increased difficulty and discrimination produced the least bias results. In the middle of the distribution of abilities, the difference between the best performing anchor set and the traditional mini anchor was larger than the DTM threshold under most conditions for the Tucker and Frequency Estimation methods. It's important to note that the Tucker and Frequency Estimation methods produced reasonable bias results when the ability standard deviation of the alternative form group was 1.00, however the bias results were large when the alternative form group was more homogeneous.

For the Equipercentile method, the most consistent bias results across all ability differences of 0.50 were observed for two types of anchor sets: 1) a midi anchor set with increased discrimination or 2) a midi anchor set with increased mean item difficulty and increased discrimination. For conditions where the standard deviation of the alternative form group was 1.00, the best anchor set construction condition was not clear.

Figures 4.23 and 4.24 present the RSME results for all equating methods when the mean item discriminations were 0.60 and 1.00, respectively. The RMSE results presented somewhat of a dichotomy. The Equipercentile, Frequency Estimation, and Tucker methods produced the lowest RMSE results using two types of anchors: 1) a midi

anchor with increased mean item discrimination or 2) a midi anchor that included difficulty and discrimination. The results were consistent for both discrimination conditions, with the exception of the Tucker method when the mean item discrimination was 1.00.

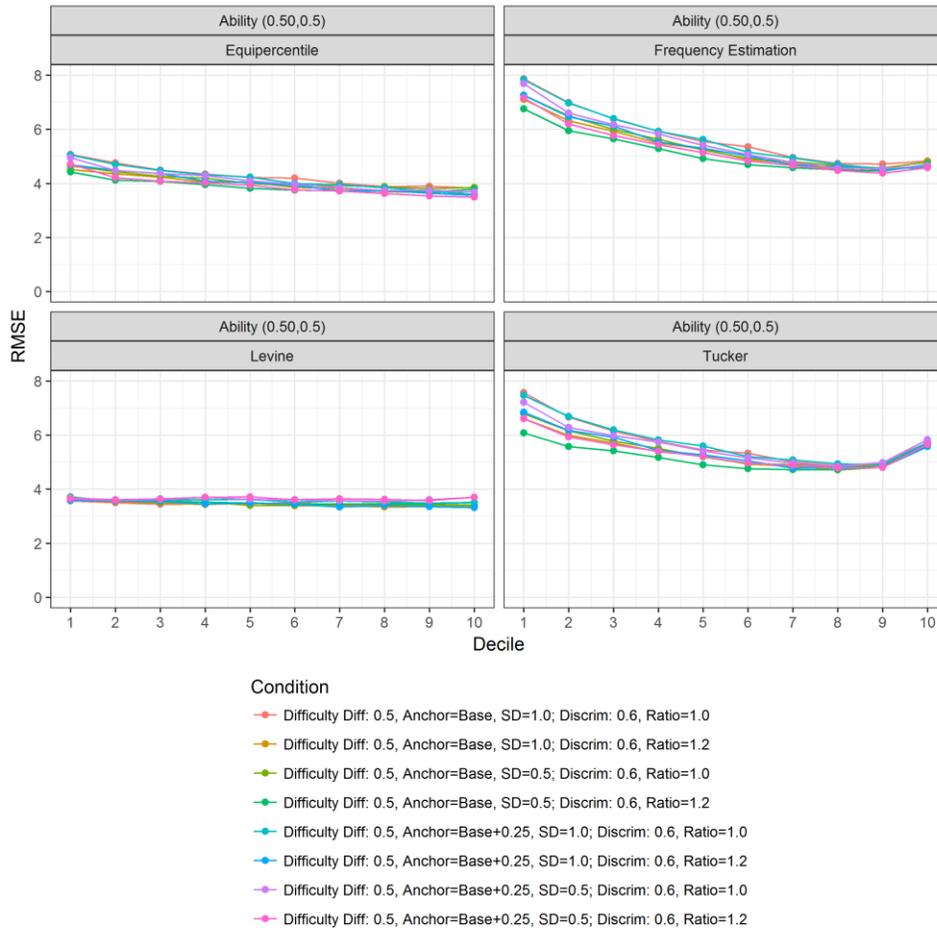


Figure 4.23. Achievement Tests: RMSE Results for All Anchor Conditions when Homogeneous Ability Differences were 0.50 for All Equating Methods when the Mean Item Discrimination was 0.60

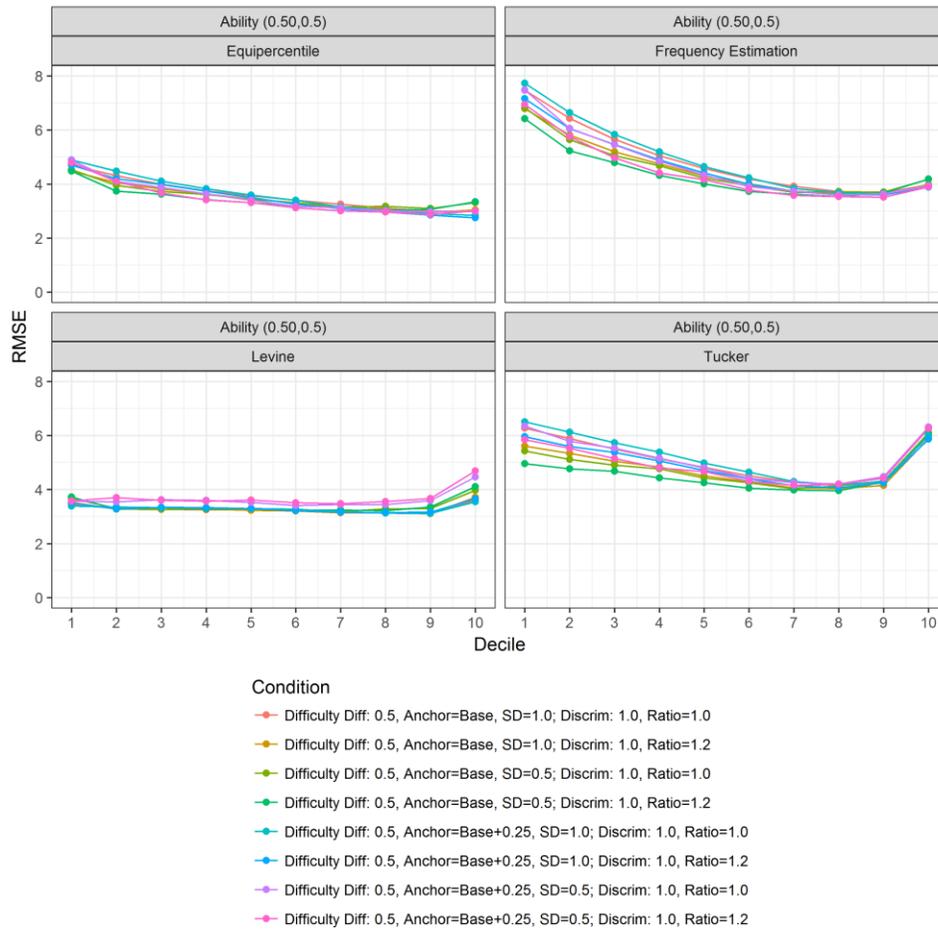


Figure 4.24. Achievement Tests: RMSE Results for All Anchor Conditions when Homogeneous Ability Differences were 0.50 for All Equating Methods when the Mean Item Discrimination was 1.00

The Levine method produced a number of similar RMSE values with the various anchor conditions. Although there was not a clear best anchor set, the largest RMSE was consistently produced by two specific anchor sets: 1) a midi anchor with a shifted mean difficulty and 2) a midi anchor with increased difficulty and discrimination.

## **Achievement Summary**

The purpose of this section is to summarize the results in a meaningful way to answer the second research question which reads: with respect to the test purpose and specifications, can anchor set assembly rules be established for linear (Tucker and Levine Observed Score) and nonlinear (Frequency Estimation and Equipercentile) equating methods when differences in group characteristics are expected?

### **Ability Mean Differences and Similar Form Difficulty**

When ability differences between the base and alternative form were large and the forms were similar in difficulty, the Equipercentile method produced the least amount of systematic error and total error. As groups became more different, bias results improved under the Equipercentile method, when compared to the bias results when group abilities were the same. With respect to total error, all equating methods produced essentially the same results as the ability groups became more different.

### **Ability Mean Differences and Dissimilar Form Difficulty**

The Equipercentile method produced the least amount of systematic error across the ability distribution when both the ability differences and form difficulty differences were large for the base and alternative forms. The Levine method produced similar results in the middle of the distribution, although in the tails the bias results were quite large. With respect to total error, the nonlinear methods tended to produce the smallest RMSE for middle and high scoring examinees while the linear methods produced the smallest RMSE at the lower end of the distribution. Therefore, the best equating method was location-specific when forms were not similar in difficulty with respect to RMSE.

### **Ability Mean and Standard Deviation Differences and Similar Form Difficulty**

The results of this study suggested that equating with a more homogeneous alternative form group created substantial equating bias for the Tucker, Frequency Estimation, and Equipercentile methods, even when the mean ability remained the same. The most consistent equating method when the alternative for group was more homogeneous was the Levine method. The Levine method produced nearly equivalent bias results across the ability distribution and the smallest RMSE results. Although the Equipercentile method was similar under some conditions, the Levine method was almost always the most consistent method. The Tucker and Frequency Estimation methods produced much larger error, and the results suggest that they should not be used to equate test forms with a more homogeneous group.

### **Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty**

The results suggested that the Levine and Equipercentile methods produced the least amount of equating error when the alternative form ability group was more homogeneous and test forms differed in difficulty. The Levine method tended to produce the smallest bias and RMSE results when test forms had smaller mean item discrimination. The results were somewhat mixed for higher discriminating tests near the upper end of the ability distribution, where the Equipercentile method produced the smallest amount of total equating error. The results provided evidence for using either the Levine or Equipercentile method with a more homogeneous alternative form group, depending upon the desired location of precision.

### **Anchor Set Construction Techniques with Ability Mean Differences and Similar Form Difficulty**

Anchor set construction techniques had little impact on equating error when only mean ability differences existed. Therefore, the results support the ability to relax some of the long-held rules for constructing anchor sets if groups simply differ with respect to mean ability.

### **Anchor Set Construction Techniques with Ability Mean Differences and Dissimilar Form Difficulty**

Equating with the Equipercentile method produced the smallest amount of bias across the ability distribution compared to all other equating methods, although the RMSE results were similar for all four equating methods. The bias results for the Equipercentile method were improved using two types of anchor sets: 1) a midi with increased item discrimination or 2) a midi anchor with increased difficulty and discrimination. It's important to note that the bias results for the Tucker and Frequency Estimation methods were also improved by the two midi anchor sets.

### **Anchor Set Construction Techniques with Ability Mean and Standard Deviation Differences and Similar Form Difficulty**

Overall, the Levine method tended to produce the lowest equating error results when the alternative form abilities were more homogeneous. Although a number of anchor set construction techniques reduced the Levine method equating error, the best technique was not clear. In summary, the Levine method was the most consistent equating method with a more homogeneous sample, and the results suggest that

practitioners could potentially be somewhat flexible when constructing anchor sets under the equating method.

Equating error results for the Equipercentile method were clearly reduced for conditions with two specific anchor set construction types: 1) a midi anchor set with increased discrimination, and 2) a midi anchor set with increased difficulty and discrimination. As a result, the two anchor construction techniques could be considered when an alternative form group is expected to be both more able and more homogeneous.

### **Anchor Set Construction Techniques with Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty**

When form, anchor, and ability conditions were varied the Levine method produced the smallest bias results under the lower discrimination condition when an anchor with increased difficulty was used. For higher discriminating tests, an anchor with a larger mean item discrimination produced the least amount of bias across the majority the ability distribution. There was not a clear best anchor construction method with respect to RMSE.

For the Equipercentile method the two anchors conditions which reduced bias and RMSE were: 1) a midi anchor with increased discrimination or 2) a midi anchor with increased difficulty and discrimination. These two anchor types consistently produced the lowest amount of equating error for the Equipercentile methods across all conditions.

## **Evaluating Certification Test Forms**

Certification tests are designed with cut score in mind. Therefore, equating error near the cut score is more important for certification exams than equating error in other areas of the ability distribution.

In this study, tests were designed to have a pass rate of approximately 90%, so equating error in the first and second deciles were the most important when alternative form examinee abilities were similar to the base form examinees. Under conditions where the alternative form group was more able or more homogeneous, or both, equating error within the first decile was the most important. Although equating error near the cut score is the focus of the following sections, differences as they appear within each decile of the ability distribution are also reported.

### **Certification Test Form Difficulty**

The base form certification tests were generated to have a mean IRT difficulty of -0.64 and a standard deviation of 1.00, to emulate a test with a pass rate of approximately 90%. The alternative form conditions included difficulty differences of 0.00, 0.25, and 0.50 with mean discrimination conditions of 0.60 and 1.00. This section presents the equating systematic and total error results under the test difficulty conditions. Bias results for all four equating methods are presented in Figure 4.25, and RMSE results are provided in Figure 4.26.

When the base and alternative forms were generated to have the same difficulty and the mean item discrimination was 0.60, all four equating methods produced essentially the same amount of bias near the cut score, as well as across the ability

distribution. However, as difficulty differences between the forms increased the linear and non-linear methods produced different patterns of bias. For instance, the Equipercentile and Frequency Estimation methods produced a relatively consistent amount of positive bias across the entire ability distribution, while the Levine and Tucker methods produced a curve. In the middle of the ability distribution the linear methods produced near-zero bias, while large positive bias was observed in the tails of the distribution. As a result, when form difference increased the nonlinear equating methods produced less bias than the linear methods in the first decile, and approximately the same amount of bias in the second decile. When form differences were 0.50, differences between nonlinear and linear methods near the cut score in the first decile were larger than the DTM threshold. The pattern of results was similar when the discrimination was 1.00, with larger differences between linear and nonlinear methods observed.

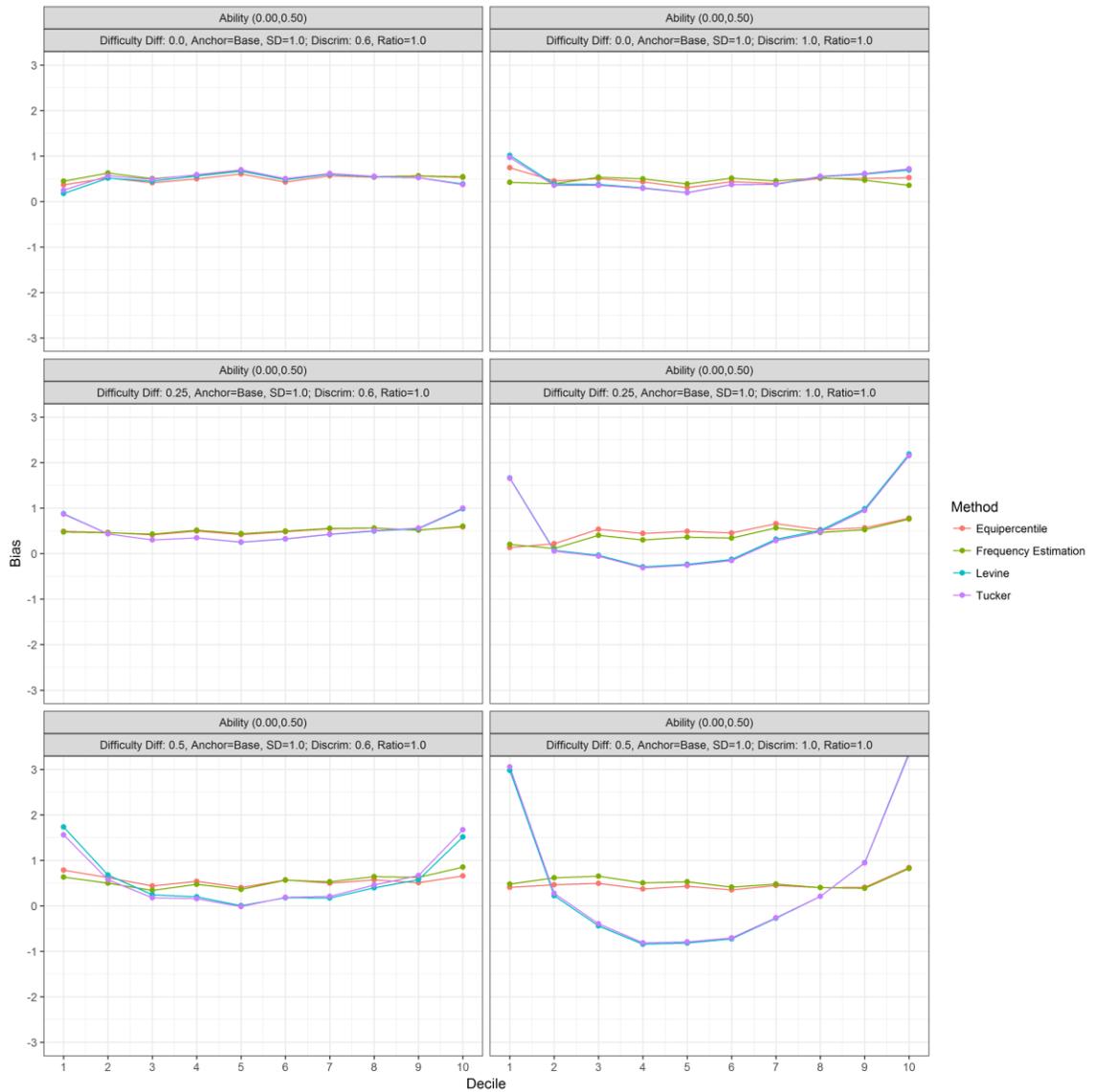


Figure 4.25. Certification Tests: Bias Results for All Equating Methods when Form Difficulty Differences were 0.00 and 0.50 for both Discrimination Conditions

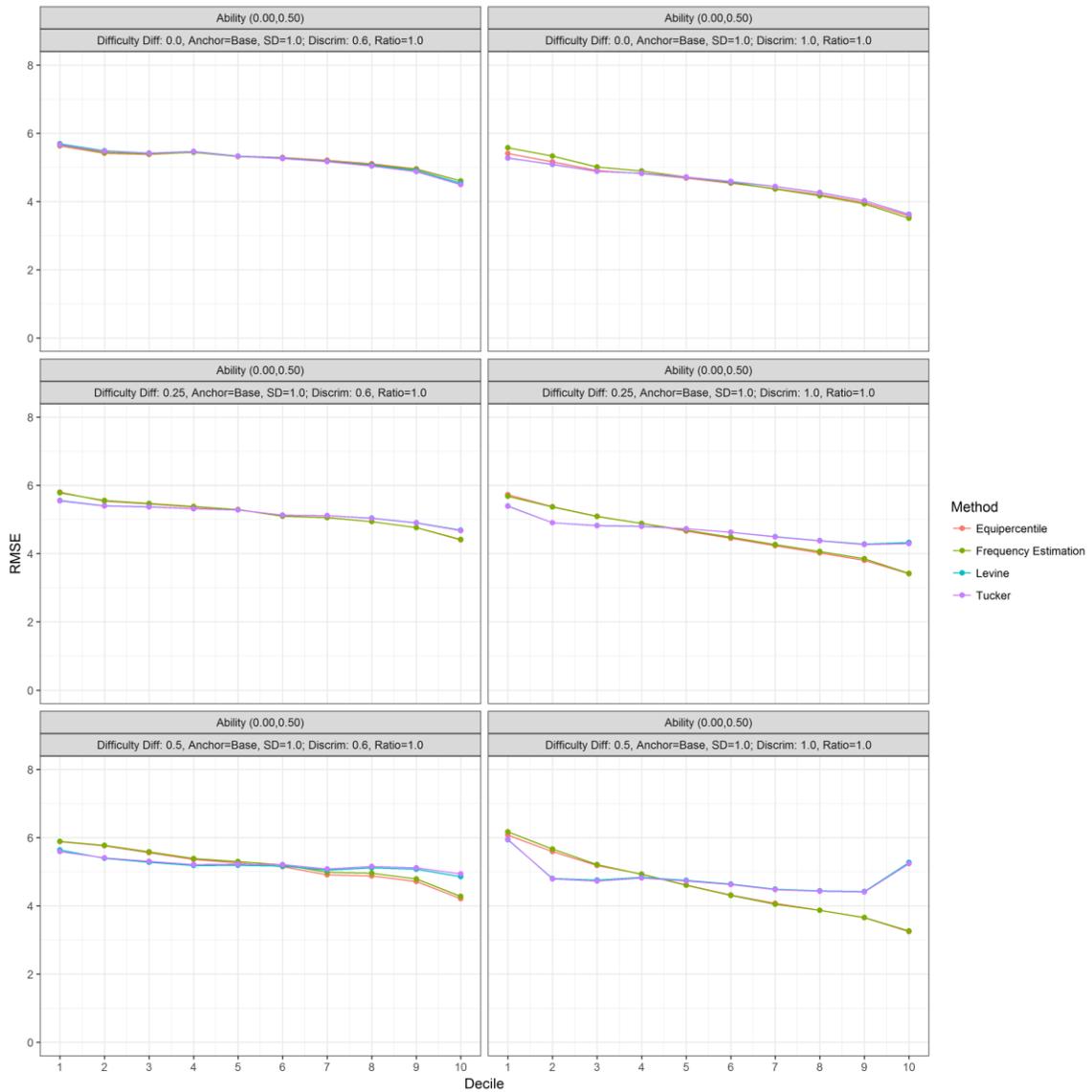


Figure 4.26. Certification Tests: RMSE Results for All Equating Methods when Form Difficulty Differences were 0.00 and 0.50 for both Discrimination Conditions

The RMSE results were similar for all equating methods when form difficulties were the same and the mean item discrimination was 0.60. Method differences were more pronounced as the difficulty differences between the forms increased. The linear methods produced slightly less RMSE at the low end of the ability distribution and the nonlinear

methods produced less RMSE at the upper end of the ability distribution. In the middle of the distribution, the linear and nonlinear equating methods produced approximately the same amount of RMSE. Larger RMSE differences between linear and nonlinear methods were observed when the discrimination was 1.00, although the overall patterns were the same.

### **Certification Anchor Difficulty Alignment**

Two conditions were generated to examine equating error based on anchor set alignment with the base form. Anchor sets were generated to either: 1) align with the base form, or 2) have a mean difficulty 0.25 greater than the base form. The off-target anchor set condition produced similar results to the well aligned anchor, when other conditions were held constant.

### **Certification Anchor Difficulty Standard Deviation**

Two conditions were generated to examine equating error related to spread of item difficulty within the anchor set. Two standard deviation conditions were generated: 1) an anchor set with a difficulty standard deviation of 1.00, and 2) a midi anchor set with a standard deviation of 0.50. The RMSE results were similar across the ability distribution for both anchor set spread of item difficulty conditions for all equating methods. Although a slight bias improvement was observed for all equating methods near the cut score, differences were less than the DTM threshold.

### **Certification Anchor Discrimination**

Two anchor set discrimination conditions were included in the study: 1) an anchor set with the same discrimination as the base and alternative forms of the test, and 2) an anchor with increased discrimination, by 20%. The bias and RMSE results were similar for both discrimination conditions, holding other conditions constant.

### **Certification Ability Conditions**

An important interest of this study was to examine equating error when the alternative form ability group was more able or more homogeneous, or both, compared to the base form group of examinees. Therefore, mean ability differences of 0.00, 0.25, and 0.50 were included in the study along with ability standard deviations of 0.50 and 0.25. Figures 4.27 and 4.28 include bias results for the aforementioned conditions, when the mean item discrimination was 0.60 and 1.00. Figures 4.29 and 4.30 display the RMSE results.

When examinee groups had the same ability mean and standard deviation and the discrimination was 0.60, all four equating methods produced approximately the same magnitude of bias near the cut score and across the ability distribution. As the mean ability of the alternative form was increased and the discrimination was held constant, a few trends were observed along the left column of Figure 4.27. First, the bias results for all four equating methods shifted in the negative direction as group differences increased. Second, the Levine and Equipercentile methods produced approximately the same amount of bias near the cut score, and across all deciles, when the ability difference was 0.25, although the Levine method produced positive bias and the Equipercentile method

produced negative bias. Third, when the difficulty difference was 0.50, the Levine method produced the smallest bias results near the cut score and across the ability distribution. Similar bias results were observed when the mean item discrimination of the test was 1.00.

The right columns of Figures 4.27 and 4.28 display results for conditions where the ability standard deviation of the alternative form group was reduced to 0.25. The Levine method produced the most consistent bias results both near the cut score and across the ability distribution for all conditions. No other method produced bias results similar to the Levine method for abilities in the first few deciles of the ability distribution.

RMSE results were essentially the same for all equating methods when ability differences between the groups was 0.00, the ability standard deviation was the same for both groups, and the mean item discrimination was 0.60. As group ability differences increased, the Tucker and Frequency Estimation methods produced much larger RMSE values, specifically when group differences were 0.50. The Equipercentile and Levine methods consistently produced the smallest RMSE results across all three mean ability conditions, particularly near the cut score. The Levine method consistently produced the smallest amount of RMSE across the ability distribution. The RMSE results were similar when the mean discrimination was 1.00.

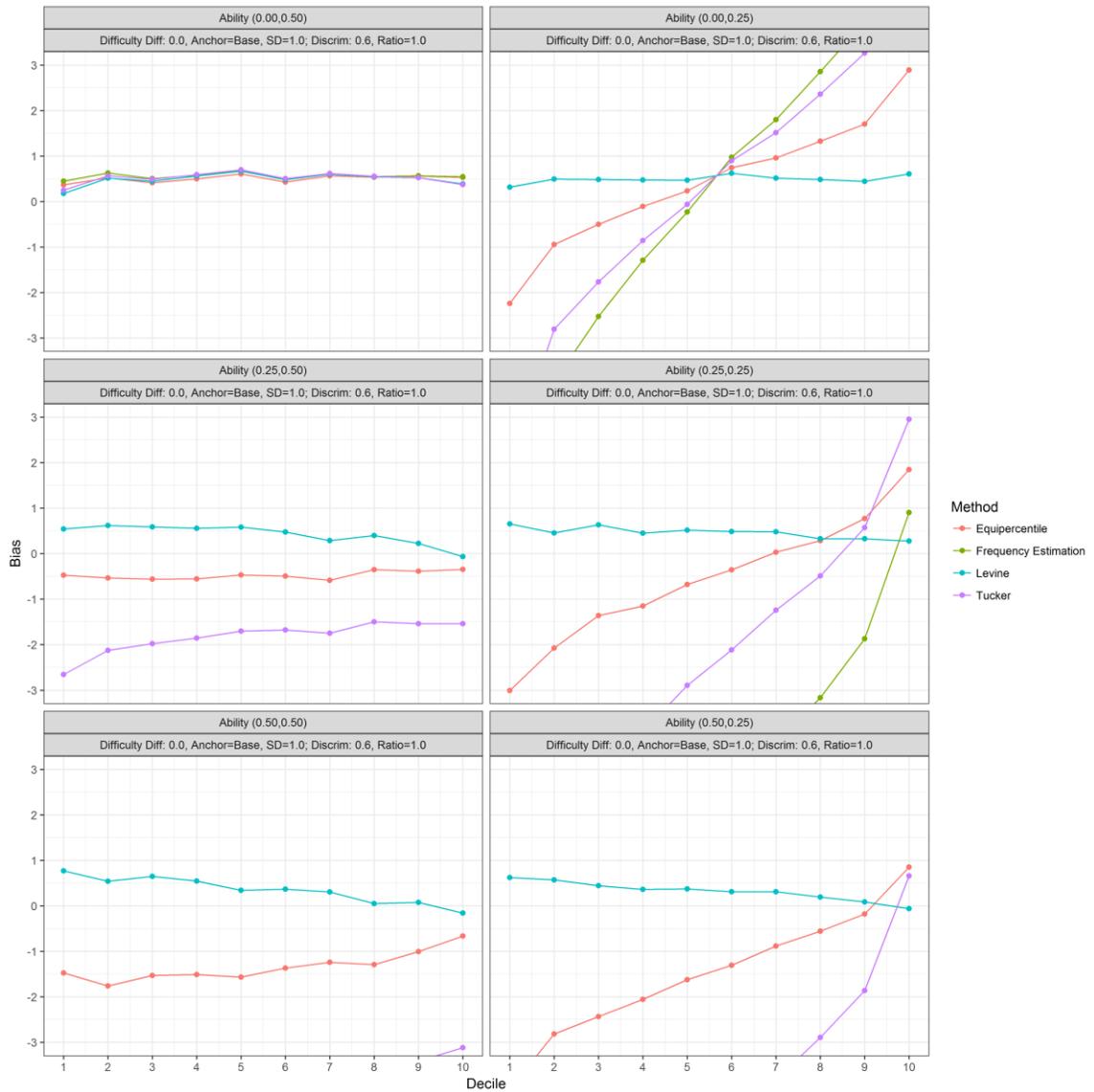


Figure 4.27. Certification Tests: Bias Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 0.60

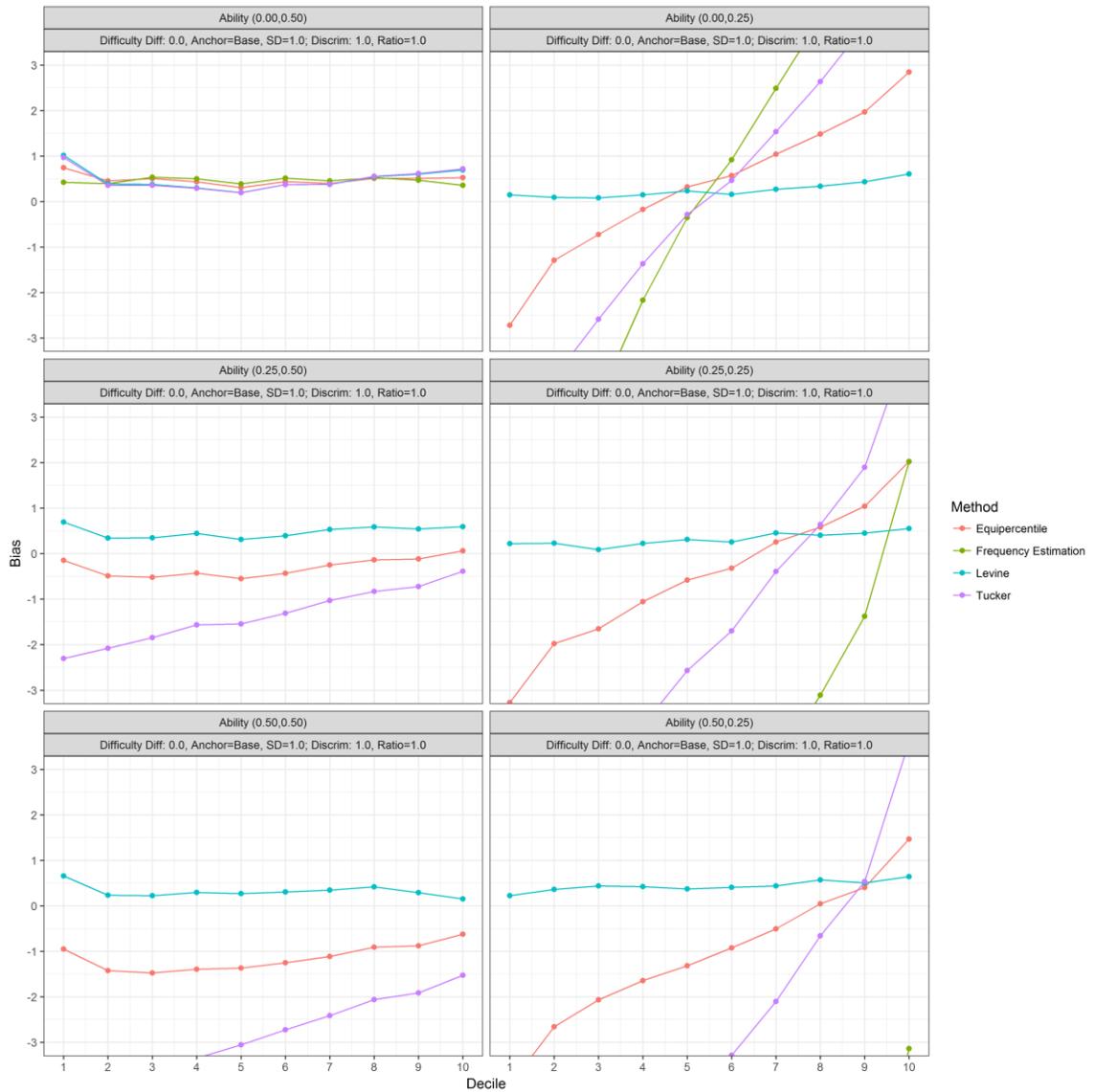


Figure 4.28. Certification Tests: Bias Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 1.00

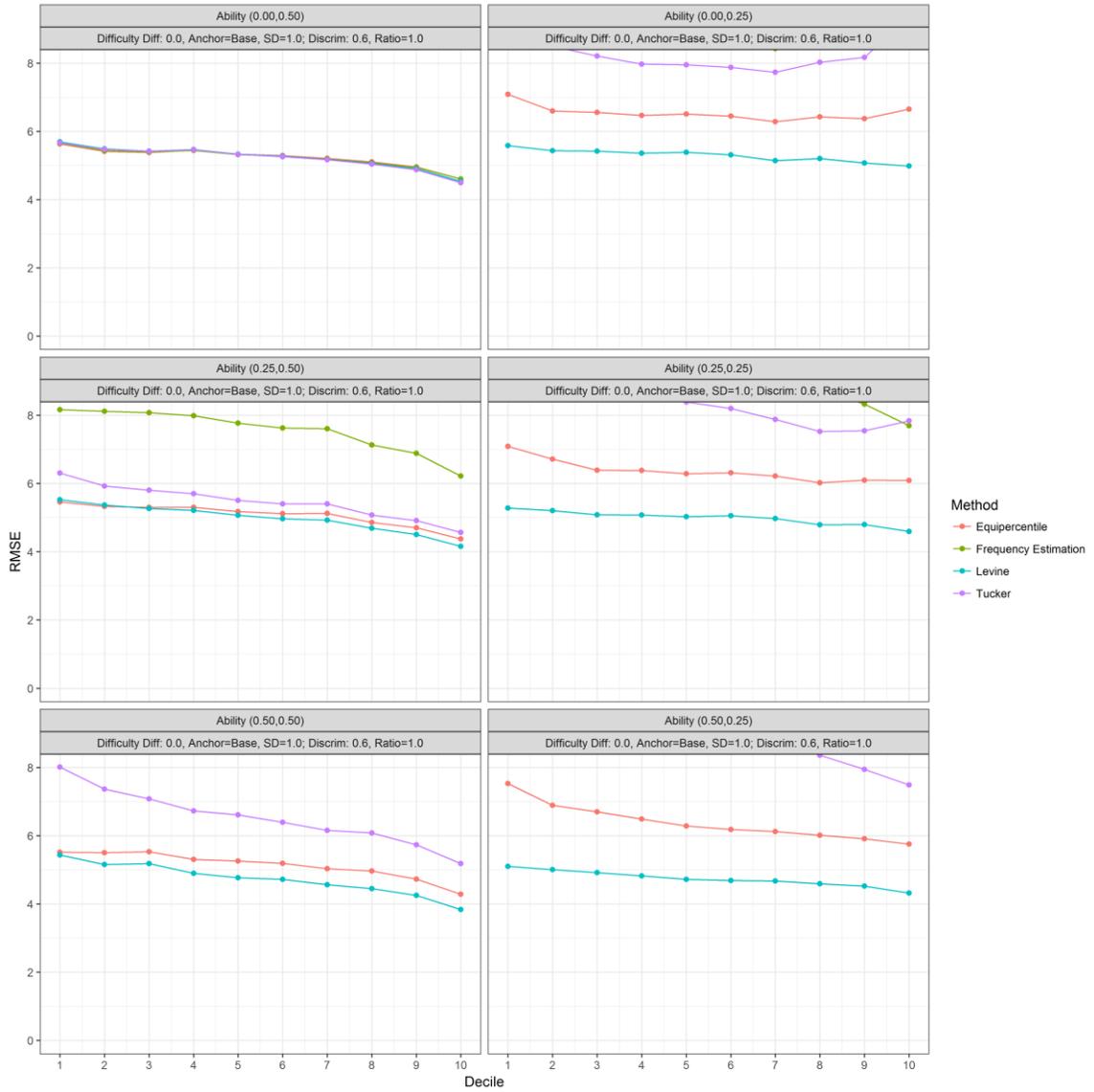


Figure 4.29. Certification Tests: RMSE Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 0.60

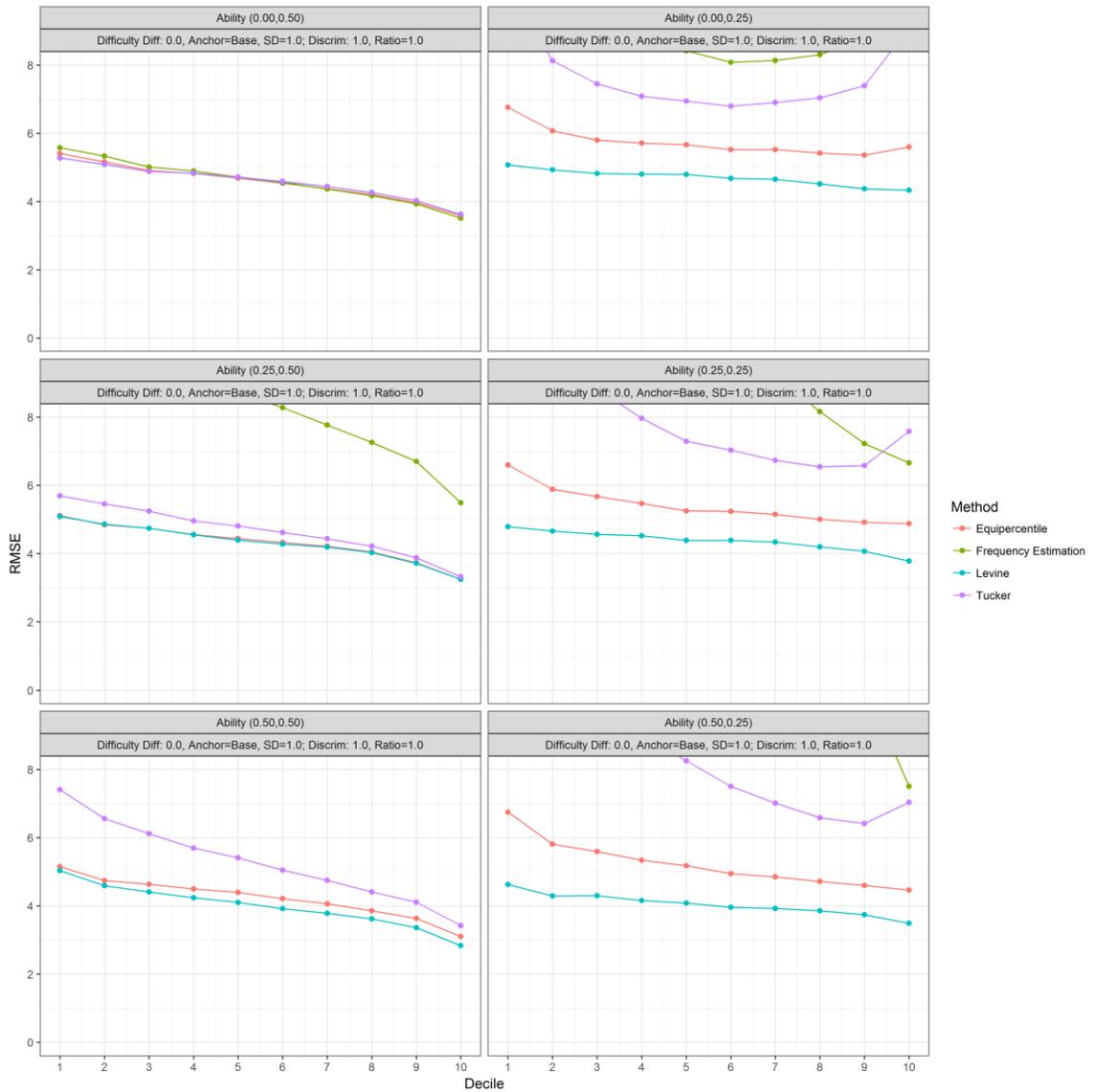


Figure 4.30. Certification Tests: RMSE Results for All Equating Methods and All Ability Conditions when Mean Item Discrimination was 1.00

For the more homogeneous ability condition displayed along the ride side of Figures 4.29 and 4.30, the Levine method produced the lowest and most consistent RMSE results both near the cut score and across the ability distribution. The Tucker and

Frequency Estimation produced much larger RMSE than the Levine and Equipercntile methods, under all conditions.

### **Achievement Condition Interactions**

Within the certification testing scenario there were a number of interaction conditions, which are discussed in this section. First, base and alternative form difficulty differences are discussed within the context of changes in group ability. Then, the results from varying the ability and anchor set conditions are presented. Finally, equating error results from conditions related to form differences, anchor set conditions, and ability differences are discussed.

#### **Form Differences and Examinee Ability Differences**

The interaction between off target exams and ability distributional differences was an important aspect of this study. Three alternative form mean difficulty differences were included, 0.00, 0.25, and 0.50, as well as three mean ability differences of 0.00, 0.25, and 0.50. However, trends for the middle difficulty and ability conditions, 0.25, are not presented, as they were similar, but smaller in magnitude, than the results when differences were 0.50. Figures 4.31 and 4.32 present bias results for test form mean difficulty differences of 0.00 and 0.50 with examinee mean abilities of 0.00 and 0.50, and mean item discriminations of 0.60 and 1.00, respectively. The top two rows of each figure present results when the alternative form ability standard deviation was 0.50, while the bottom two rows include results when the standard deviation was 0.25.

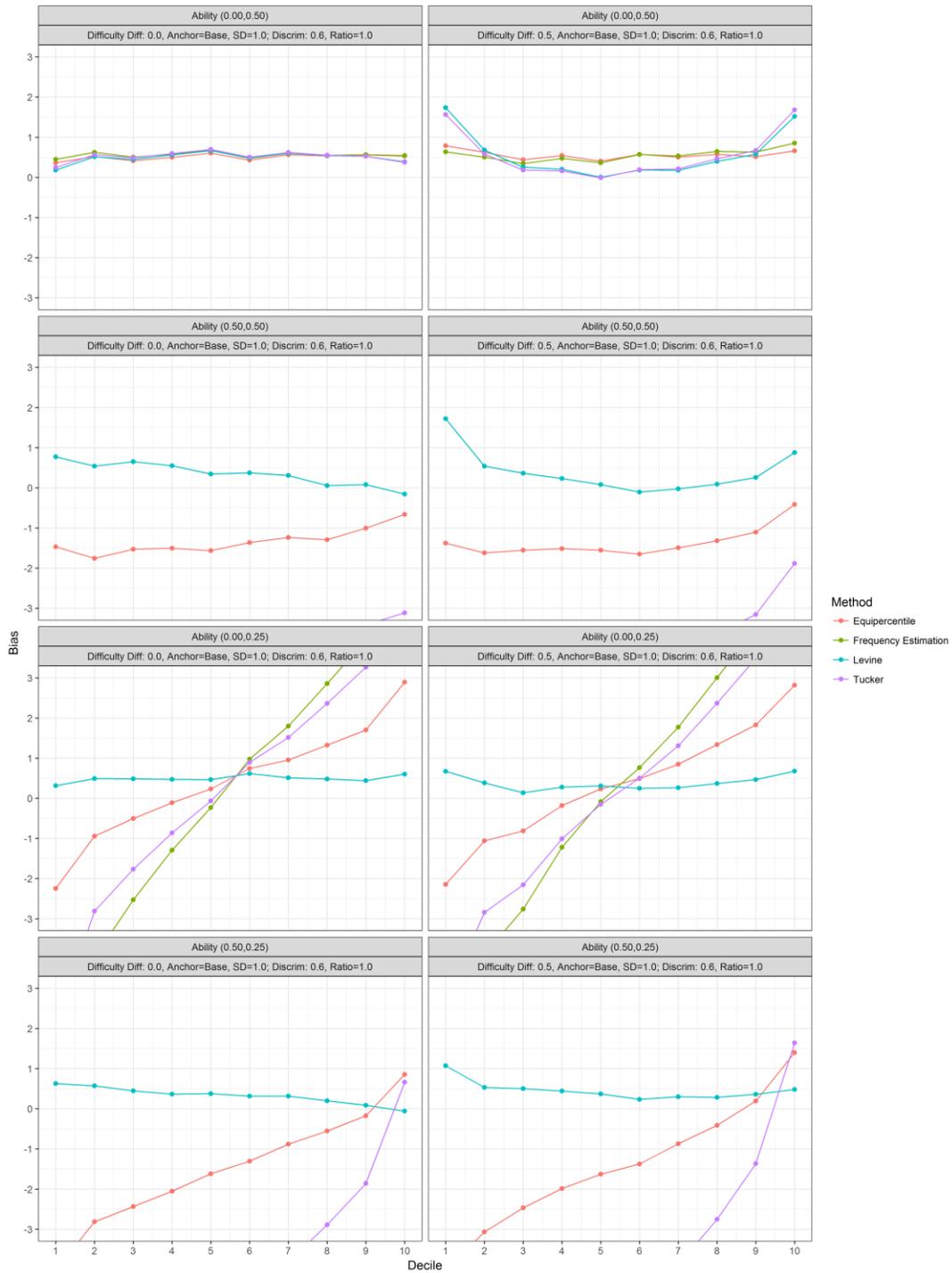


Figure 4.31. Certification Tests: Bias Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 0.60

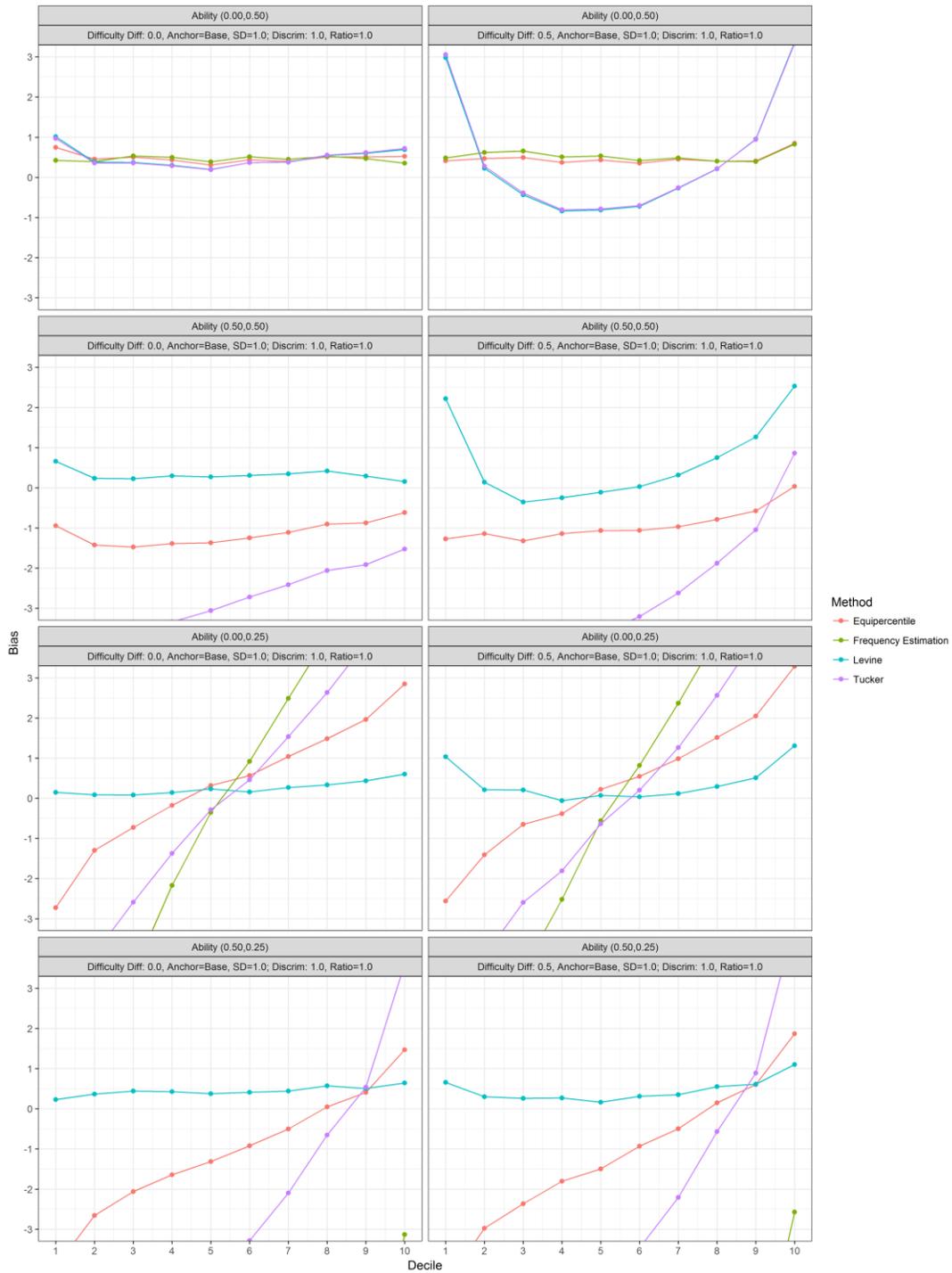


Figure 4.32. Certification Tests: Bias Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 1.00

When the alternative form examinee ability standard deviation was 0.50, test form differences influenced bias results for the linear equating methods near the cut score. Larger differences were observed when the discrimination was 1.00, compared to 0.60. Bias differences were larger than the DTM threshold for linear methods in the first decile when form difficulty was increased. Under the same conditions, differences in mean ability impacted the bias results for the nonlinear methods. It is important to acknowledge that nonlinear methods were less influenced by the discrimination of the test.

When both test form difficulty and mean ability differences were large, the Levine and Equipercentile methods produced similar bias near the cut score, and across the ability distribution, when the discrimination was 0.60. For the higher discrimination condition, the Equipercentile method produced less bias near the cut score than the Levine method, by an amount larger than the DTM threshold.

For conditions where the ability was more homogeneous, the patterns were much different for the Tucker, Equipercentile, and Frequency Estimation methods, while the Levine method remained stable. The Tucker, Equipercentile, and Frequency Estimation methods produced a large amount of negative bias near the cut score, near-zero bias near the middle of the ability distribution, and a large amount of positive bias for examinees with high abilities. The bias near the cut score for the Levine method was reduced for the 0.25 standard deviation conditions, and the bias magnitude and direction was consistent across the entire ability distribution.

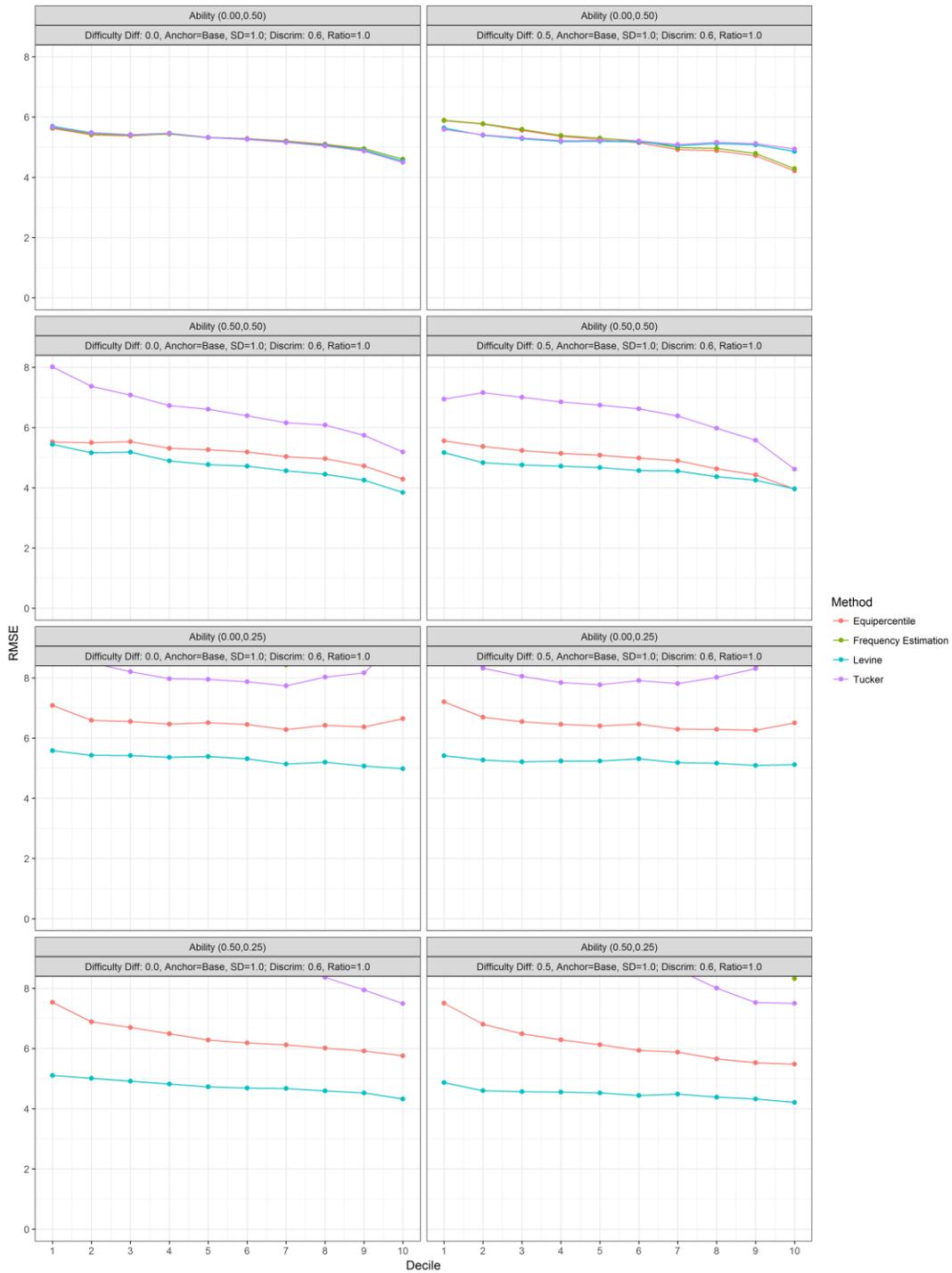


Figure 4.33. Certification Tests: RMSE Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 0.60

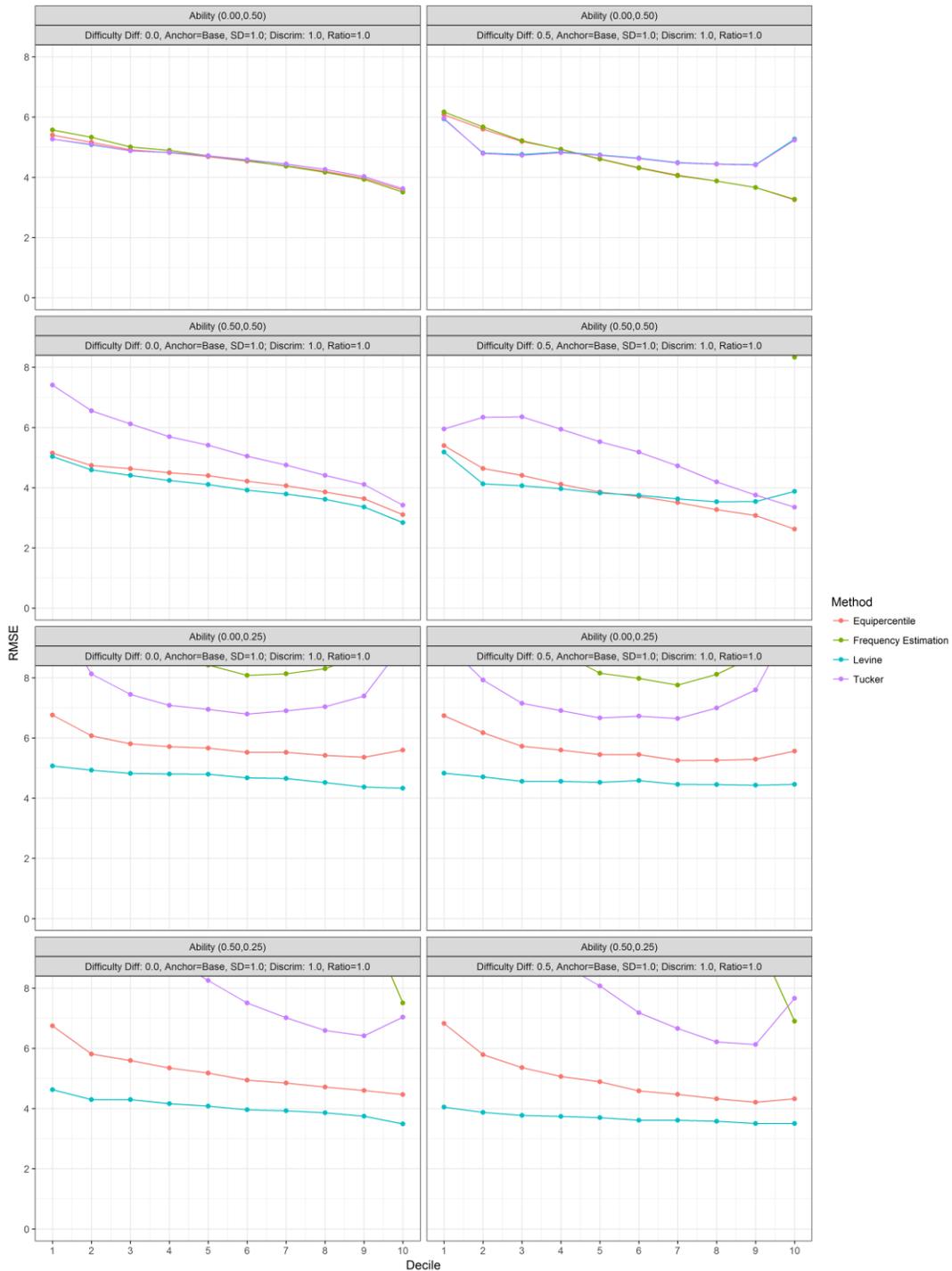


Figure 4.34. Certification Tests: RMSE Results for Form and Ability Differences of 0.00 and 0.50, All Equating Methods, when Mean Item Discrimination was 1.00

Figures 4.33 and 4.34 display RMSE results for the interaction between form difficulty and examinee ability differences. Figures 4.33 and 4.44 present RMSE results for both discrimination conditions, 0.60 and 1.00, respectively. In both figures, the top and bottom rows present results when the ability standard deviations were 0.50 and 0.25, respectively.

The Levine method produced the smallest RMSE near the cut score for all conditions. As form difficulty differences increased and mean abilities became more able and more homogeneous, the RMSE differences between the Levine method and the other three methods increased. Under some conditions the Equipercentile method produced the best results in some locations of the ability the distribution. However, near the cut score the Levine method always produced the smallest RMSE.

### **Anchor Differences and Examinee Ability Differences**

Another aspect of this study was to examine the relationship between anchor set specifications when group ability differences are observed. This section summarizes the equating error results after manipulating the two conditions. Figures 4.35 and 4.36 present the bias results for the linear equating methods, and 4.37 and 4.38 display the bias results for the nonlinear methods, for the 0.60 and 1.00 discrimination conditions, respectively.

For the Levine method, the most consistent anchor construction method overall was the mini. However, under some conditions other anchor types produced similar, and sometimes slightly less, bias near the cut score. For situations where the ability difference was 0.50, an anchor with increased difficulty performed either similarly, or slightly better

than the traditional mini anchor. The trend was true for both ability standard deviation conditions. For scenarios where the alternative form group was more homogeneous, but the mean abilities for the groups were the same, using a midi and with increased difficulty produced similar, or slightly less bias than the traditional mini, near the cut score. It is important to note that these differences rarely less than the DTM threshold.

Two anchor types produced the lowest bias results for the Equipercentile method, regardless of the ability standard deviation conditions: 1) a midi anchor with increased discrimination and 2) a midi anchor with increased difficulty and discrimination. The difference between each anchor and the traditional mini anchor near the cut score was greater than the DTM threshold when the alternative form ability distribution was different than the base form group.

The bias results for the Tucker and Frequency equating methods were large for all anchor types under all conditions, with the exception of when both groups had the same ability distributional characteristics.

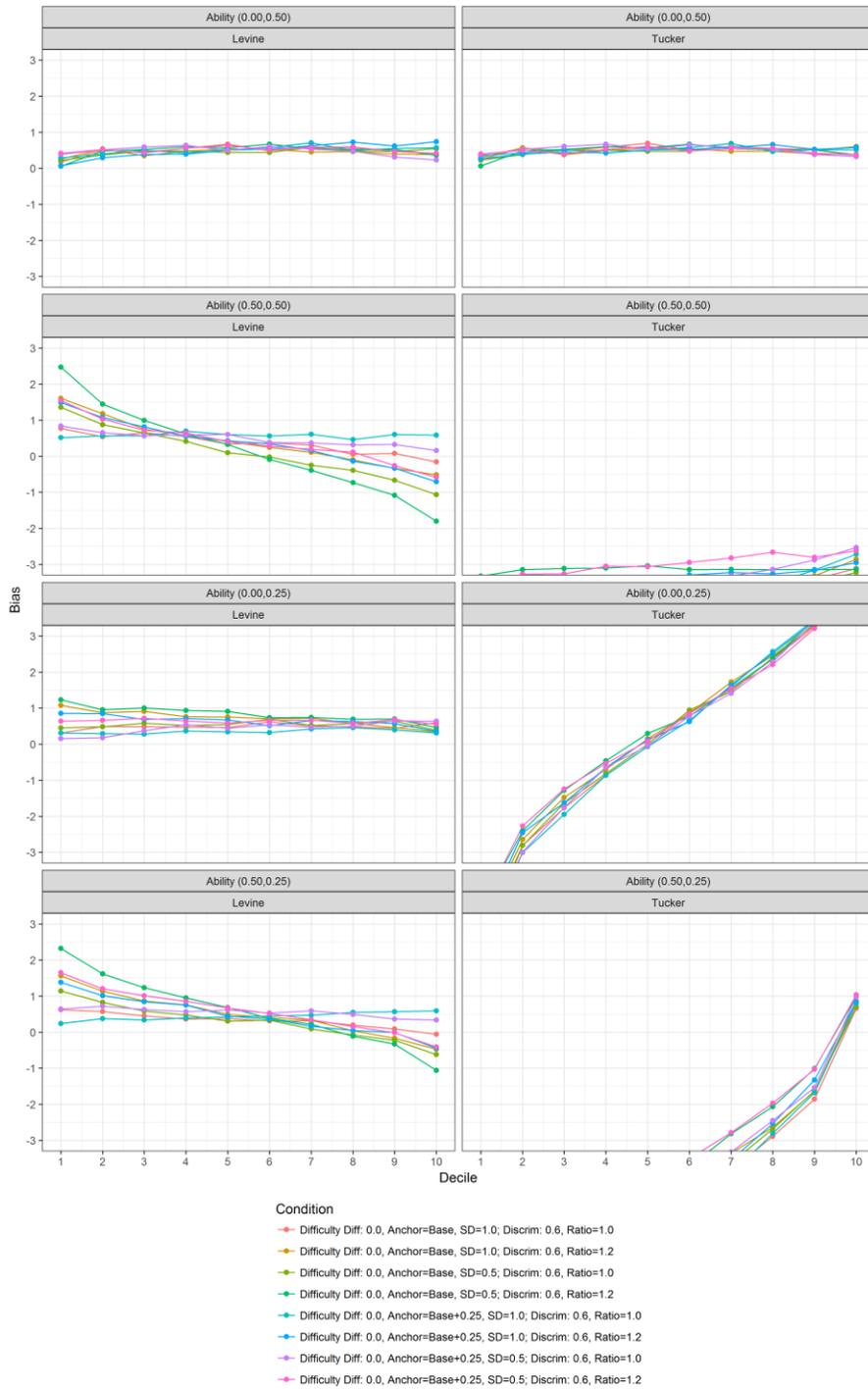


Figure 4.35. Certification Tests: Bias Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 0.60

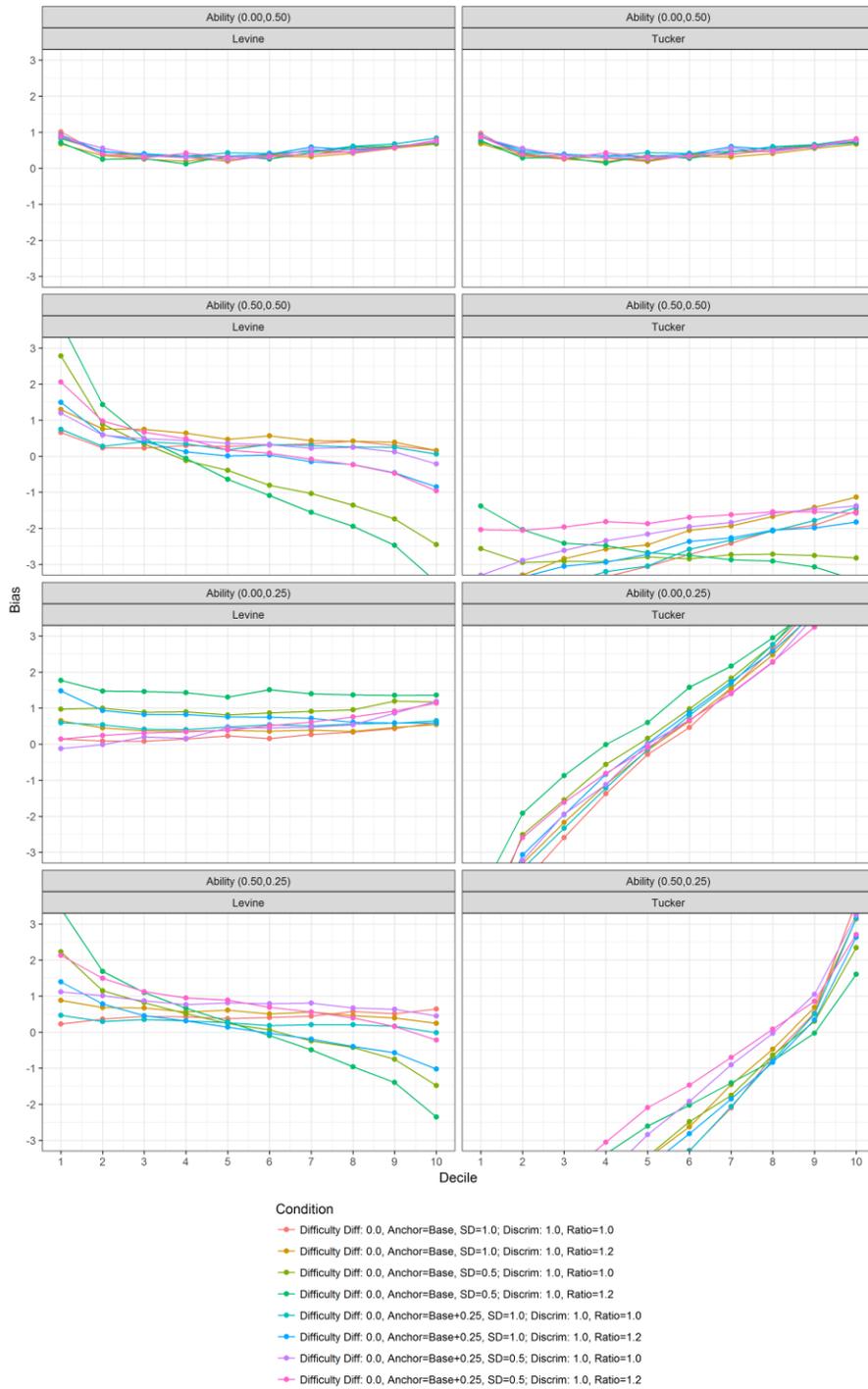


Figure 4.36. Certification Tests: Bias Results for All Anchor Conditions and Selected Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 1.00

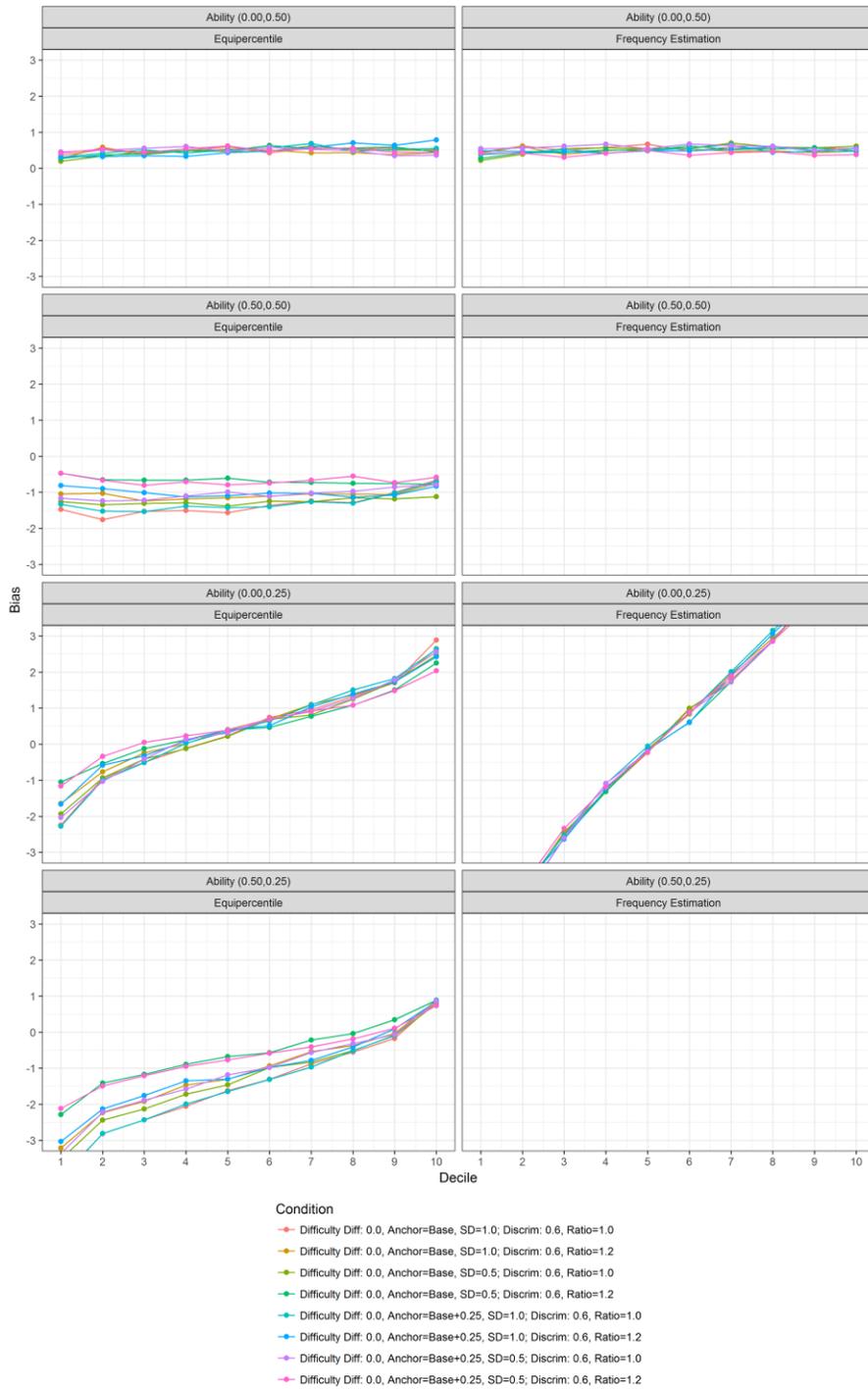


Figure 4.37. Certification Tests: Bias Results for All Anchor Conditions and Selected Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60

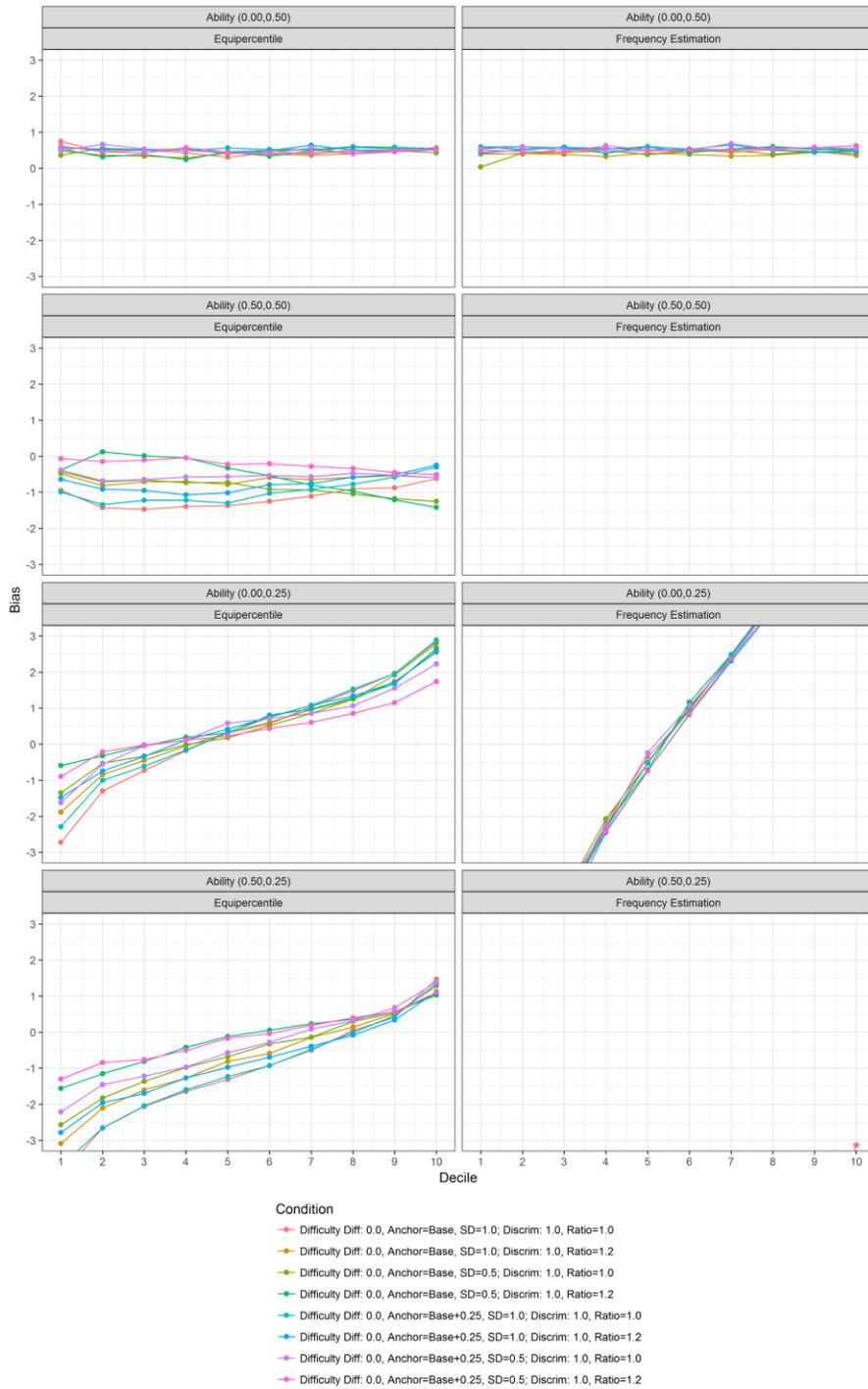


Figure 4.38. Certification Tests: Bias Results for All Anchor Conditions and Selected Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00

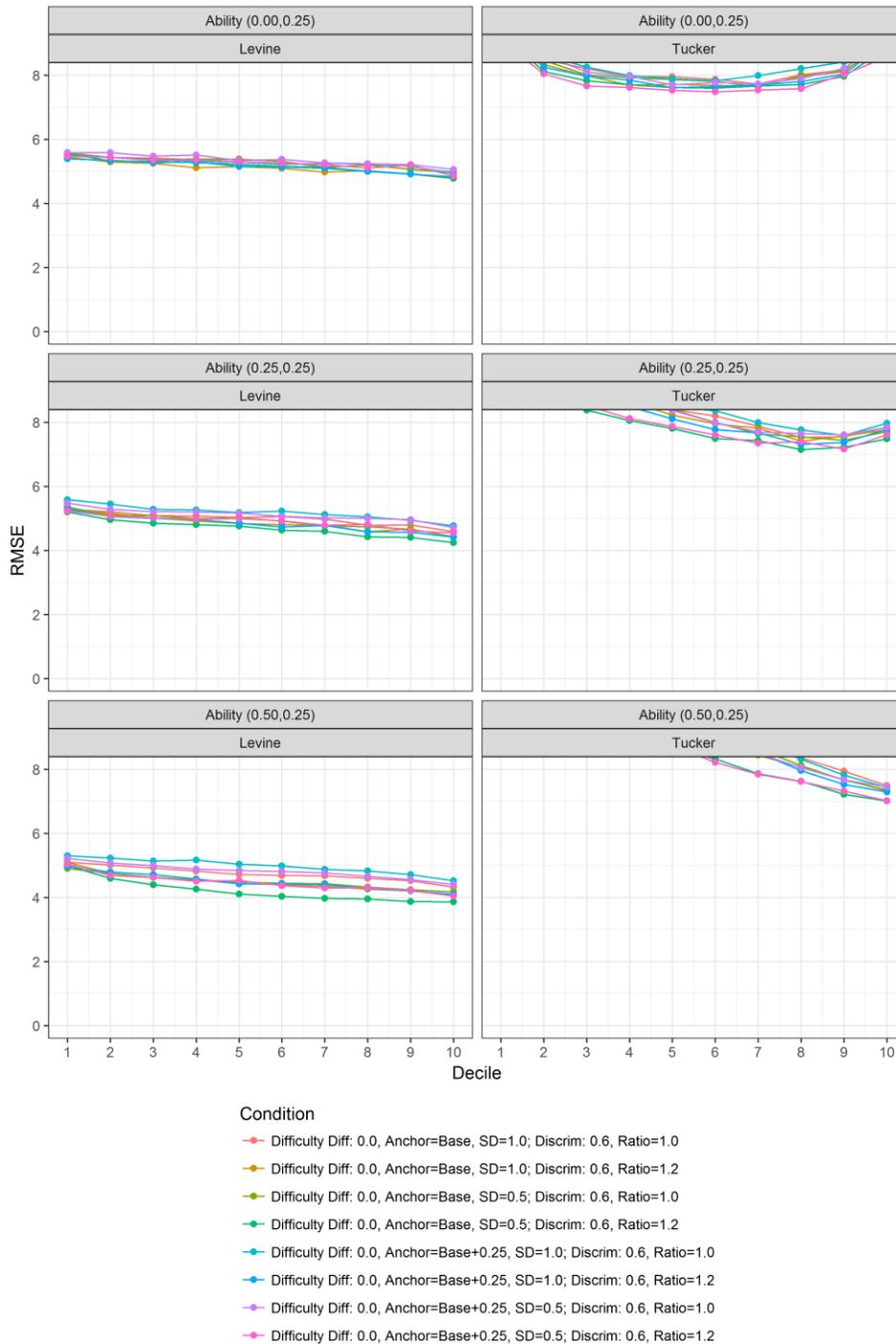


Figure 4.39. Certification Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 0.60

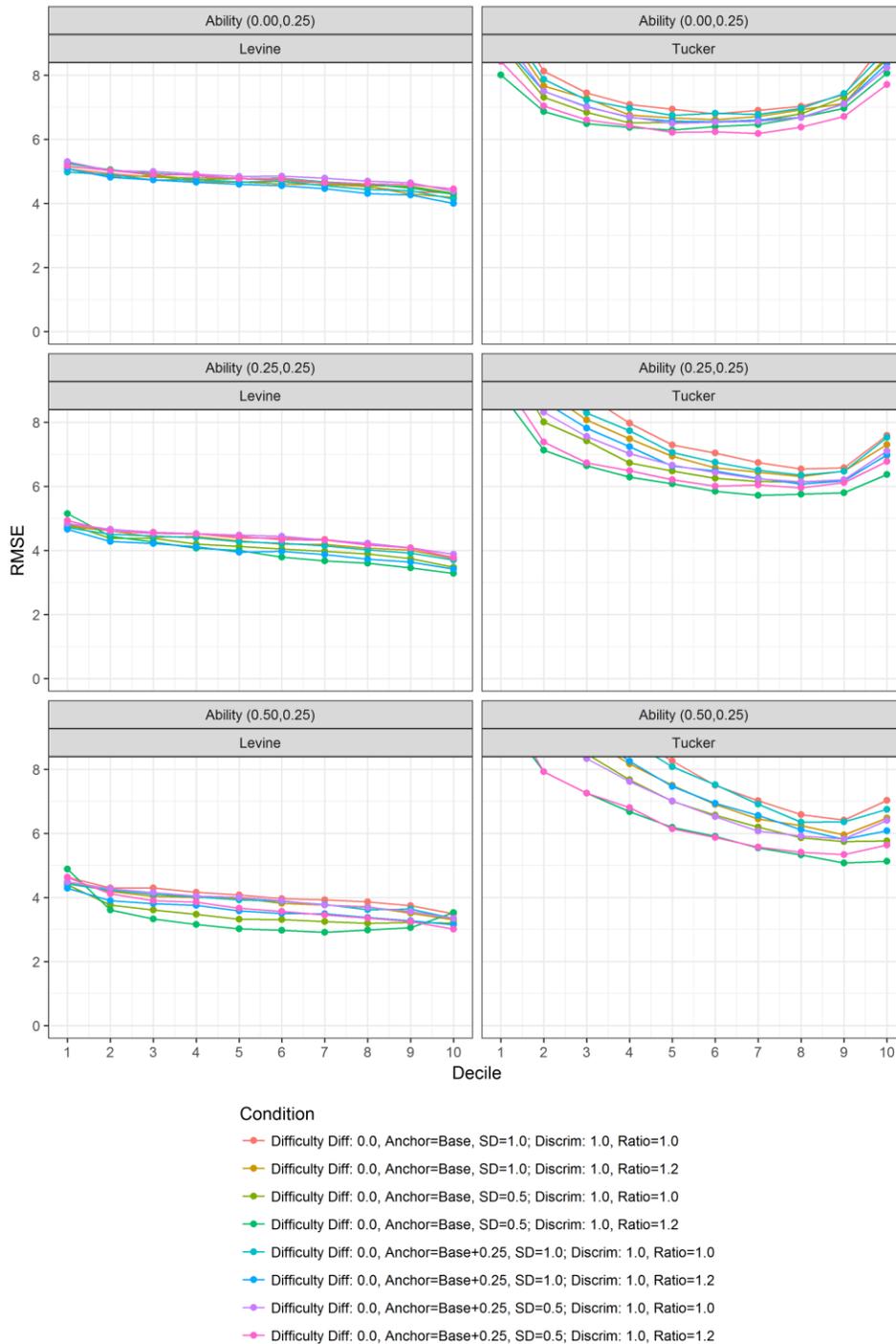


Figure 4.40. Certification Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Linear Equating Methods when the Mean Item Discrimination was 1.00

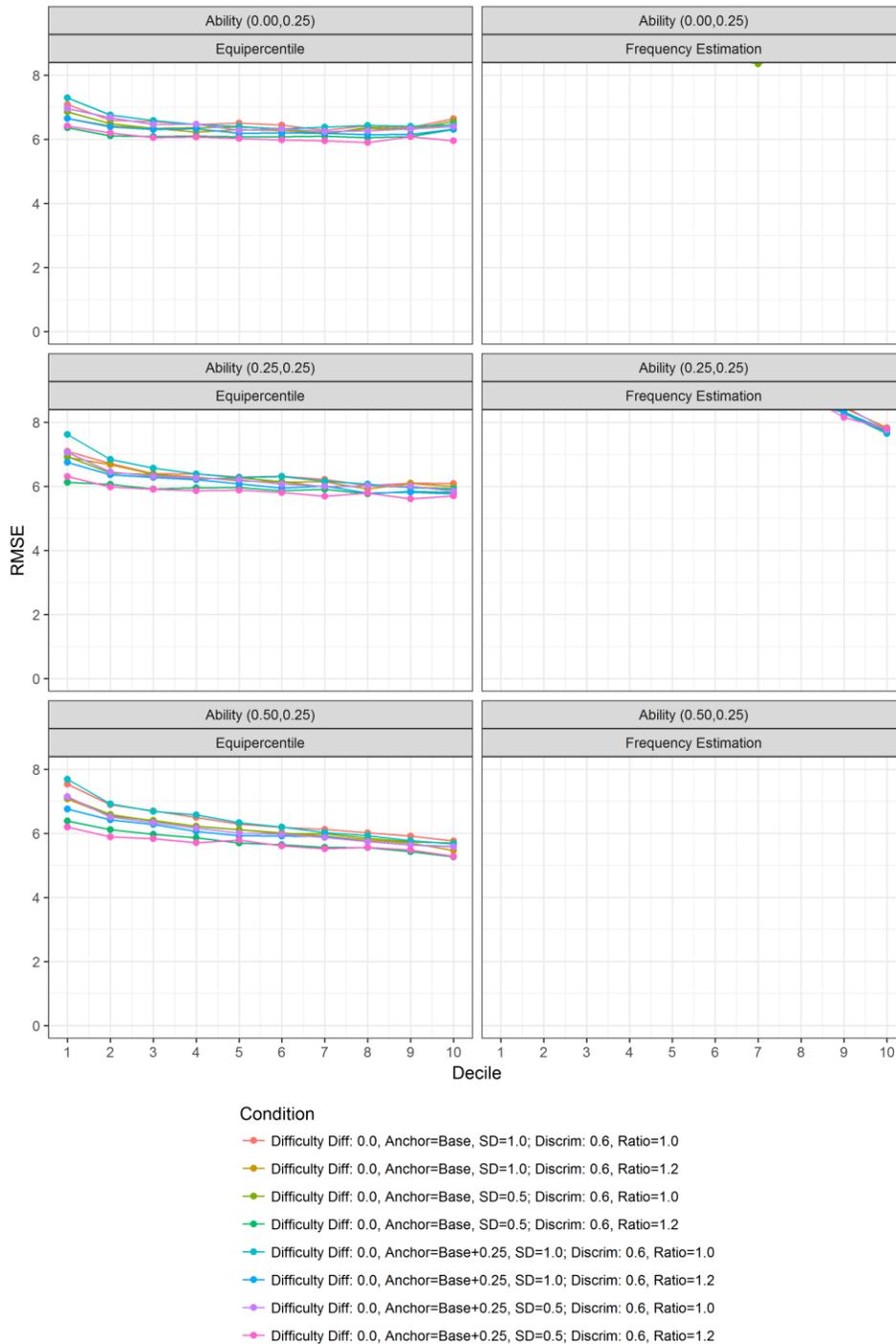


Figure 4.41. Certification Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 0.60

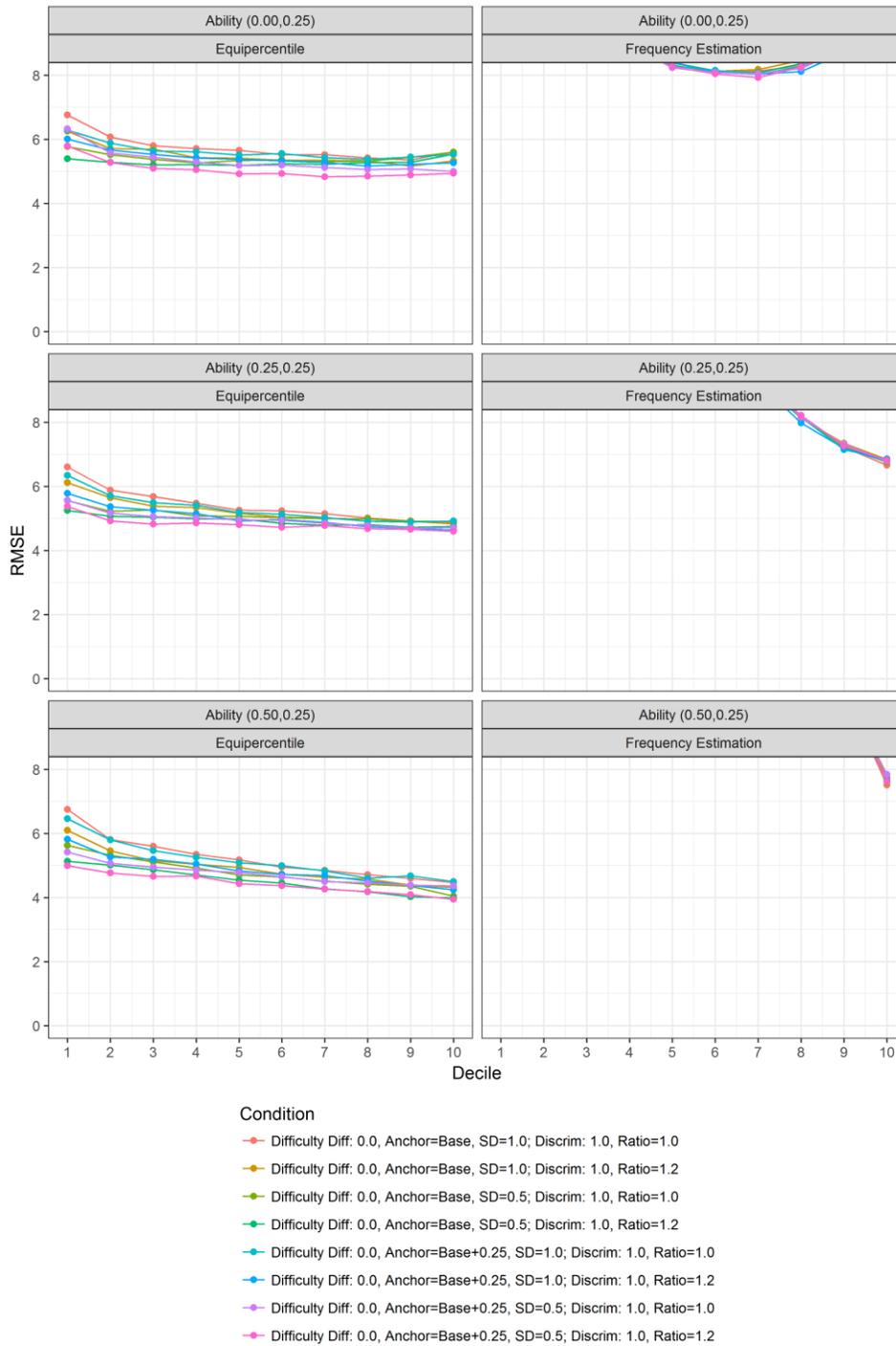


Figure 4.42. Certification Tests: RMSE Results for All Anchor Conditions and All Homogeneous Ability Conditions for Nonlinear Equating Methods when the Mean Item Discrimination was 1.00

The RMSE results suggested that anchor set conditions had little impact on equating error near the cut score when forms had the same difficulty and the groups had the same ability standard deviation. However, more homogeneous groups produced larger RMSE differences when anchor conditions were manipulated. Figures 4.39 and 4.40 present the RSME results for the linear equating methods and Figures 4.41 and 4.42 present RMSE results for the nonlinear methods when the alternative form group was more homogeneous.

The Levine method produced similar RMSE results near the cut score for all anchor set conditions and ability differences when the mean item discrimination was 0.60. Throughout most of the ability distribution, a midi anchor set with increased discrimination produced the smallest RMSE for the Levine method. When the mean item discrimination was 1.00, an anchor set with increased difficulty and discrimination produced slightly lower RMSE compared to the other methods near the cut score, although the difference was small. Across the middle and upper part of the ability distribution, a midi anchor set with increased discrimination produced the smallest RMSE.

The Equipercentile method produced the lowest RMSE near the cut score and across the ability distribution for both discrimination conditions with two types of anchor sets: 1) a midi with increased item discrimination or 2) a midi with increased difficulty and discrimination. The observed trend was true under all conditions where the alternative form group was more homogeneous.

Similar to the large bias results produced by the Tucker and Frequency Estimation methods, the observed RMSE was also much larger than the RMSE produced by other equating methods under the same conditions.

### **Form Differences, Anchor Differences, and Examinee Ability Differences**

This section summarizes the results for scenarios where the alternative form was more difficult, the anchor set conditions were altered, and examinee ability mean and standard deviations were manipulated. Other than when the ability groups had the same mean and standard deviation, the Tucker and Frequency Estimations methods produced results with large amounts of bias. Therefore Figures 4.43 and 4.44 present the bias results for the Levine equating method when the mean item discriminations were 0.60 and 1.00, respectively, and Figures 4.45 and 4.46 display the bias results for the Equipercentile method. Each figure presents results for all form and ability differences for the alternative form group.

For the Levine anchor set conditions, a few trends were observed when the group abilities were different. First, for the low discrimination testing condition, an anchor set with increased difficulty consistently produced the lowest amount of bias near the cut score, as displayed in Figure 4.43. For conditions where the test discrimination was 1.00, a traditional mini anchor set produced the smallest amount of bias, which is displayed in Figure 4.44.

For the Equipercentile method, the most consistent bias near the cut score was almost always observed when using an anchor set that was either: 1) a midi anchor set with increased discrimination or 2) a midi anchor set with increased difficulty and

discrimination. For conditions where the standard deviation of the alternative form group was 0.50 and the mean test discrimination was 1.00, the best anchor set construction condition with respect to bias near the cut score was less clear.

The Tucker and Frequency Estimation methods produced much larger RMSE results than the Levine and Equipercentile methods. Therefore, the Levine and Equipercentile RMSE results are presented in Figures 4.47 and 4.48 for tests with mean item discriminations of 0.60 and 1.00, respectively.

The best anchor set conditions for the Levine and Equipercentile methods were similar to the bias results. Near the cut score, the best anchor set for the Levine method was unclear, although a midi anchor set with increased discrimination tended to produce the smallest RMSE results throughout most of the ability distribution.

The Equipercentile method produced the lowest RMSE results using two types of anchors: 1) a midi anchor with increased mean item discrimination or 2) a midi anchor with increased difficulty and discrimination.

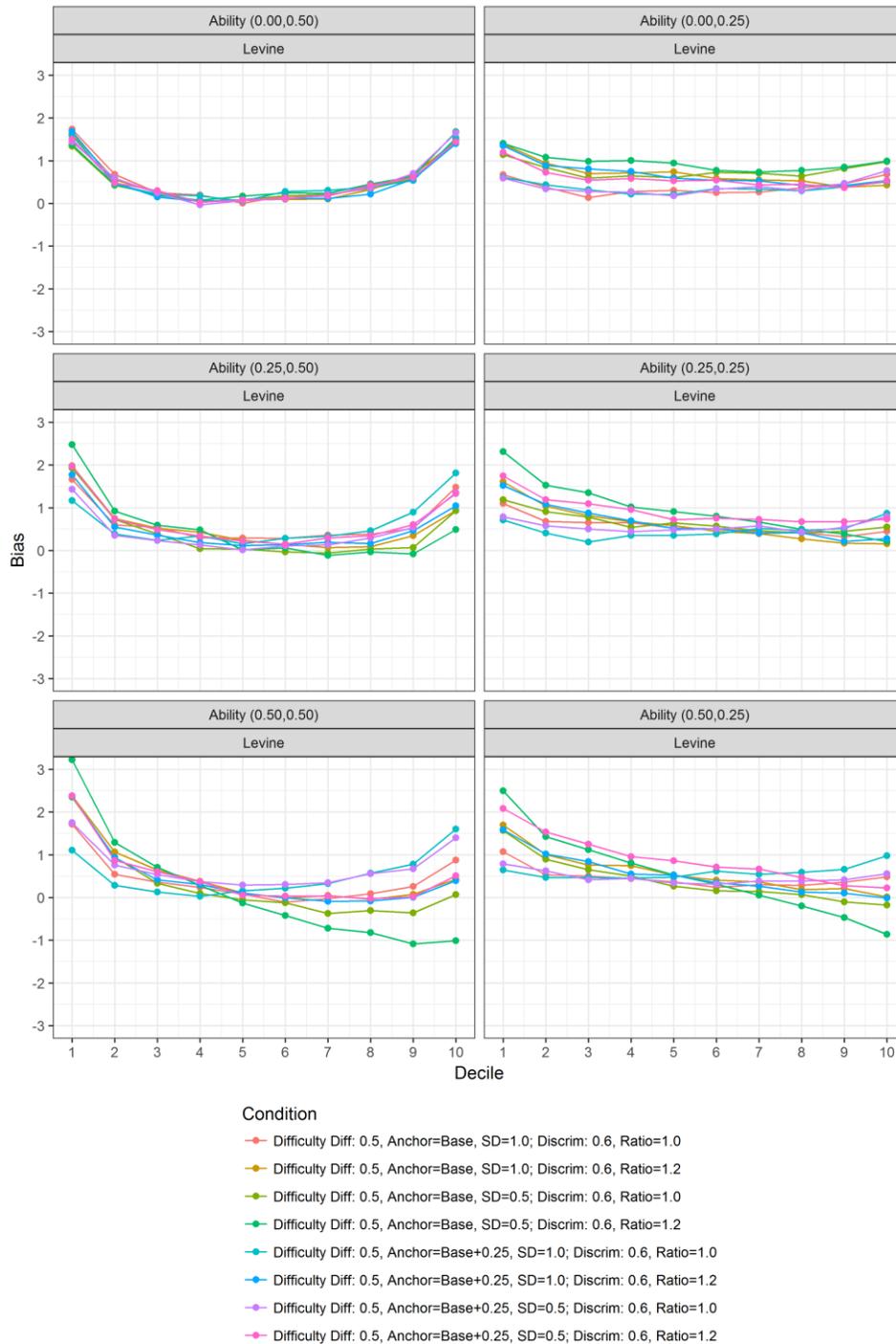


Figure 4.43. Certification Tests: Bias Results for All Anchor Conditions for All Ability Differences for the Levine Equating Method when the Mean Item Discrimination was 0.60

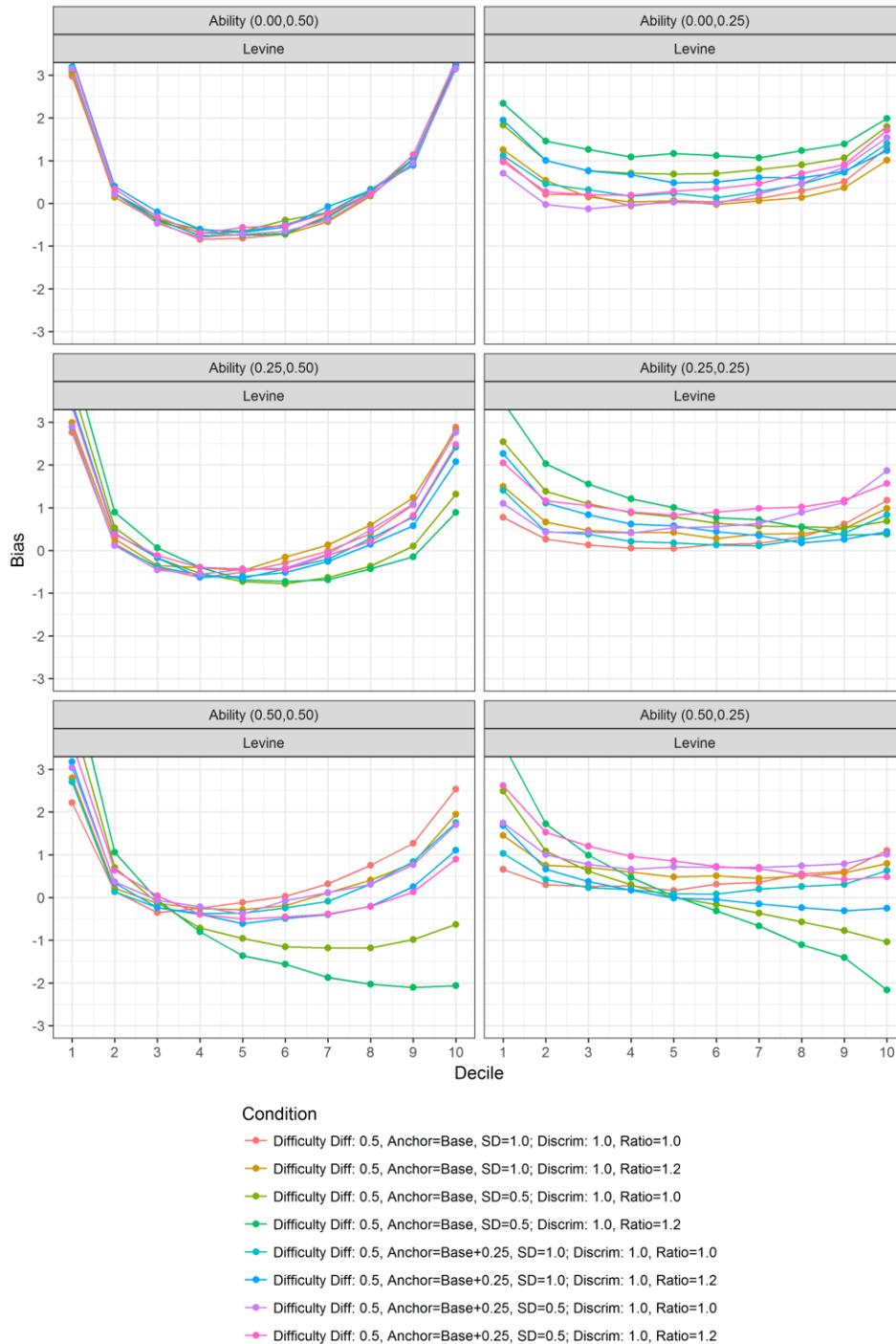


Figure 4.44. Certification Tests: Bias Results for All Anchor Conditions for All Ability Differences for the Levine Equating Method when the Mean Item Discrimination was 1.00

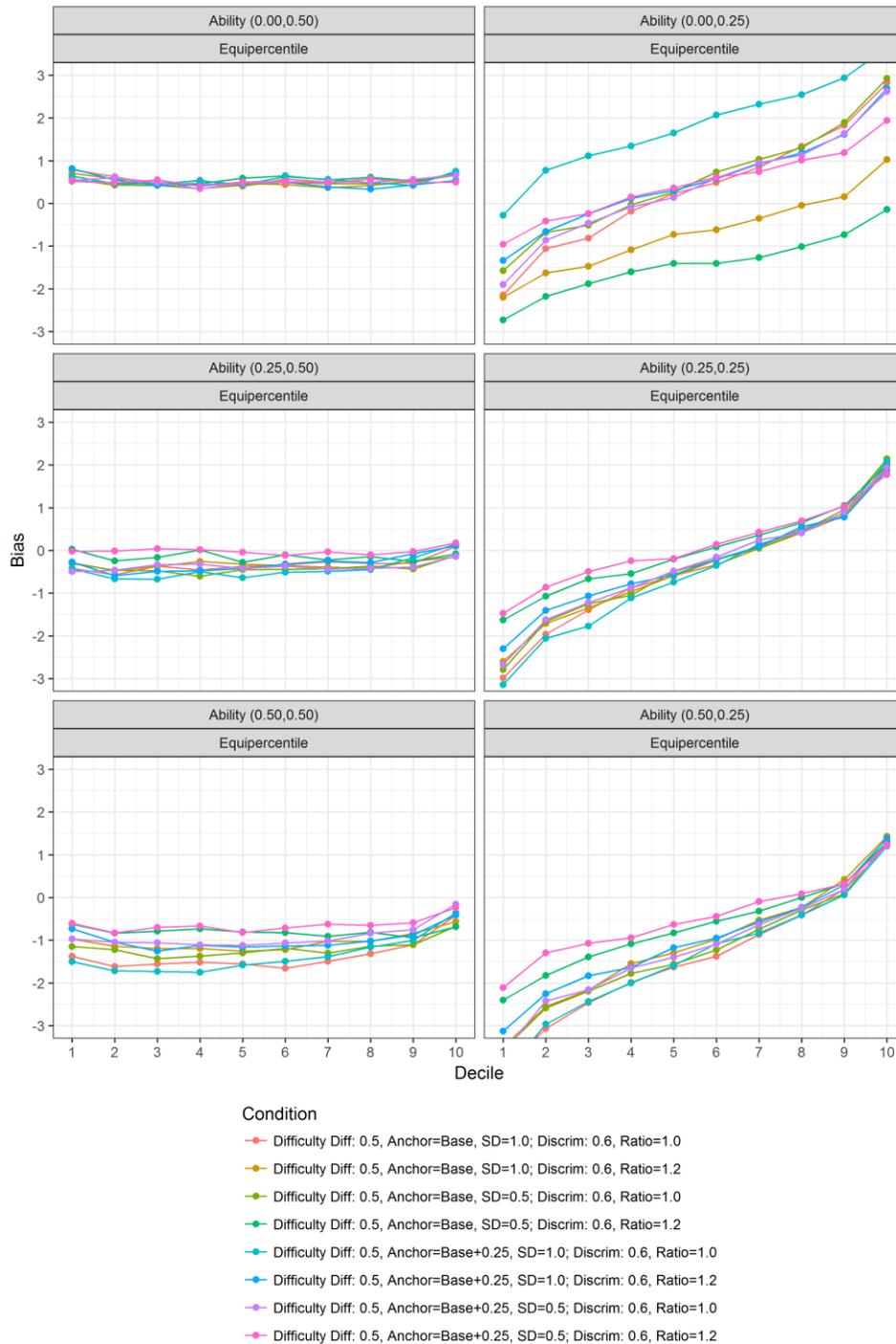


Figure 4.45. Certification Tests: Bias Results for All Anchor Conditions for All Ability Differences for the Equipercntile Equating Method when the Mean Item Discrimination was 0.60

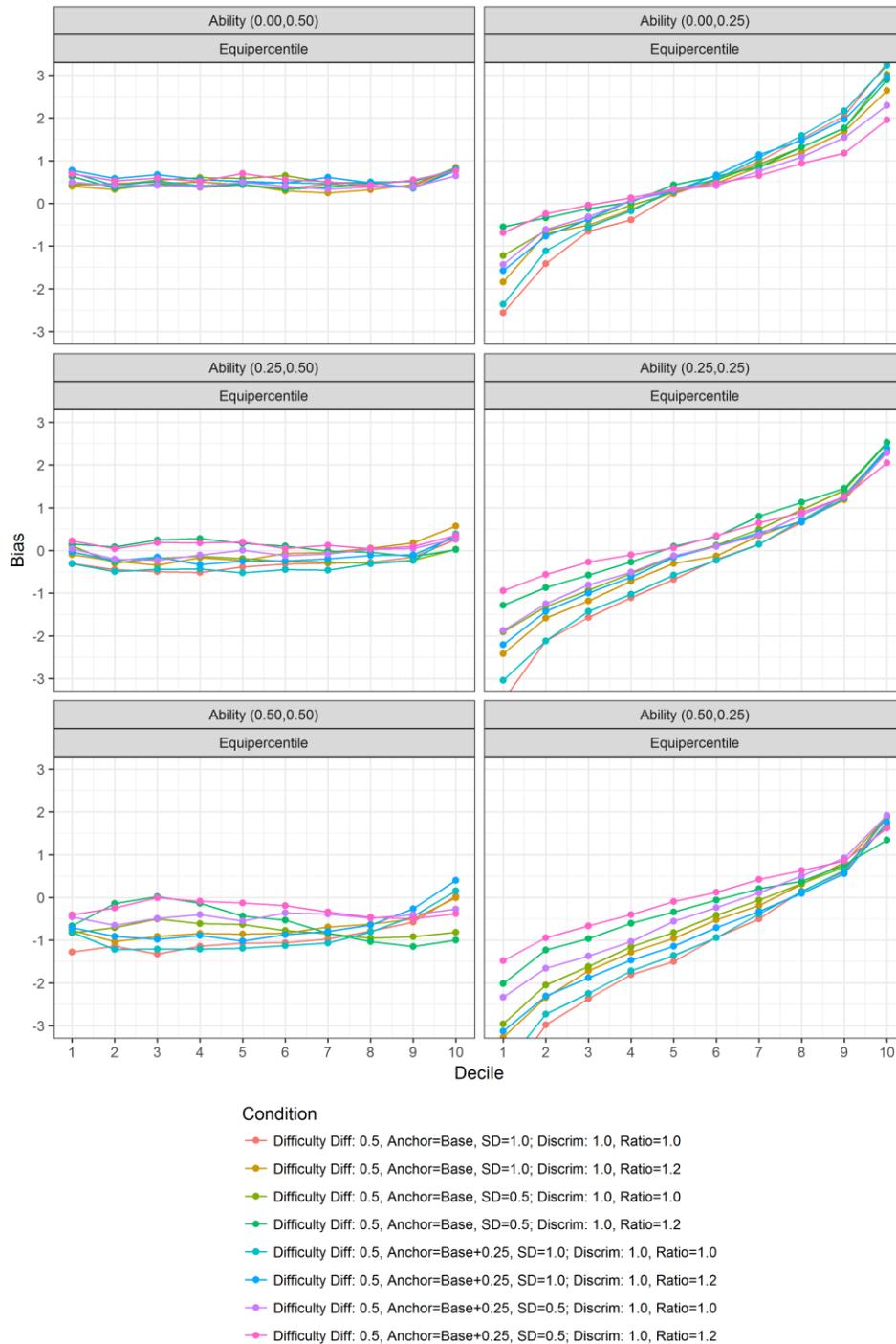


Figure 4.46. Certification Tests: Bias Results for All Anchor Conditions for All Ability Differences for Equipercntile Equating Method when the Mean Item Discrimination was 1.00

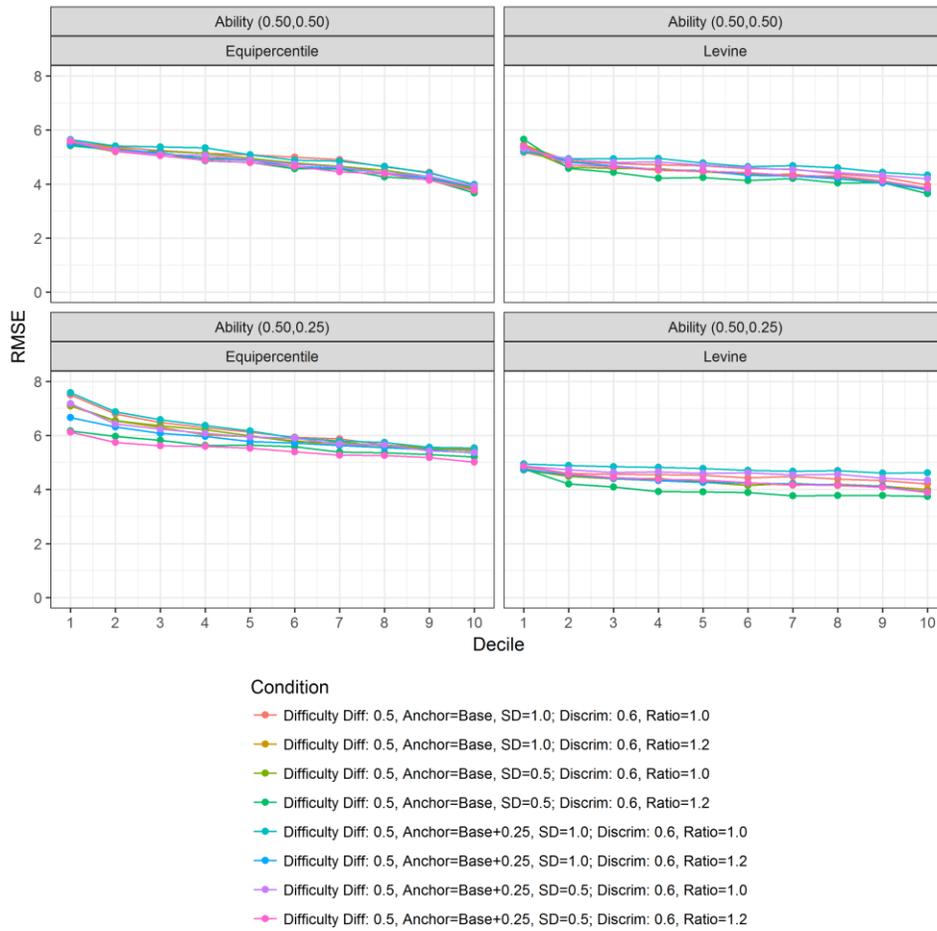


Figure 4.47. Certification Tests: RMSE Results for All Anchor Conditions when Homogeneous Ability Differences were 0.50 for the Levine and Equipercetile Methods when the Mean Item Discrimination was 0.60

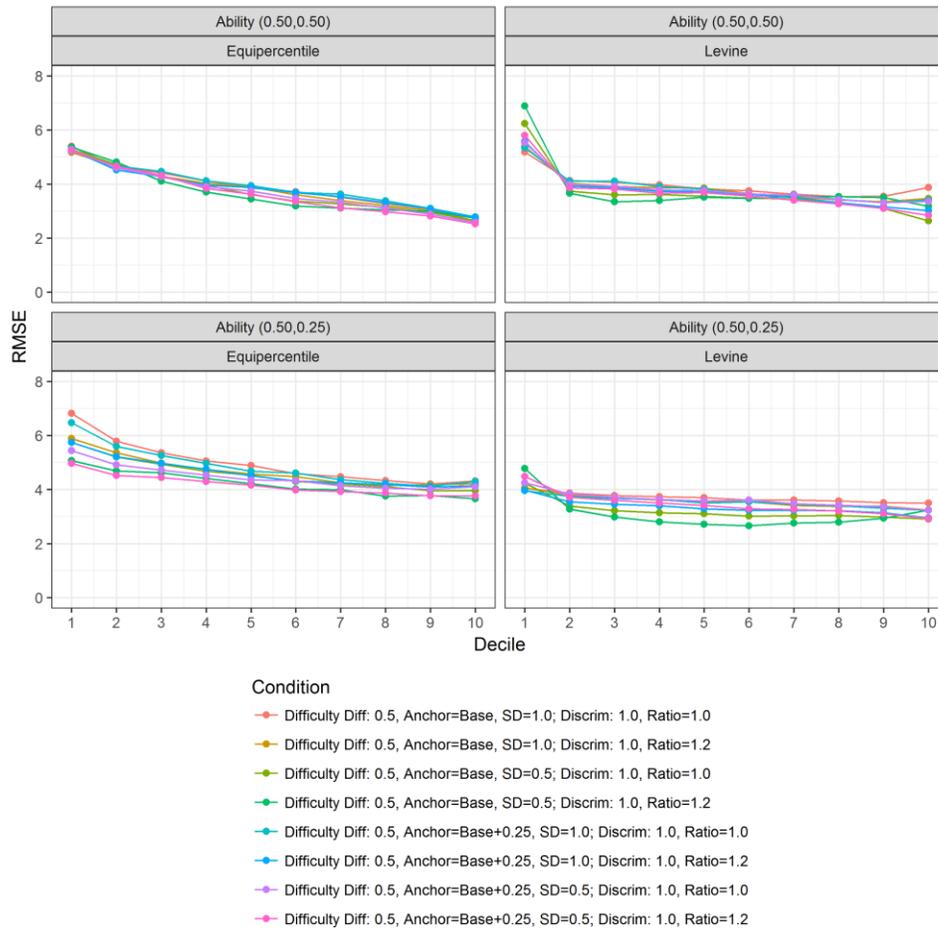


Figure 4.48. Certification Tests: RMSE Results for All Anchor Conditions when Homogeneous Ability Differences were 0.50 for the Levine and Equipercetile Methods when the Mean Item Discrimination was 1.00

## **Certification Summary**

The purpose of this section is to summarize the results to answer the second research question, which reads: with respect to the test purpose and specifications, can anchor set assembly rules be established for linear (Tucker and Levine Observed Score) and nonlinear (Frequency Estimation and Equipercentile) equating methods when differences in group characteristics are expected?

### **Ability Mean Differences and Similar Form Difficulty**

When ability differences between the base and alternative form examinees were large and the forms were similar in difficulty, the Levine method produced the least amount of equating error, both systematic and total. The Equipercentile method produced similar results, although the Levine method consistently produced the smallest bias and RMSE both near the cut score and across the ability distribution.

### **Ability Mean Differences and Dissimilar Form Difficulty**

When ability differences and form difficulty differences were large, the Levine and Equipercentile methods produced the smallest amount of equating error. Near the cut score, the Equipercentile method produced slightly less bias while the Levine method produced the smallest amount of RMSE. Across the overall ability distribution the Levine method produced the smallest amount of bias and RMSE. The equating error produced by the other methods was much larger than the Levine and Equipercentile methods.

### **Ability Mean and Standard Deviation Differences and Similar Form Difficulty**

Substantial equating bias was observed for the Tucker and Frequency Estimation methods when the alternative form examinees were more homogeneous under all mean

ability conditions. The lowest, and most consistent, error results were observed for the Levine method. The Levine method produced the smallest amount of bias and RMSE near the cut score, and for most conditions across the ability distribution. The Equipercentile method was similar under some conditions, but the Levine method was by far the most consistent method. The results suggest that the Tucker and Frequency Estimation methods should not be used to equate test forms under such conditions.

#### **Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty**

The Levine method produced the least amount of equating error, both bias and RMSE, near the cut score when the alternative form ability group was more homogeneous and test forms differed in difficulty. The Levine method also produced the smallest amount of equating error across the ability distribution. The Equipercentile, Tucker, and Frequency Estimation methods produced large amounts of equating error under the aforementioned conditions.

#### **Anchor Set Construction Techniques with Ability Mean Differences and Similar Form Difficulty**

Two observations were made when comparing equating error results for the best two equating methods, Levine and Equipercentile, when the mean abilities were large. The Levine method was the most consistent method across all conditions included in the study, and there was no clear best anchor set for all conditions. For large differences in mean ability, an anchor with the increased difficulty produced similar bias results to the traditional mini, while the RMSE was similar for all anchor designs near the cut score.

With respect to the Equipercentile method, the best two anchors to reduce equating error near the cut score when ability differences were large were: 1) a midi with increased item discrimination or 2) a midi anchor with increased difficulty and discrimination. These two anchor sets improved the Equipercentile method results to match the bias results of the Levine method.

### **Anchor Set Construction Techniques with Ability Mean Differences and Dissimilar Form Difficulty**

The Equipercentile method produced the smallest amount of bias near the cut score when group ability and form difficulty differences existed. Both the Equipercentile and Levine methods produced smaller RMSE near the cut score compared to the other methods, although the best method was unclear. For low discrimination tests the Levine method produced the smallest bias near the cut score with an anchor with increased difficulty, but for high discrimination tests the traditional mini anchor set performed the best with respect to bias. The bias and RMSE results for the Equipercentile method were improved near the cut score by using two types of anchor sets: 1) a midi with increased item discrimination or 2) a midi anchor with increased difficulty and discrimination. The Levine method produced similar RMSE results near the cut score regardless of anchor set specifications.

### **Anchor Set Construction Techniques with Ability Mean and Standard Deviation Differences and Similar Form Difficulty**

The Levine method tended to produce the lowest, and most consistent, equating error results when the alternative form abilities were more homogeneous. Not only did

the method have the most consistent results, there was no clear best anchor set condition near the cut score for all conditions. For tests with lower mean discrimination, increasing the difficulty of the anchor set slightly improved equating bias near the cut score when the alternative forms were more able and more homogeneous. For higher discrimination forms, the traditional mini produced the lowest amount of bias error near the cut score. Total equating error was similar near the cut score for all anchor types under the same conditions.

For the Equipercentile method the two anchors conditions which reduced equating error near the cut score when ability differences were large were: 1) a midi with increased item discrimination or 2) a midi anchor with increased difficulty and discrimination. These two anchor types consistently produced the lowest amount of bias and RMSE for the Equipercentile method, regardless of condition.

### **Anchor Set Construction Techniques with Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty**

When form, anchor, and ability conditions were varied the Levine method produced the smallest equating error results near the cut score and across the distribution of ability. However, there was not a clear best anchor construction method for the Levine method. For low discrimination tests an anchor with increased difficulty produced the best results, while the traditional mini anchor produced the best results for high discrimination tests. Overall, the Levine method appeared to be the most flexible equating method when other conditions were varied.

For the Equipercentile method the two anchor set conditions which reduced equating error near the cut score when the alternative form abilities were more able and more homogeneous were: 1) a midi anchor with increased item discrimination or 2) a midi anchor with increased difficulty and discrimination. These two anchor types consistently produced the lowest equating error for the Equipercentile method across all conditions included in the study.

## **CHAPTER V**

### **CONCLUSIONS AND DISCUSSION**

This chapter presents implications of the major findings of this study, provides test equating and anchor construction rules for practitioners, acknowledges limitations of the study, and suggests future directions for research investigating anchor set construction practices.

#### **Design Considerations**

Equating tests when form and group differences exist is not ideal, particularly when large differences exist, but sometimes is necessary operationally due to a number of factors. This study was designed to emulate reality as closely as possible, and therefore the design was different than other recent studies in a number of ways.

While other studies have calculated equating error by comparing equated scores to a criterion equating function, this study compared equated scores to examinee true scores generated by GENEQUATE (Luecht, 2014). The process was similar to the framework described by Luecht and Ackerman (2018). By computing equating error results with true scores, there was no confounding from the criterion equating method chosen by the researcher, a limitation of the other studies.

Bootstrapping, a commonality among other studies, was also not used in this study. Instead 30 replications, each with different test forms and examinee samples, were

generated under two specific scenarios: achievement and certification testing. For achievement testing, the item difficulty and examinee ability means were aligned on a 60 item test. For certification tests, forms with 150 items were assembled with a mean difficulty that aligned with a pass rate of 90%. These considerations were built into the design to allow it to be as realistic and generalizable as possible for operational psychometricians.

### **Research Question 1**

The purpose of the first research question in this study was to examine equating error that results from interactions between ability, form, and anchor set conditions. Specifically the first research question was: how do examinee ability distributional characteristics, test development specifications, and anchor set properties interact to impact total equating error (root mean squared error) and systematic equating error (bias) when equating with linear (Tucker and Levine Observed Score) and nonlinear (Frequency Estimation and Equipercntile) equating methods under the NEAT design?

There were six important observations in this study related to the interplay between test difficulty and examinee ability, three of which were influenced by specifications of the anchor set. The first three scenarios that are presented are essentially main effects, where only one condition was altered for achievement and certification testing situations. The final three scenarios are true interactions, where at least two aspects of the study were altered with respect to examinee ability, form difficulty, and anchor set specifications.

First, when no differences existed between examinee groups and test forms the choice of equating method did not make a practical difference to systematic or total equating error. Likewise, the anchor set specifications did not impact equating error either. These findings support recent research on relaxing anchor set construction techniques when examinee samples are similar (Fitzpatrick & Skorupski, 2016; Holland, Feigenbaum, & Curley, 2009, 2011; Sinharay, Holland, Curley, & Feigenbaum, 2011).

Second, when test forms differed in mean difficulty and group abilities were the same, the location of the least amount of systematic equating error was dependent upon the equating method used. Nonlinear equating methods produced a consistent amount of positive systematic error across the ability distribution, while linear methods produced a curve with positive bias in the tails of the ability distribution, and near-zero, or negative bias near the middle of the ability distribution. When test discrimination was low, the linear methods produced near-zero bias in the middle of the ability distribution. For high discrimination tests, the curve was more pronounced, with larger amounts of positive bias in the tails of the ability distribution and negative bias in the middle of the distribution. These findings were true across both achievement and certification testing scenarios, and were not influenced by anchor set construction specifications.

With respect to total equating error, the location of the least amount of total equating error was also method dependent. Although, the root mean squared error (RMSE) results were similar for nonlinear and linear methods under the aforementioned conditions, linear methods produced slightly less RMSE at the lower end of the ability distribution and nonlinear methods performed better for higher performing examinees.

Third, when groups differed in mean ability, the results indicated that two specific equating methods, the Levine and Equipercentile methods, consistently produced the lowest systematic and total equating error compared to the Tucker and Frequency Estimation methods. There were some differences with regard to anchor construction conditions for the both methods under achievement and certification testing situations, which are presented later in the guidelines for practitioners.

Fourth, when test form and group differences existed, the Equipercentile and Levine methods produced the smallest amount of systematic and total equating error. The Equipercentile method produced the least amount of equating error when either a midi anchor with increased discrimination, or a midi anchor with increased difficulty and discrimination, was included. It's important to note that the Tucker and Frequency Estimation results were also improved by the same two anchor types in the achievement testing scenario. These findings support previous research that has shown that midi anchor sets produce similar results to traditional mini anchor sets when large ability differences exist between examinee groups (Holland, Feigenbaum, & Curley, 2009, 2011; Sinharay, Holland, Curley, & Feigenbaum, 2011).

Fifth, when the alternative form group of examinees was both more able and more homogeneous, the Levine method produced the smallest amount of systematic and total equating error. The results were improved by using a traditional mini anchor for certification tests with high discrimination, an anchor with increased difficulty for certification tests with low discrimination, and the preferred anchor set construction method was unclear for achievement testing conditions. The finding is consistent with the

research of Trierweiler, Lewis, and Smith (2016) which found that mini anchors produce the highest correlation between the anchor and total test for high discrimination tests.

Sixth, alternative form ability mean and standard deviation differences had a large impact on systematic and total equating error, when combined with test form difficulty. The Levine method was the most consistent equating method for all conditions when the alternative form group was more homogeneous, although the Equipercentile method produced similar results under some achievement testing conditions. Ideal anchor sets for equating with the Levine and Equipercentile methods are provided in the guidelines for practitioners. Equating with the Frequency Estimation or Tucker methods under conditions with a more homogeneous alternative form group produced large amounts of systematic error, regardless of form difficulty alignment, and are not recommended for use under such conditions.

## **Research Question 2**

Practical implications of the results are addressed by answering the second research question, which reads: with respect to the test purpose and specifications, can anchor set assembly rules be established for linear (Tucker and Levine Observed Score) and nonlinear (Frequency Estimation and Equipercentile) equating methods when differences in group characteristics are expected?

Although the second research question focuses on construction guidelines for ability difference conditions, two specific conditions are relevant but do not depend on examinee groups being different, and are included in the guidelines. A synopsis of the

recommendations are provided in Table 5.1 for achievement and certification testing scenarios.

The results of this study suggest that when test forms have the same specifications and examinee groups have the same ability distribution, any equating method will arrive at essentially the same equated results, which is captured in the first row of Table 5.1. Anchor set specifications did not contribute a practical amount of systematic and total equating error to alter scores under the aforementioned conditions. Therefore, the findings confirm recent research on anchor set construction that suggests that the requirement of anchor sets to be a miniature version of the overall test is not warranted under all conditions (Fitzpatrick & Skorupski, 2016; Holland, Feigenbaum, & Curley, 2009, 2011; Sinharay, Holland, Curley, & Feigenbaum, 2011).

When difficulty differences exist between forms, nonlinear equating methods are recommended, as indicated in the second row of Table 5.1. However, it's important to acknowledge that the choice of equating method for many practitioners is limited, at least to some degree, by sample size. The results of this study support using Equipercntile or Frequency Estimation equating methods when differences between forms exist, examinee populations are stable, and sample sizes are large enough.

For programs with smaller sample sizes, it's important to understand where limitations of the linear equating methods exist. Generally, the largest amount of systematic error for the linear methods was observed in the extremes of the distribution of scores, particularly for higher discriminating tests. Although it would likely not present a problem for achievement tests, the lower end of the ability distribution is where

practitioners expect the cut score to fall on certification tests, and where more precision is desired. Therefore, practitioners in certification testing that use linear equating methods should avoid constructing tests which are not built with the same difficulty.

Table 5.1. Guidelines for Practitioners

<u>Conditions</u>			<u>Achievement</u>		<u>Certification</u>	
Alt. Form Mean Difficulty	Alt. Group Ability Mean	Alt. Group Ability SD	Suggested Equating Method	Suggested Anchor Construction	Suggested Equating Method	Suggested Anchor Construction
Equal	Equal	Equal	Any	Any	Any	Any
Harder	Equal	Equal	Nonlinear	Any	Nonlinear	Any
Equal	More Able	Equal	Equipercentile	Any	Levine	Any
Harder	More Able	Equal	Equipercentile	Midi*	Equipercentile	Midi*
Equal	More Able	Less	Levine	Unclear	Levine	<u>Low Discrim.</u> More Difficult
Harder	More Able	Less	<u>Low Discrim.</u> Levine	<u>Low Discrim.</u> More Difficult	<u>Low Discrim.</u> Levine	<u>Low Discrim.</u> More Difficult
			<u>High Discrim.</u> Equipercentile	<u>High Discrim.</u> Midi*	<u>High Discrim.</u> Levine	<u>High Discrim.</u> Mini

\*Two specific types of midi anchor sets produced the best results: 1) a midi with increased item discrimination or 2) a midi anchor with increased difficulty and discrimination

### **Ability Mean Differences and Similar Form Difficulty**

Practitioners should be aware that changes in group ability impact equating error in a major way. Systematic error is the most troublesome for practitioners, as it indicates an introduction of bias into the results. Therefore, the Levine and Equipercentile methods are recommended to practitioners under such conditions, as indicated in the third row of the Table 5.1.

Practitioners in achievement testing that expect a more able group of examinees from one administration to the next are advised to use the Equipercentile method to equate test forms, as it produced the least bias results. The choice of anchor set made little difference in improving the results for the Equipercentile methods under achievement testing conditions, and the finding supports relaxing the mini requirement for the anchor set. For those with smaller sample sizes that must use a linear method, the Levine method produced similar results and is recommended.

For certification testing, the Levine equating method is recommended for practitioners that expect a more able group of alternative form examinees. The Levine method produced the smallest amount of systematic and total equating error near the cut score, followed closely by the Equipercentile method. Although the Levine method always produced the smallest amount of total error, the bias produced by the Equipercentile method was improved to be equivalent to the Levine method near the cut score by using two types of anchors: 1) a midi with increased item discrimination and 2) a midi anchor with increased difficulty and discrimination. The finding is consistent with previous research which has shown that midi anchor sets perform well when large ability

differences exist between examinee samples (Holland, Feigenbaum, & Curley, 2009, 2011; Sinharay, Holland, Curley, & Feigenbaum, 2011).

### **Ability Mean Differences and Dissimilar Form Difficulty**

Practitioners in achievement testing that expect a more able group of examinees from one administration to the next are advised to use the Equipercentile method to equate test forms when form mean difficulty differences are large, as presented in the fourth row of Table 5.1. Likewise, practitioners in certification testing are also advised to use the Equipercentile method when equating under the aforementioned conditions, as it resulted in the smallest amount of systematic and total error near the cut score.

Specifically, in both scenarios the Equipercentile method produced the best results with two types of anchors: 1) a midi anchor set with increased discrimination, or 2) a midi anchor set with increased mean item difficulty and discrimination. These findings support the research of Sinharay and Holland (2006a) and Fitzpatrick and Skorupski (2016), which concluded that midi anchor sets improve equating results for tests that are not well targeted.

### **Ability Mean and Standard Deviation Differences and Similar Form Difficulty**

The Levine method is recommended for achievement testing conditions where the alternative form group is more homogenous and test forms are similar in difficulty, as presented in the fifth row of Table 5.1. The Levine method produced the lowest amount of systematic and total error across the distribution, whether the examinee abilities were the same or different, when the alternative form group was more homogeneous. A number of anchor sets provided similar results in the achievement testing scenario, with

no clear best anchor set construction method standing out across the ability distribution. However, using midi anchor sets with increased difficulty, or midi anchor sets with increased difficulty and discrimination, is not advised for equating tests with higher mean item discrimination under the Levine method. Both anchor types produced much larger systematic and total equating error compared to the other anchor sets.

For certification testing, the Levine method is also recommended when equating test forms under conditions where the alternative form group is more homogenous and test forms are similar in difficulty. The Levine method produced the lowest amount of systematic and total error near the cut score, whether the examinees had similar or dissimilar abilities. The traditional mini anchor set is recommended for use with the Levine method for high discrimination tests, and an anchor set with increased difficulty is recommended for low discrimination tests, as displayed in the fifth row of Table 5.1. The recommendation is consistent with the Trierweiler, Lewis, and Smith (2016) study which demonstrated that for high discrimination tests mini anchors produce the highest correlation between the anchor and total test.

#### **Ability Mean and Standard Deviation Differences and Dissimilar Form Difficulty**

The Levine and Equipercentile methods are recommended for practitioners when equating achievement test forms when alternative form groups are more homogenous and test forms are dissimilar with respect to difficulty, as presented in the sixth row of Table 5.1. The Levine method produced the smallest amount of systematic and total equating error for low discrimination tests, and the results were improved by using an anchor with increased discrimination. Although the results were somewhat mixed for the two equating

methods for higher discrimination tests, the Equipercentile method is recommended due to the improved bias results with two specific anchor set types which created consistent results across the ability distribution: 1) a midi anchor with increased discrimination or 2) a midi anchor with increased difficulty and discrimination.

The Levine method is recommended for practitioners as the only option for equating certification test forms under conditions where the alternative form group is more homogenous and test form difficulties are dissimilar. No other equating method produced similar systematic or total error results near the cut score. Two specific types of anchors produced the best results, which were dependent upon the mean discrimination of the test. For low discrimination tests, increasing the mean difficulty produced the best results, while for high discrimination tests a traditional mini anchor set performed best near the cut score, as indicated in the sixth row of Table 5.1

### **A Note about Some of the Results**

There were two specific results that warrant additional commentary. First, when no differences existed between examinee groups and test forms, a slight positive bias was observed across the ability distribution. The result was likely a product of the way error was calculated in this study, since for error to be calculated a score had to have been observed.

Generally, a small amount of negative bias was observed for scores at the low end of the scale, a small amount of positive bias was observed for scores at the upper end of the scale, and near zero bias was observed for scores in the middle of the scale. Upon further investigation, it was apparent that there were no scores observed at the extreme

low end of the score scale, yet scores were observed at the extreme high end of the scale. The shift towards the upper end of the scale, likely a result of the IRT pseudo guessing parameter included in item generation, is what probably caused the small amount of positive bias across the ability distribution.

Another result that warrants further explanation is the inconsistency in results for the Tucker and Levine methods, particularly when the alternative form group was more homogeneous. Generally, the two equating methods are similar. However, the Levine Observed score method under a classical congeneric model calculates the regression weights using the equations

$$\gamma_1 = \frac{\sigma_1^2(X) + \sigma_1(X, V)}{\sigma_1^2(V) + \sigma_1(X, V)}$$

and

$$\gamma_2 = \frac{\sigma_2^2(Y) + \sigma_2(Y, V)}{\sigma_2^2(V) + \sigma_2(Y, V)}.$$

Because the covariance between the scored and anchor items appeared in both the numerator and denominator, large differences in covariance influenced the results less than they did under the Tucker Method. Under the Tucker method the covariance appeared only in the numerator.

## **Limitations and Future Research**

This section acknowledges limitations of this research and indicates areas where more research is needed.

### **Sample Size**

The chosen sample size for this study was 2,000 examinees. Testing programs come in all sizes, and some may consider 2,000 to be a large sample size while others may consider it to be a rather small sample size. A sample size of 2,000 was chosen for two reasons: 1) nonlinear methods require much larger sample sizes than linear methods, and 2) to include a reasonably large enough sample size to be confident about the results for all conditions included. However, with a sample size as large as 2,000 a testing program could reasonably choose to use item response theory (IRT) equating methods over observed score equating. IRT equating methods were not included because they were beyond the scope of this study. The focus of this study was observed score equating methods, and future studies should consider including smaller sample sizes that would be more realistic for operational observed score equating.

### **Examinee Homogeneity**

The homogeneity conditions included in this study were also a limitation, and impacted the equating error results the most of any condition. Future studies should include ability conditions with standard deviations somewhere between 1.00 and 0.50 for achievement tests, and 0.50 and 0.25 for certification exams, to create more realistic shifts in the examinee population that operational psychometricians might observe.

Although, the results clearly conveyed the concerns that practitioners should be aware of when equating with a more homogeneous population.

### **Presmoothing Methods**

For the nonlinear methods included in the study, loglinear smoothing techniques were implemented to preserve the first five univariate moments and a covariance moment of the distribution. However, it was not possible to conserve the covariance moment for the certification tests, which is a reminder that smoothing introduces error into the results. The smoothing methods were chosen in this study to maintain consistency with the methods used by Sinharay and Holland (2006b, 2007). Another option was to choose the best fitting presmoothing method for each replication, which would have introduced less smoothing error within the replication, but would have led to less consistency across replications.

### **Levine Method**

In this study, the Levine method proved to be the most consistent equating method across the distribution of scores when the alternative form group was more homogenous. Although previous studies have not included the Levine method, the results of this study suggest that future research should include the Levine method when examining test and anchor set construction.

### **True Score Methods**

This study included only linear observed score equating methods for the Tucker and Levine methods. True score methods should be included in future anchor set construction studies for comparison.

## **Content**

This study did not consider the traditional content requirement associated with the mini anchor sets. Rather, for this study content areas were not included in the generation process and only the overall statistical aspects of the test forms were considered.

Therefore, the generalizability of the results are limited to tests without major statistical differences between content areas.

## **Internal Anchor Sets**

The warning of Sinharay and Holland (2006b, 2007) regarding the complexity of manipulating an internal anchor set of items was taken when designing this study, therefore only external anchor sets were included. However, many testing programs use internal anchor sets, particularly those with small item banks which would benefit the most from the ability to relax the anchor construction rules. Internal anchor sets are an important aspect to include in future research.

## REFERENCES

- Albano, A. D. (2017). *equate: An R Package for Observed-Score Linking and Equating*. R package version, 2.0.6.
- American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In RL Thorndike (Ed.), *Educational measurement* (2nd ed., pp.508-600). Washington, DC: American Council on Education.
- Antal, J., Proctor, T. P., & Melican, G. J. (2014). The Effect of Anchor Test Construction on Scale Drift. *Applied Measurement in Education*, 27 (3), 159-172.
- Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic.
- Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement*, 22(1), 13–20.

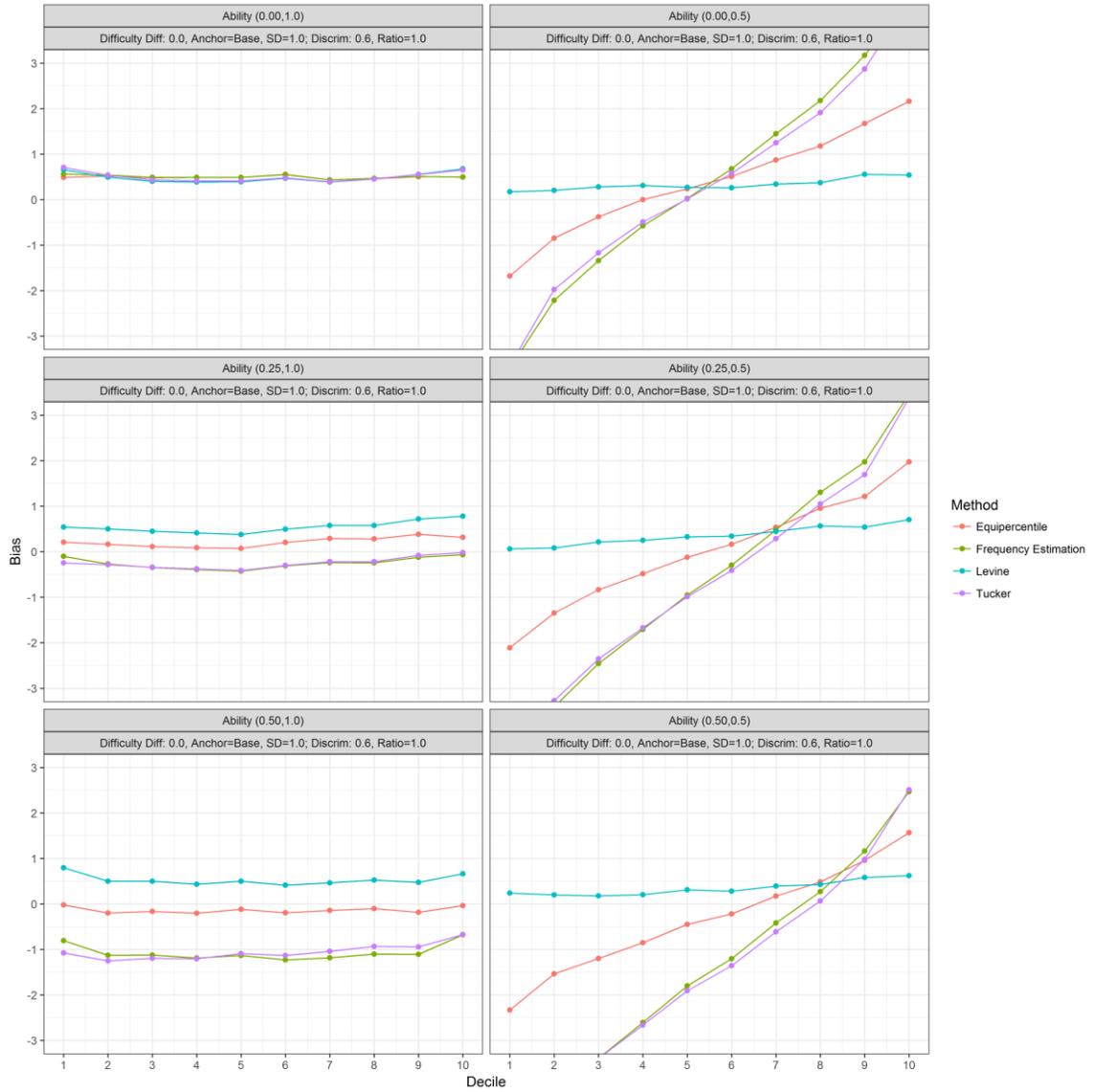
- Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, 11(3), 225-244.
- Cronbach, L. J., & Warrington, W. G. (1952). Efficacy of multiple choice tests as a function of spread of item difficulties. *Psychometrika*, 17, 127-147.
- Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education/Praeger.
- Fitzpatrick, A. R. (2008). NCME 2008 presidential address: The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice*, 27(4), 34-40.
- Fitzpatrick, J., & Skorupski, W. P. (2016). Equating With Midtests Using IRT. *Journal of Educational Measurement*, 53(2), 172-189.

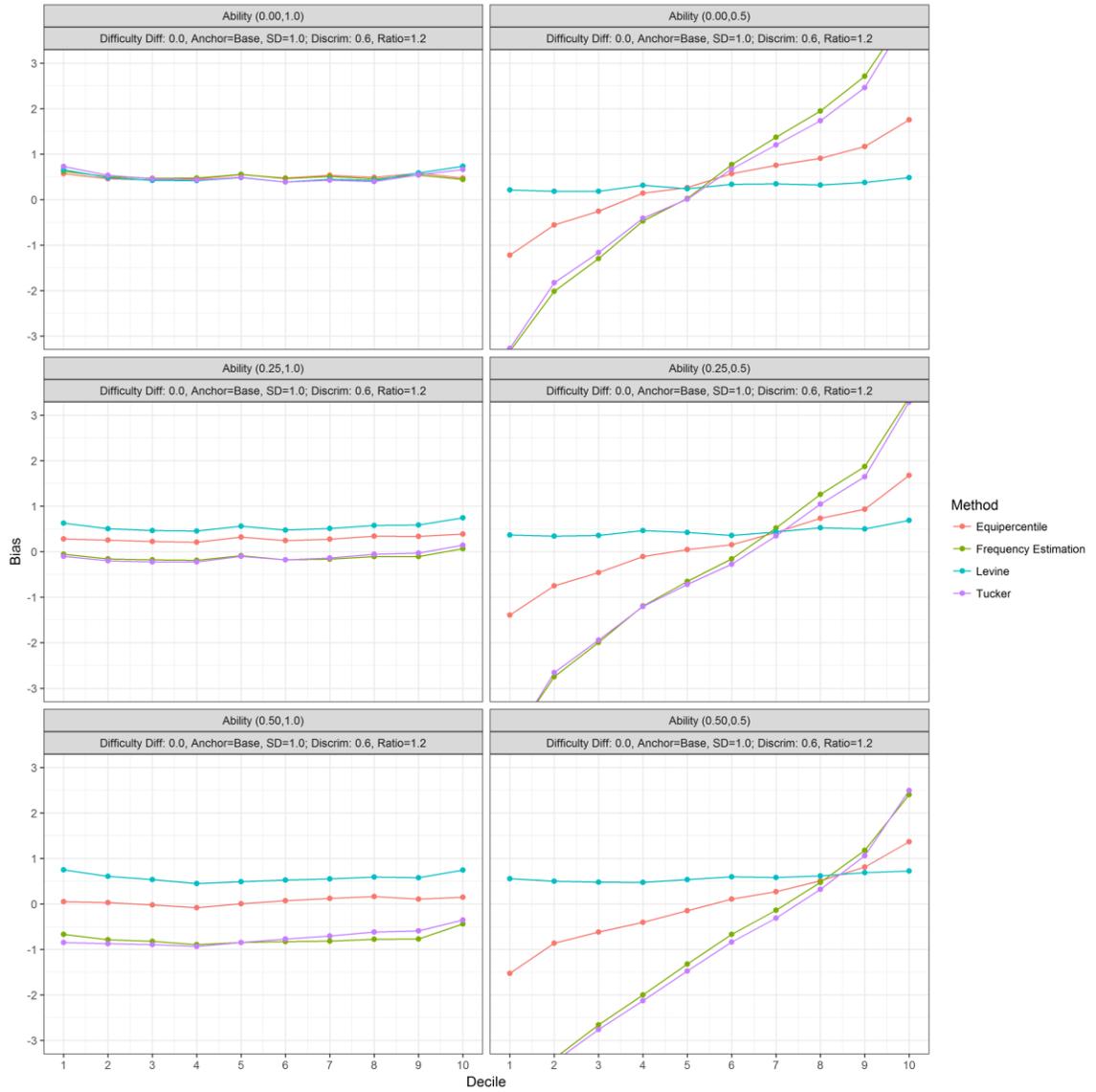
- Gulliksen, H. (1945). The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, 10, 2, 79-91.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, 22(3), 197-206.
- Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice*, 7(4), 29-37.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, linking, and scaling: Methods and Practices* (3rd Ed.). Springer-Verlag, New York.
- 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.
- Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a Mini-Version anchor and a midi anchor: A case study using SAT® data. *Journal of Educational Measurement*, 48(4), 361-379.
- Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2009). The effects of different types of anchor tests on observed score equating. *ETS Research Report Series*, 2009(2), i-46.

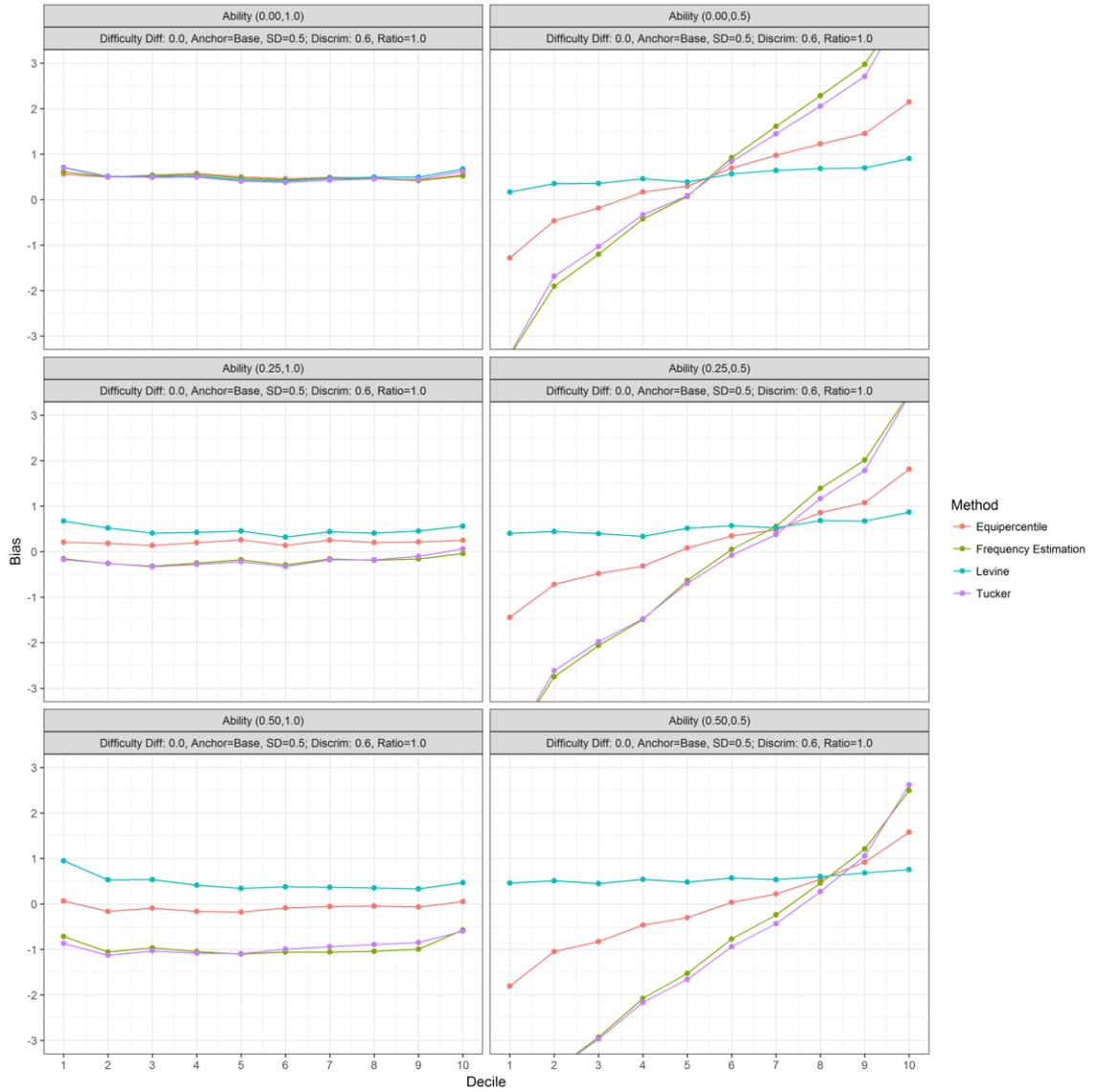
- Liu, J., Sinharay, S., Holland, P., Feigenbaum, M., & Curley, E. (2011). Observed Score Equating Using a Mini-Version Anchor and an Anchor with Less Spread of Difficulty: A Comparison Study. *Educational and Psychological Measurement*, 71, 2, 346.
- Livingston, S. A. (2014). Equating test scores (without IRT). Educational testing service.
- Lord, F. M. (1952). The relationship of the reliability of multiple-choice test to the distribution of item difficulties. *Psychometrika*, 17, 181-194.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates, Inc.
- Luecht, R. M. (2014). GENEQUATE: A 3PL Item Model-Based Data Generation. Observed-Score Version for Equating Research. Version 4.0.
- Luecht, R., & Ackerman T. A. (2018). Citation for: A Technical Note on IRT Simulation Studies: Dealing With Truth, Estimates, Observed Data, and Residuals. *Educational Measurement: Issues and Practice*.
- Magis D, & Raiche G (2011). catR: An R Package for Computerized Adaptive Testing. *Applied Psychological Measurement*, 35, 576–577.
- Moses, T., & Kim, S. (2007). Reliability and the Nonequivalent Groups With Anchor Test Design. *ETS Research Report Series*, 2007(1).
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. *Educational measurement*, 3, 221-262.
- Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. *Test equating*, 71-135.

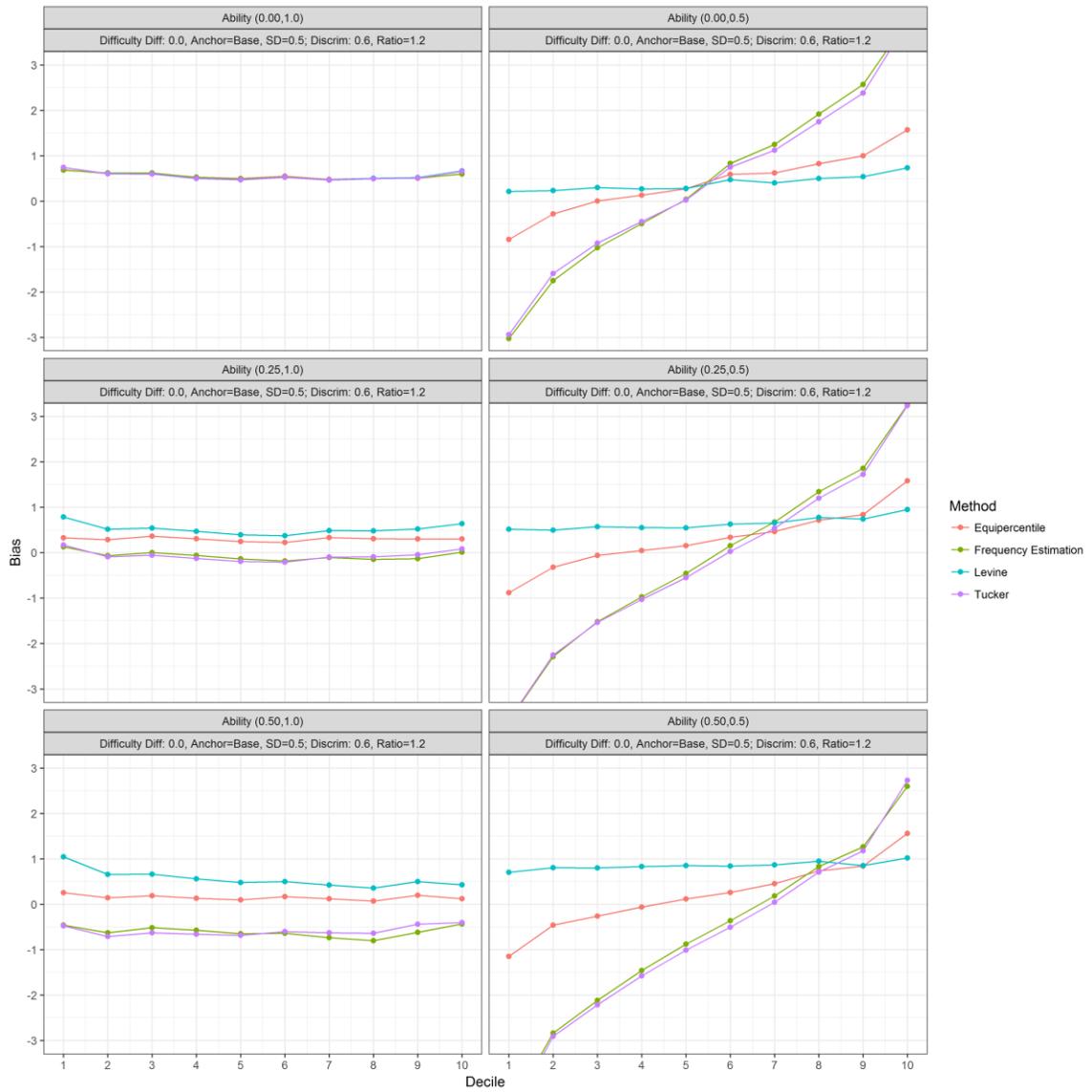
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1, 2, 33-49.
- Sinharay, S., Haberman, S., Holland, P., & Lewis, C. (2012). A note on the choice of an anchor test in equating. *ETS Research Report Series*, 2012(2).
- Sinharay, S., & Holland, P. (2006a). The correlation between the scores of a test and an anchor test. *ETS Research Report Series*, 2006(1), i-28.
- Sinharay, S., & Holland, P. (2006b). Choice of anchor test in equating. *ETS Research Report Series*, 2006(2), i-43.
- Sinharay, S., & Holland, P. W. (2007). Is It Necessary to Make Anchor Tests Mini-Versions of the Tests Being Equated or Can Some Restrictions Be Relaxed?. *Journal of Educational Measurement*, 44(3), 249-275.
- Trierweiler, T. J., Lewis, C., & Smith, R. L. (2016). Further Study of the Choice of Anchor Tests in Equating. *Journal of Educational Measurement*, 53(4), 498-518.

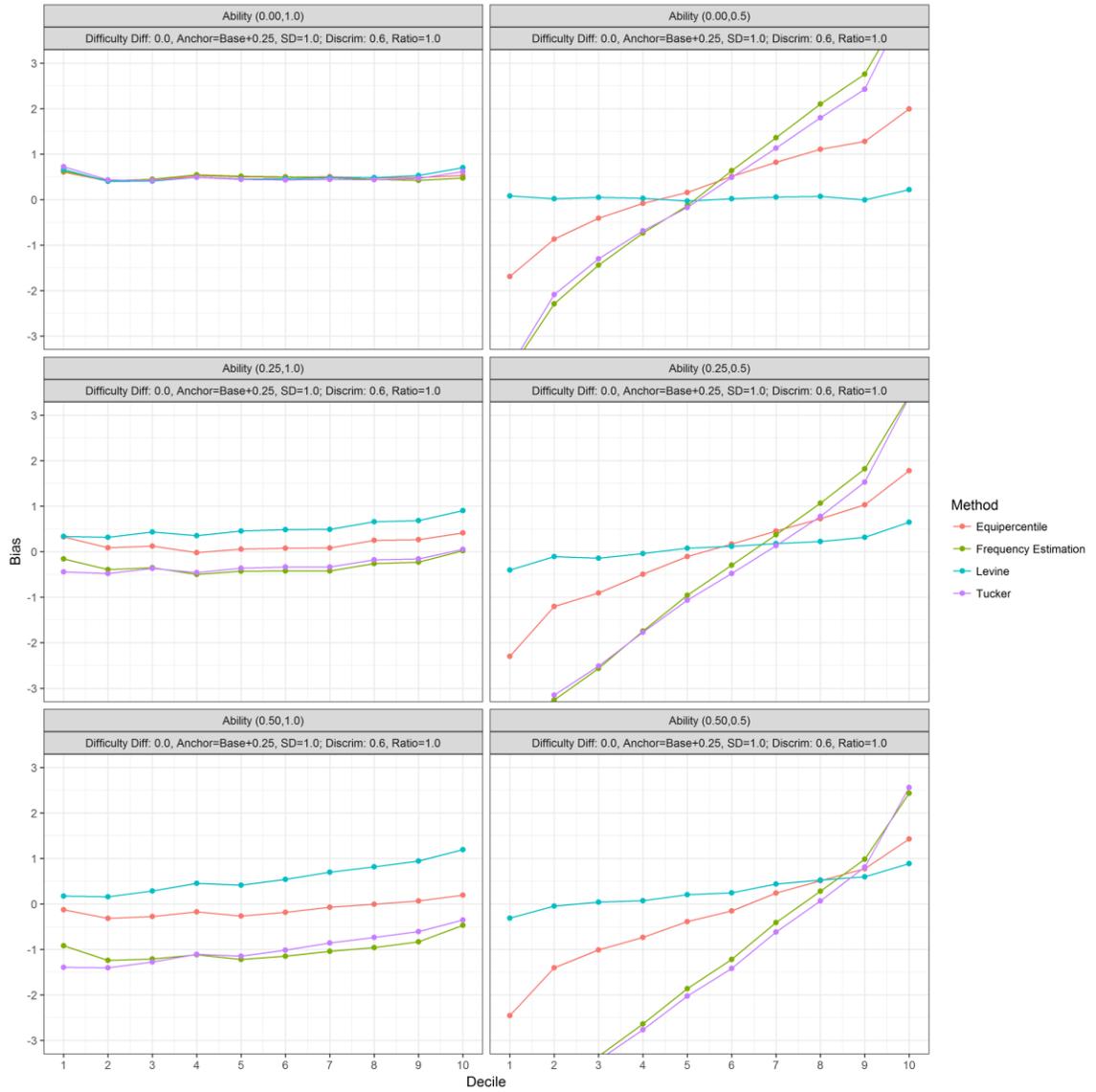
APPENDIX A  
ACHIEVEMENT BIAS RESULTS

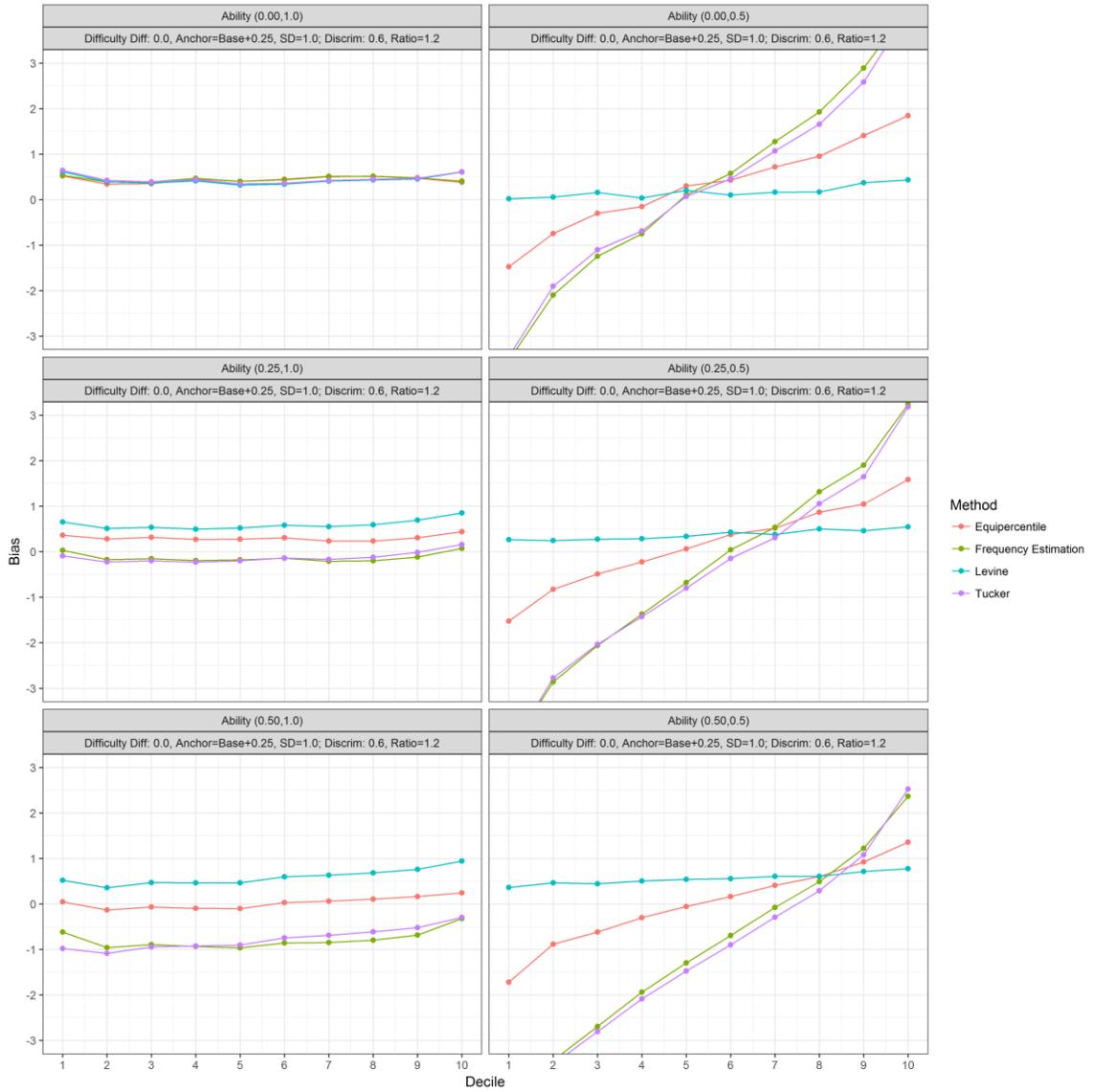


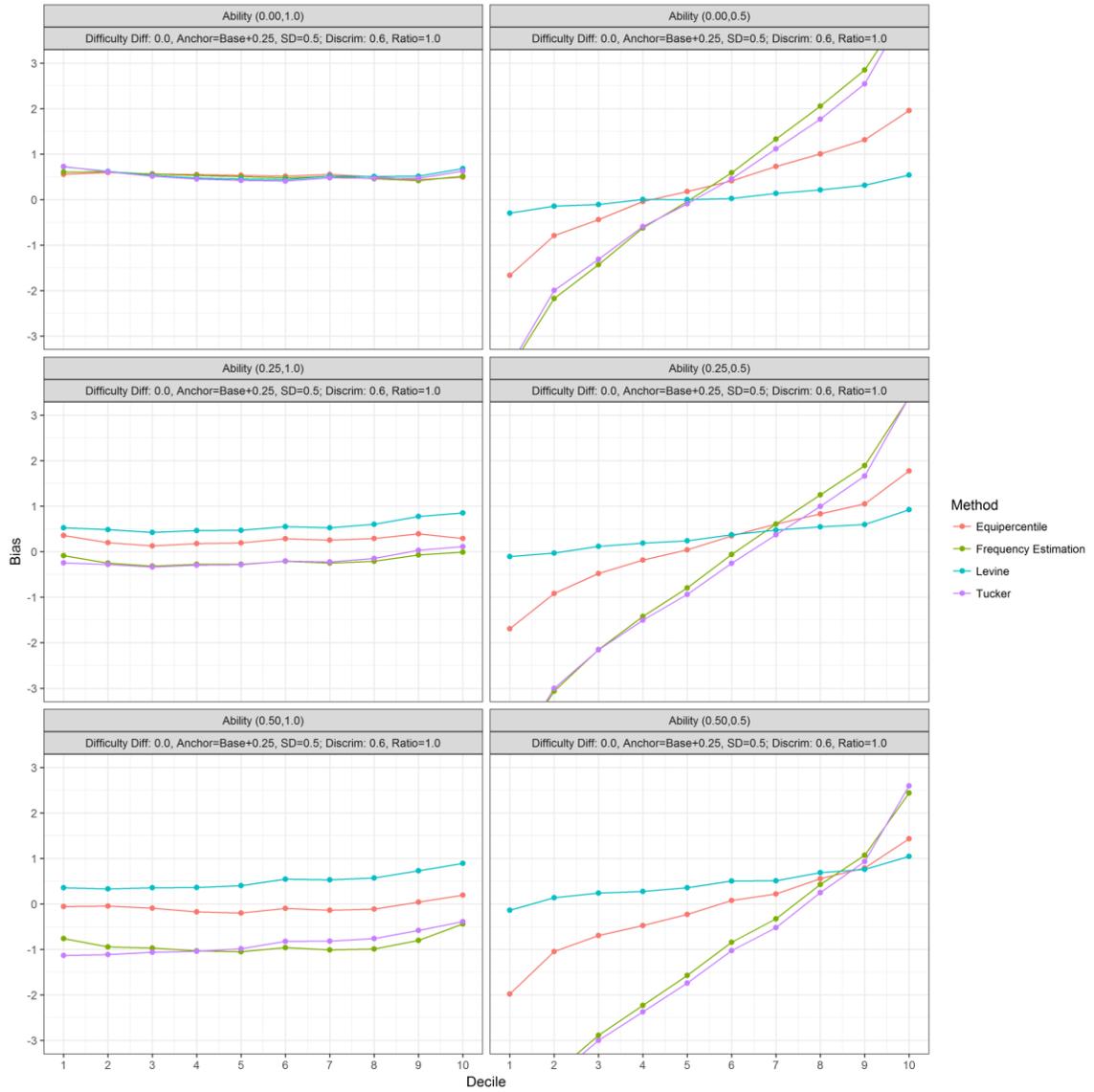


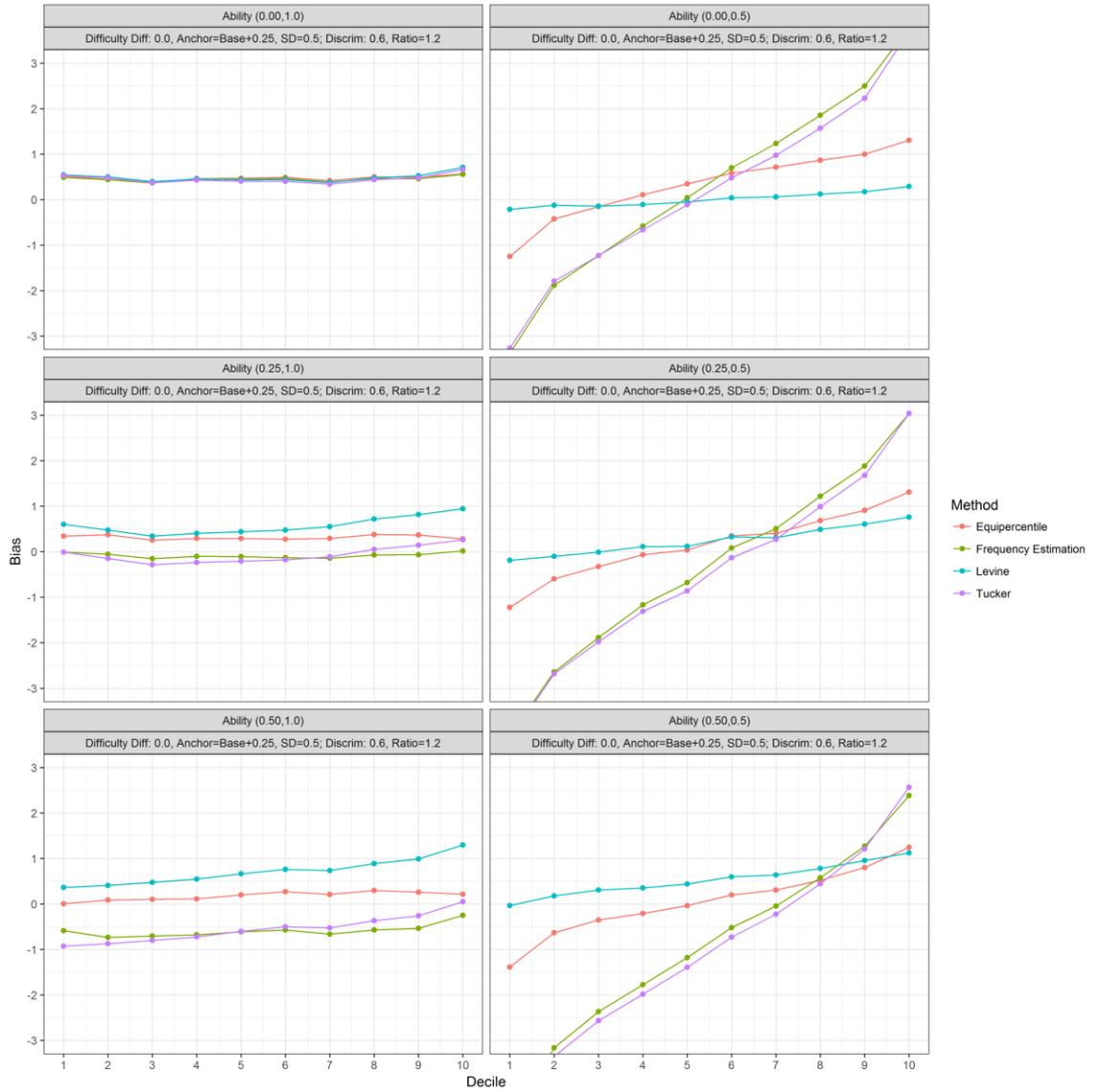


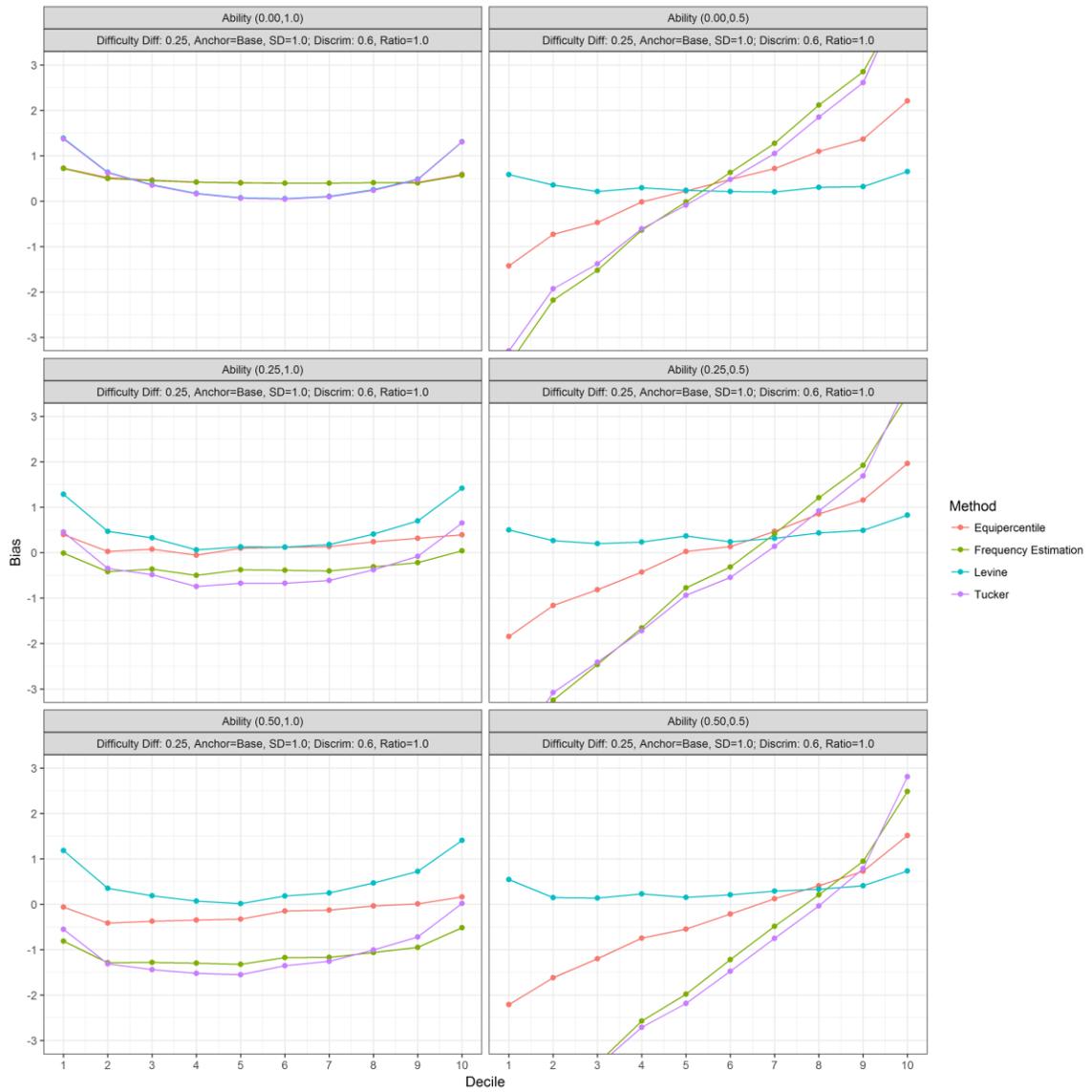


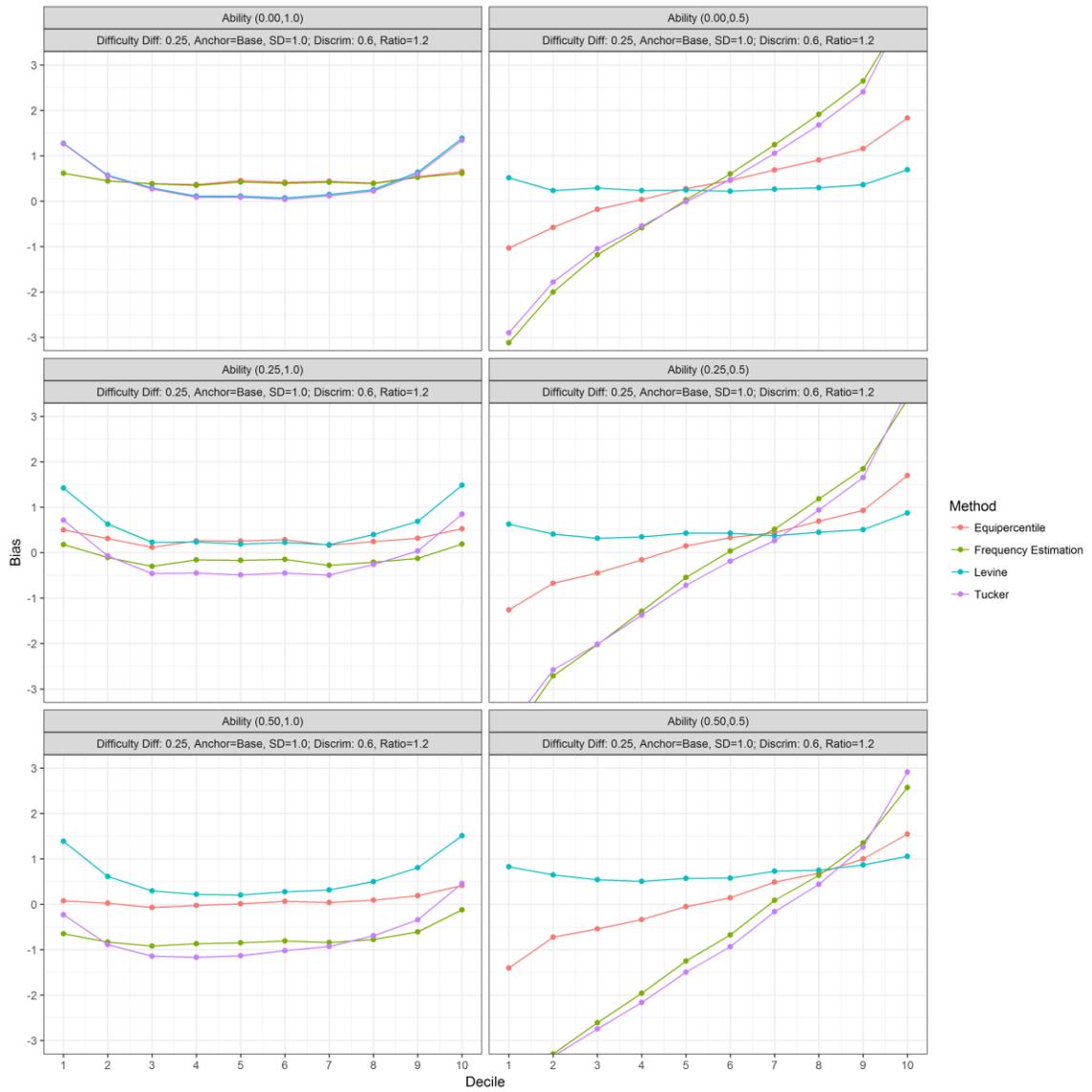


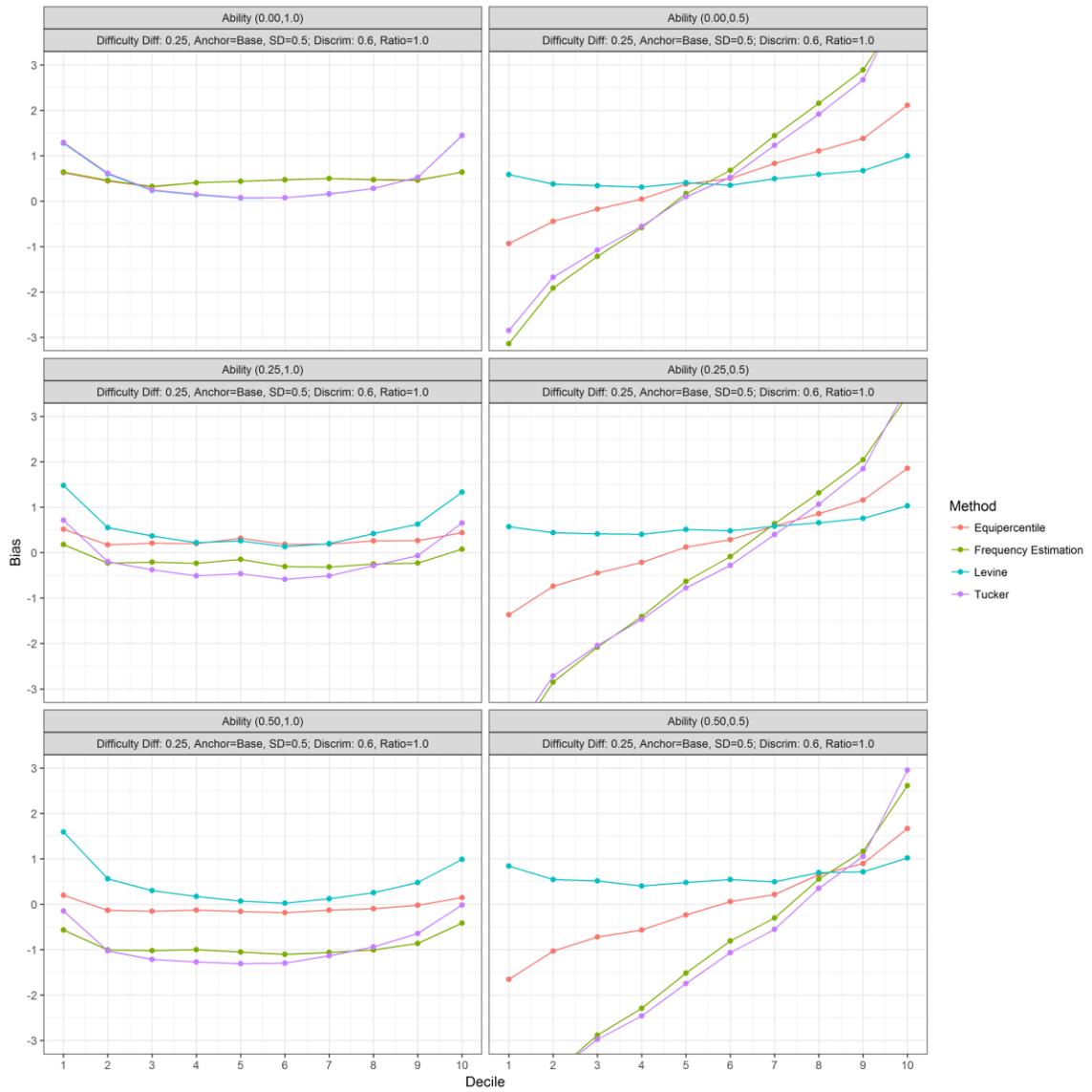


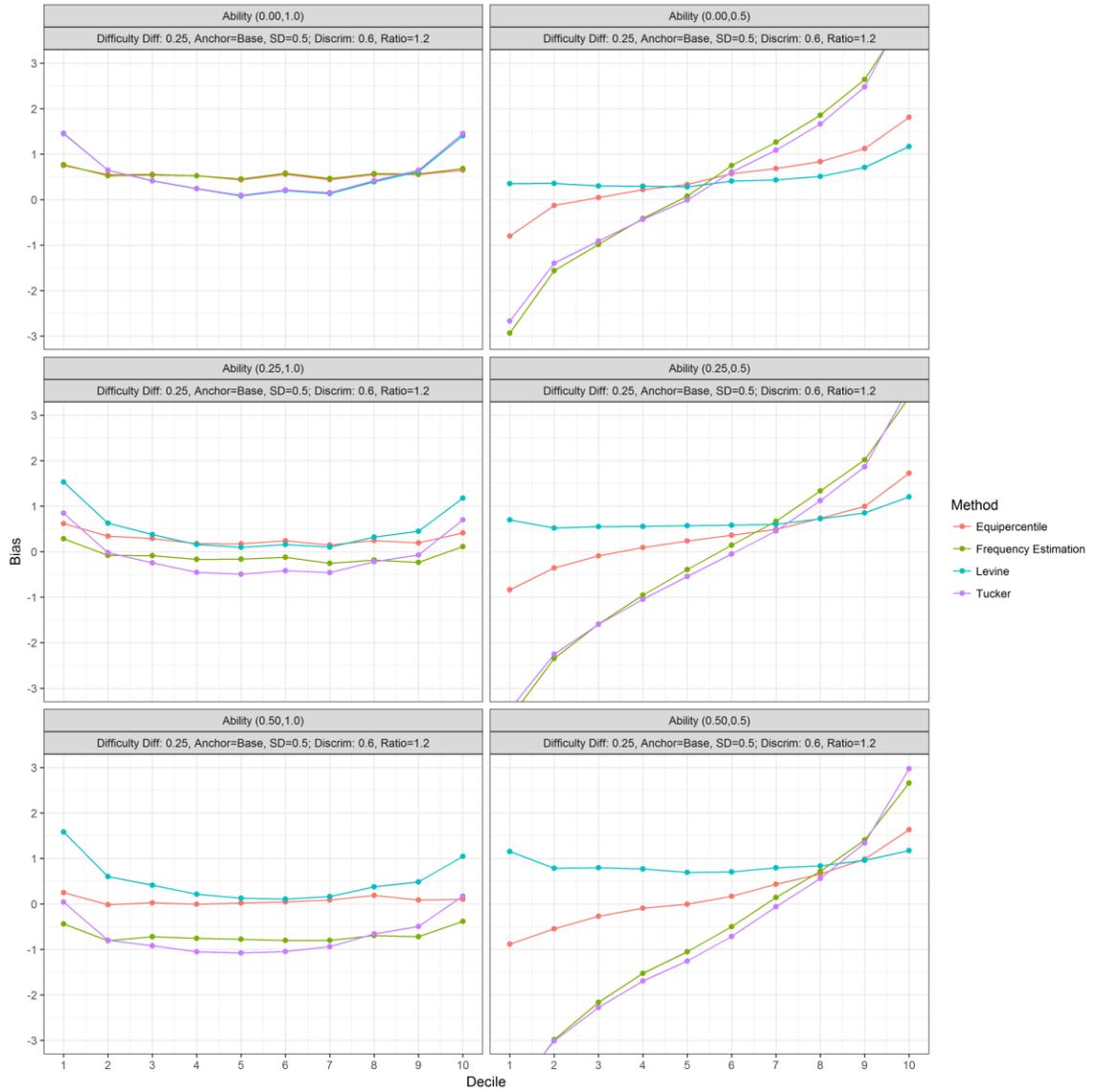


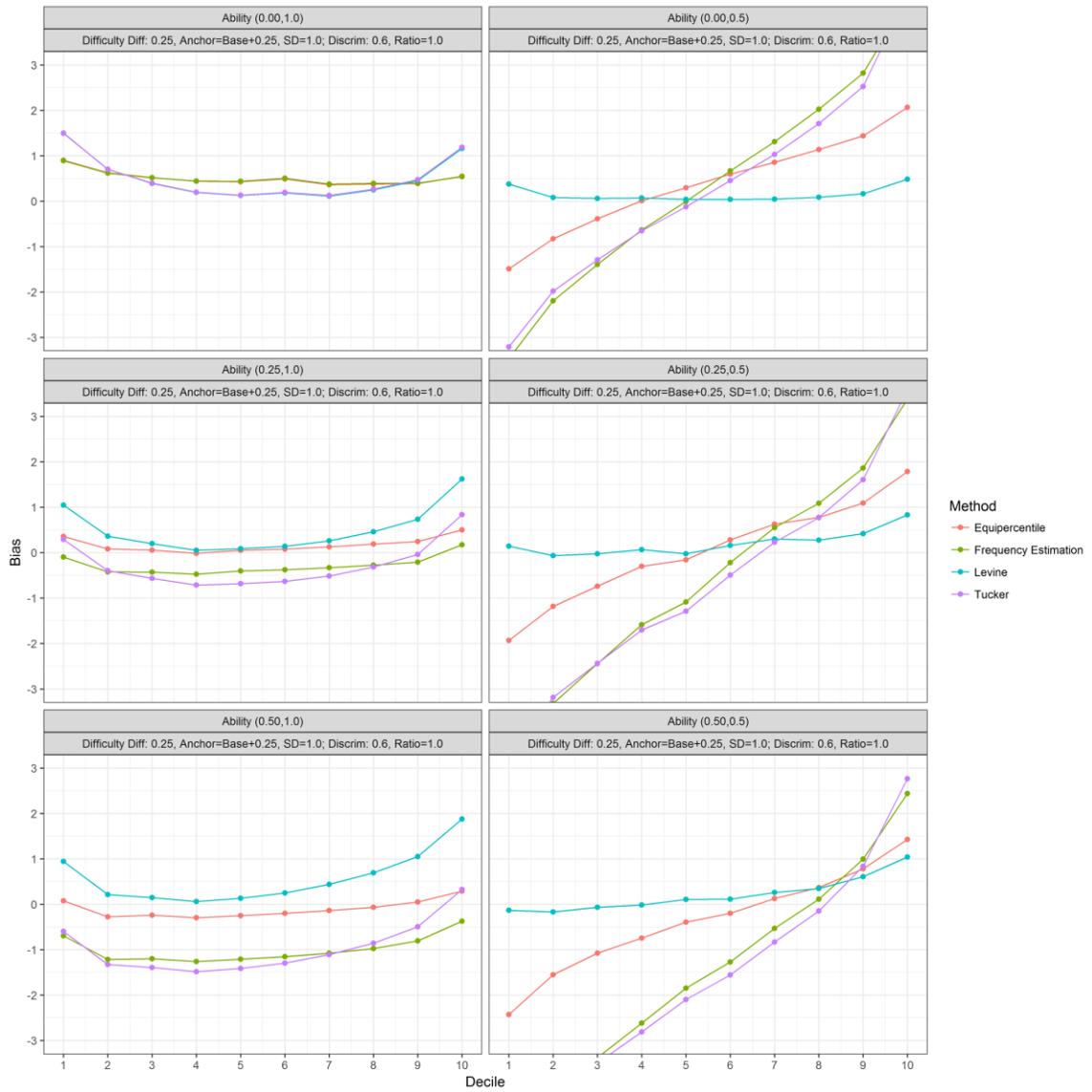


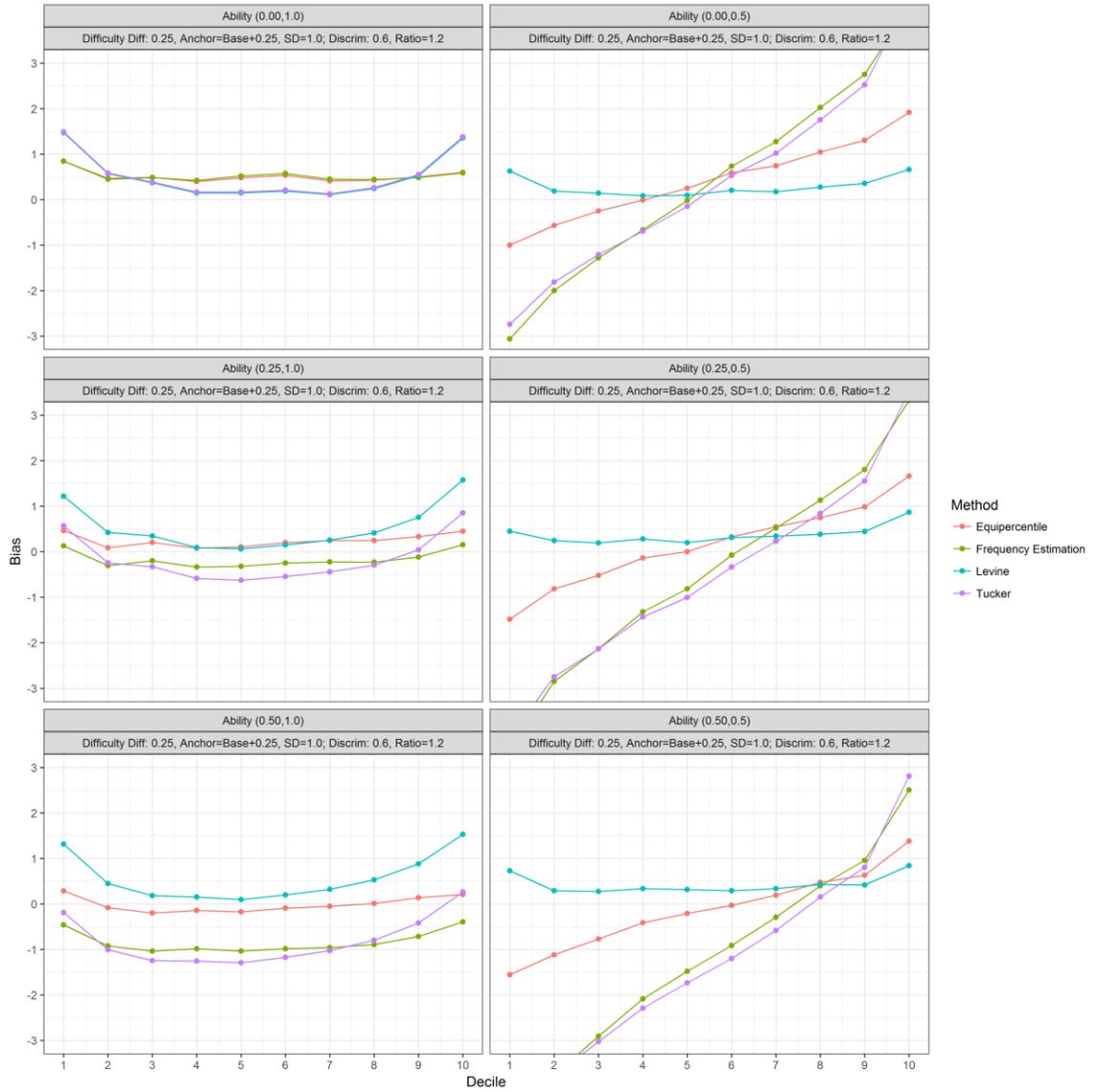


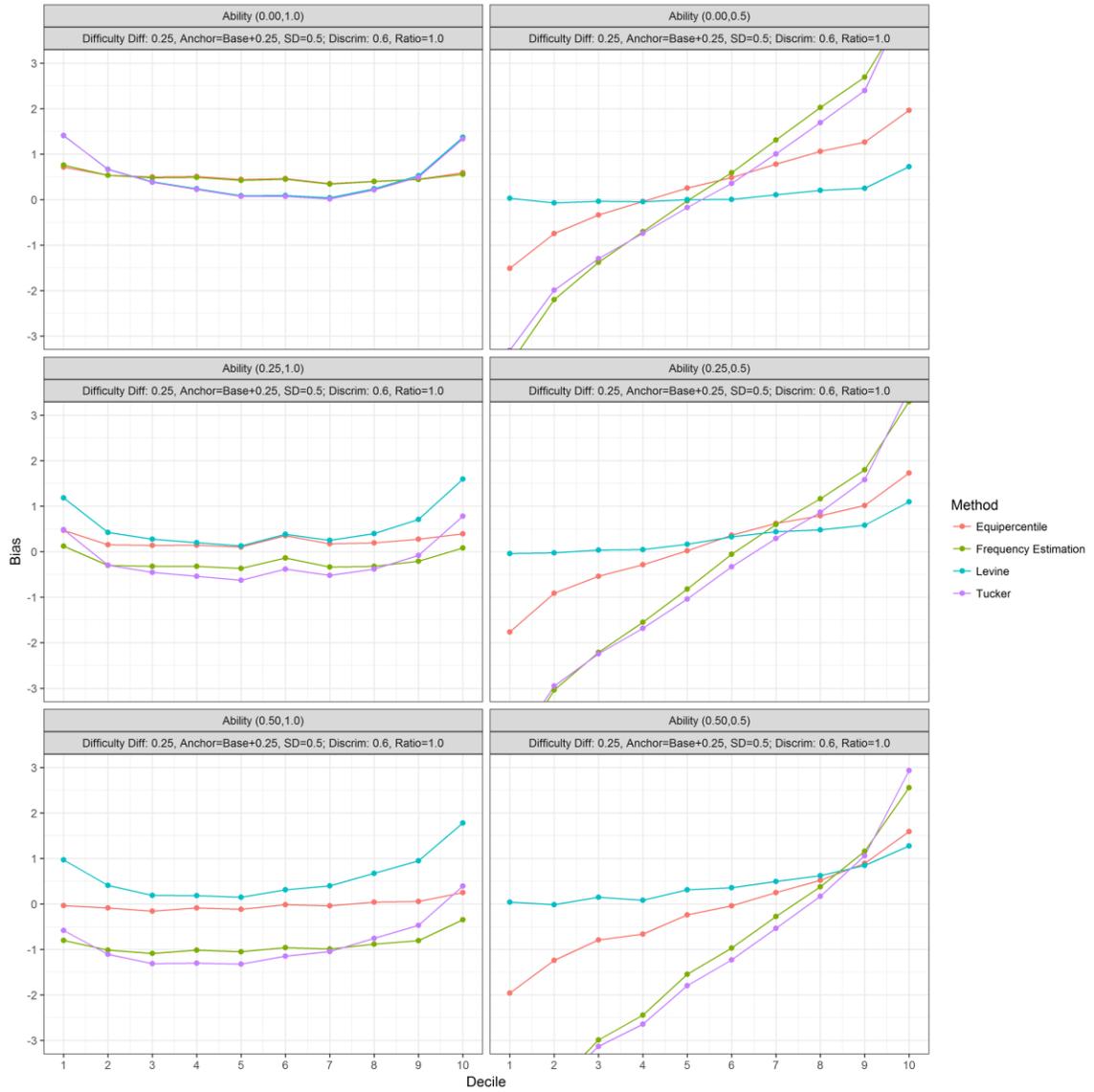


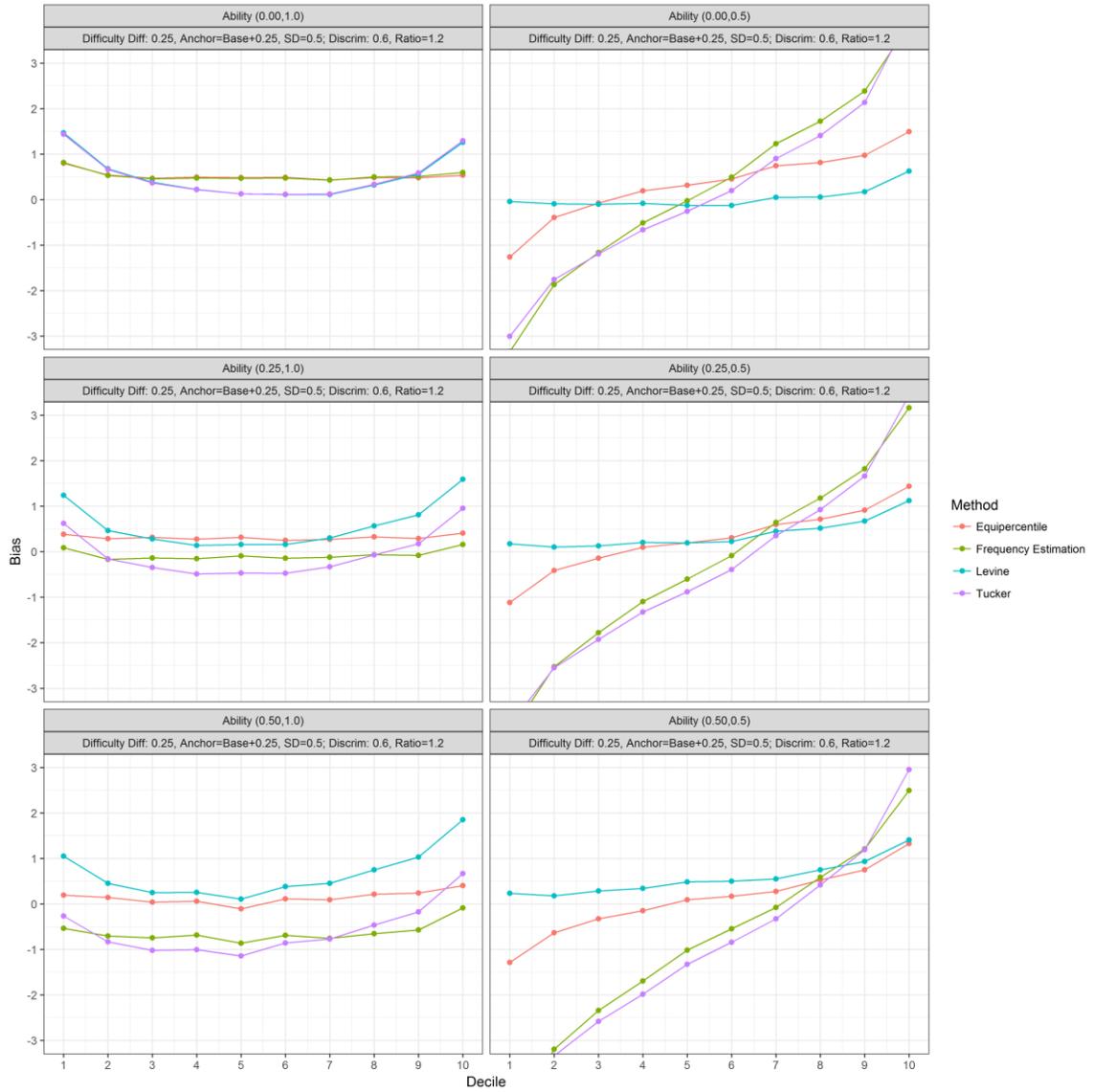


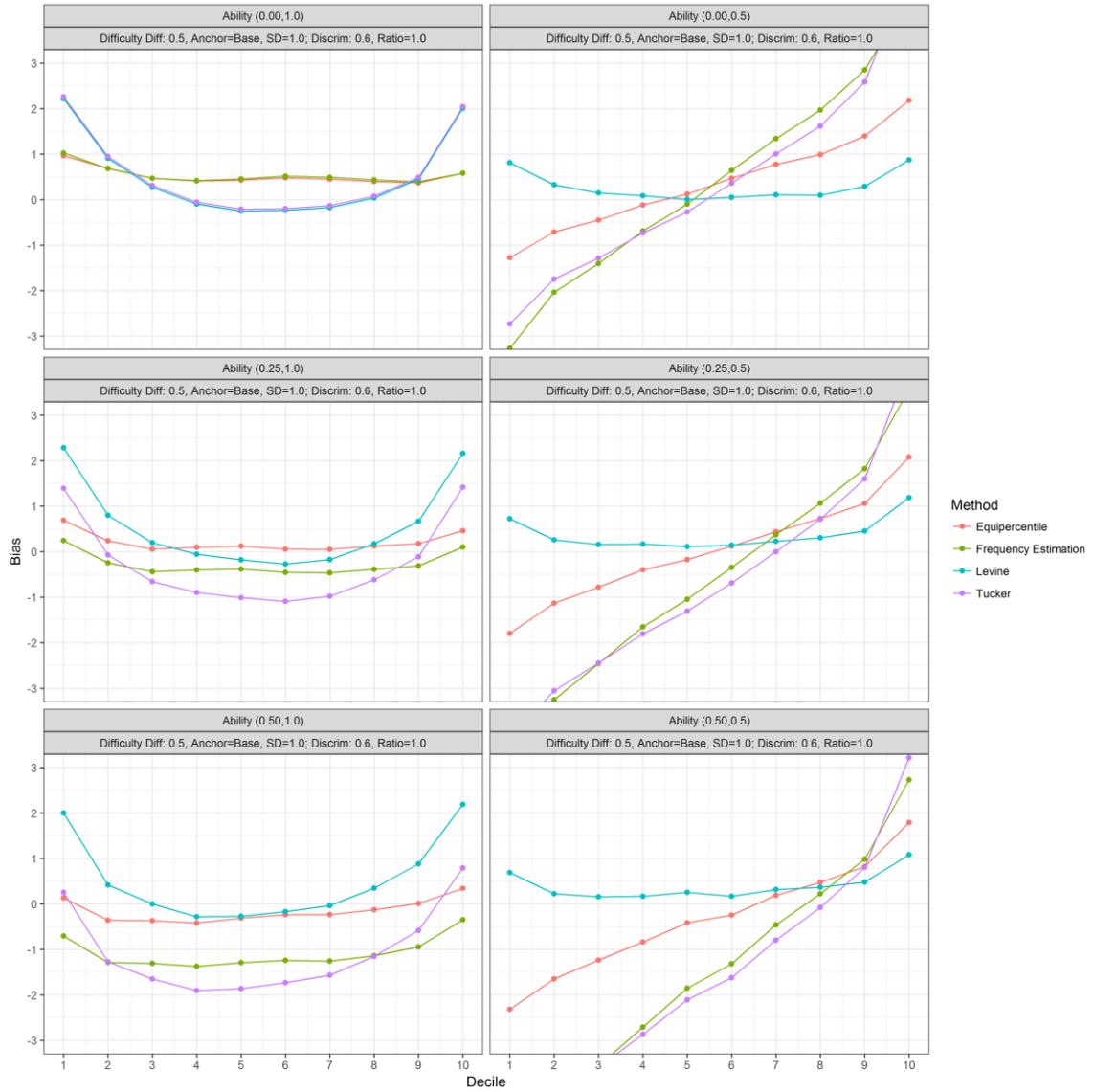


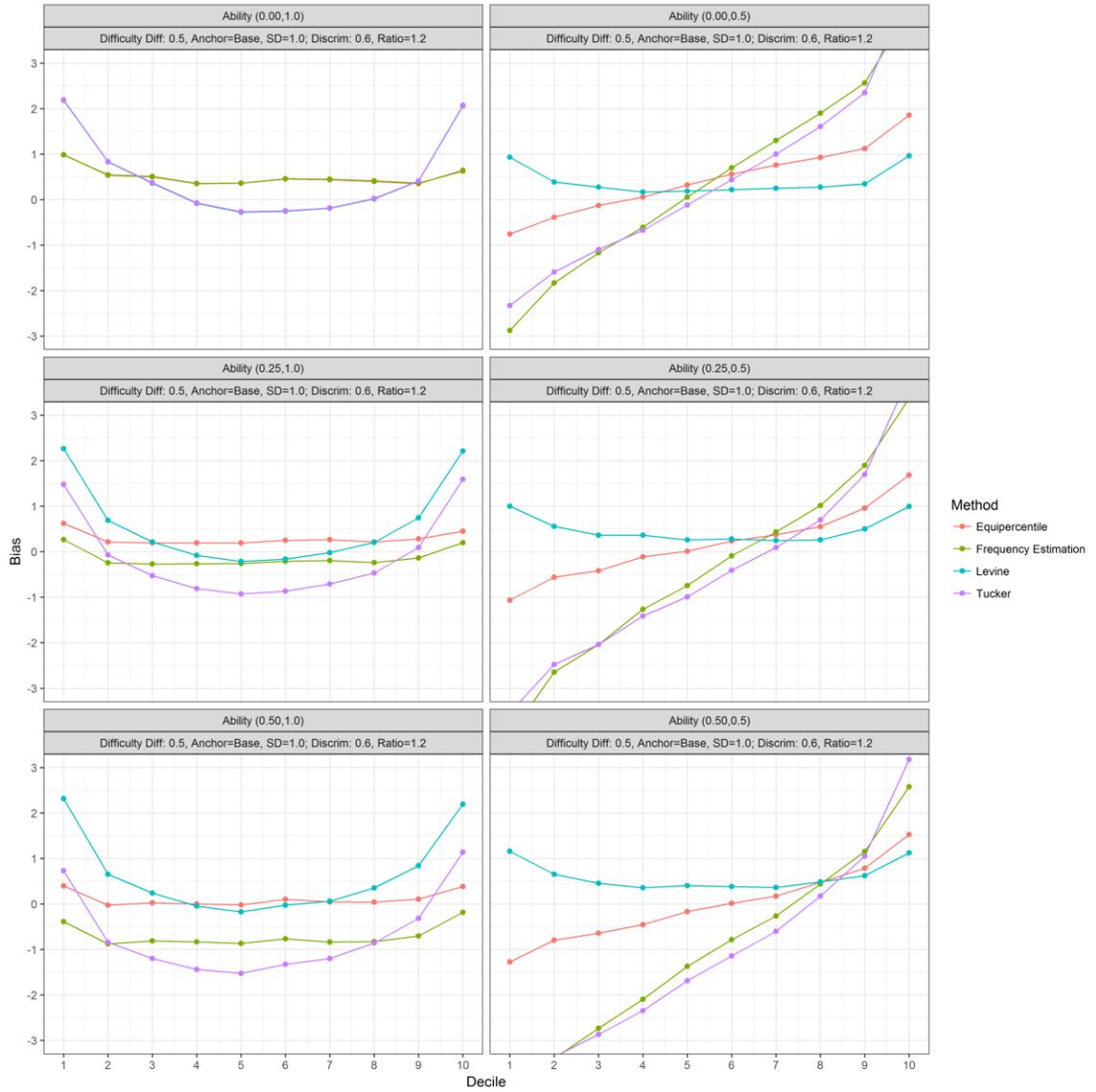


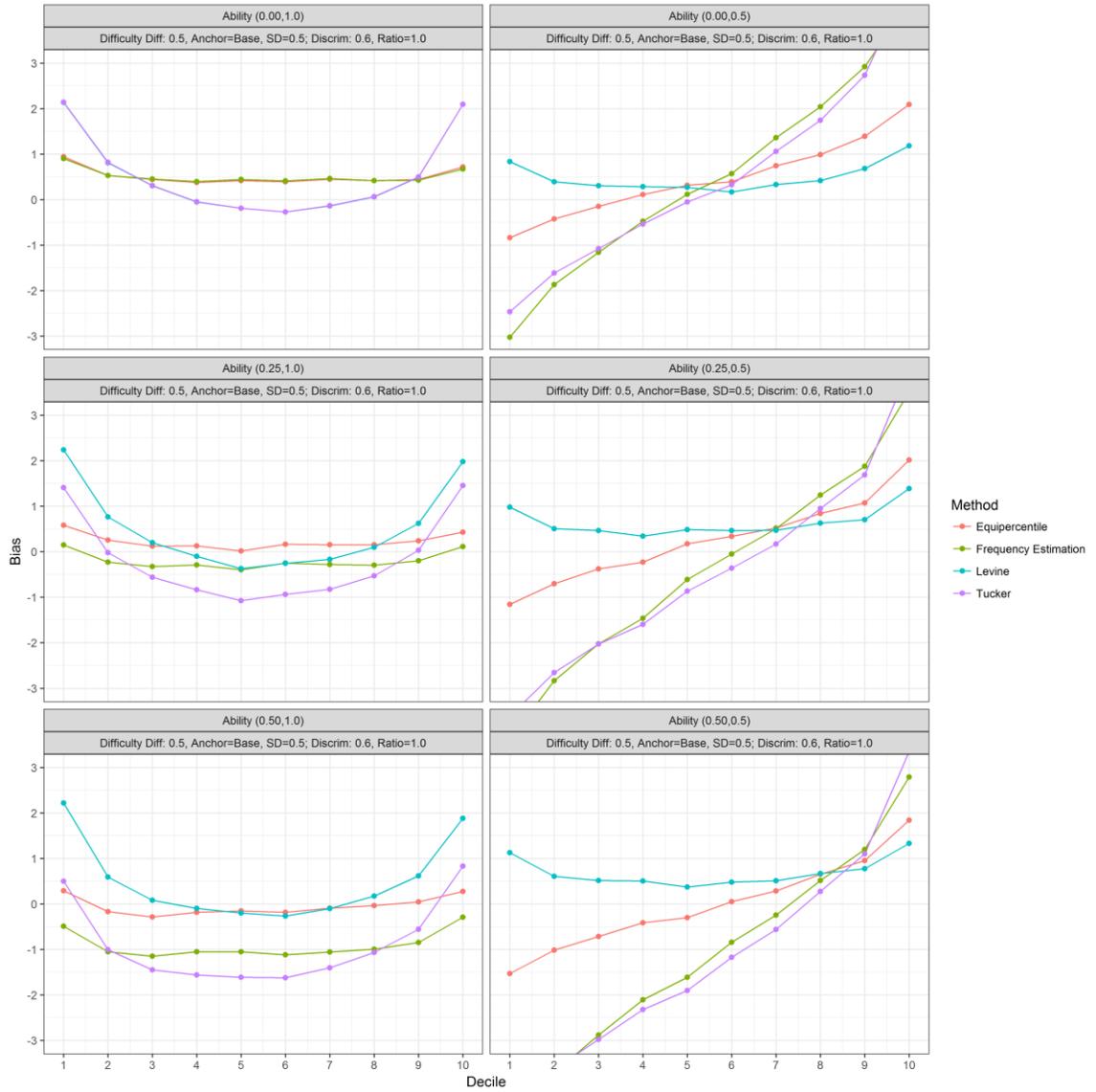


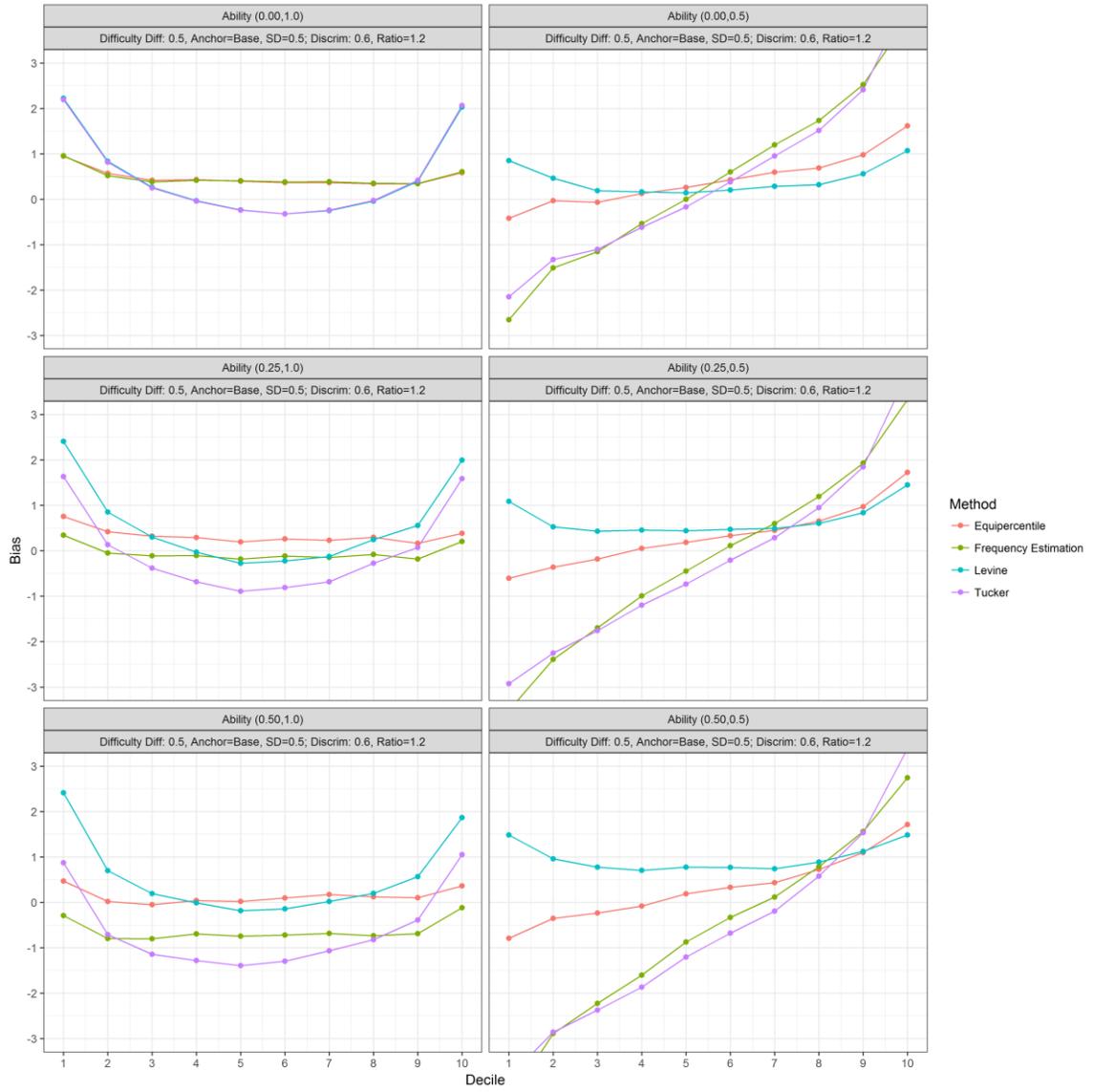


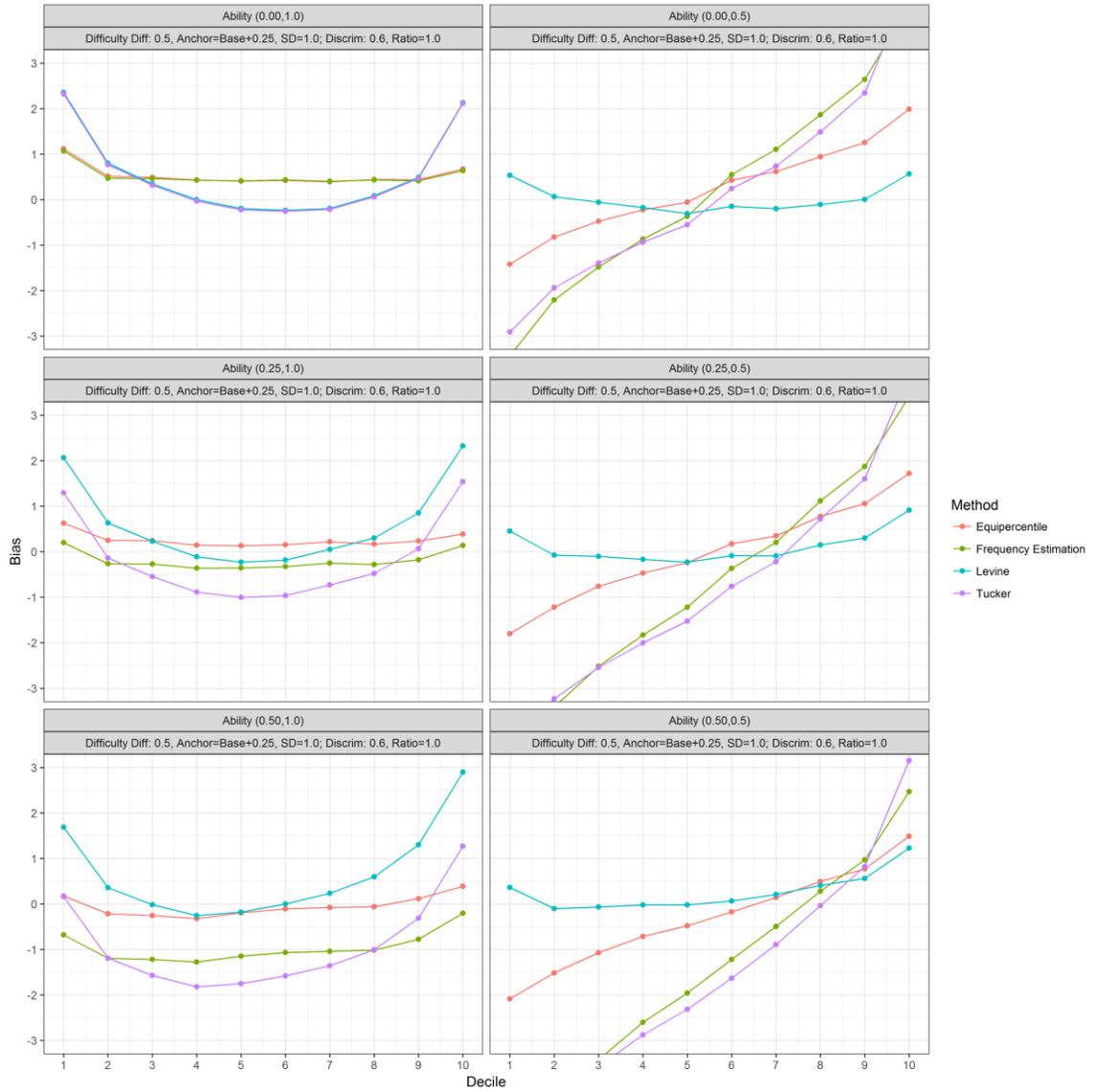


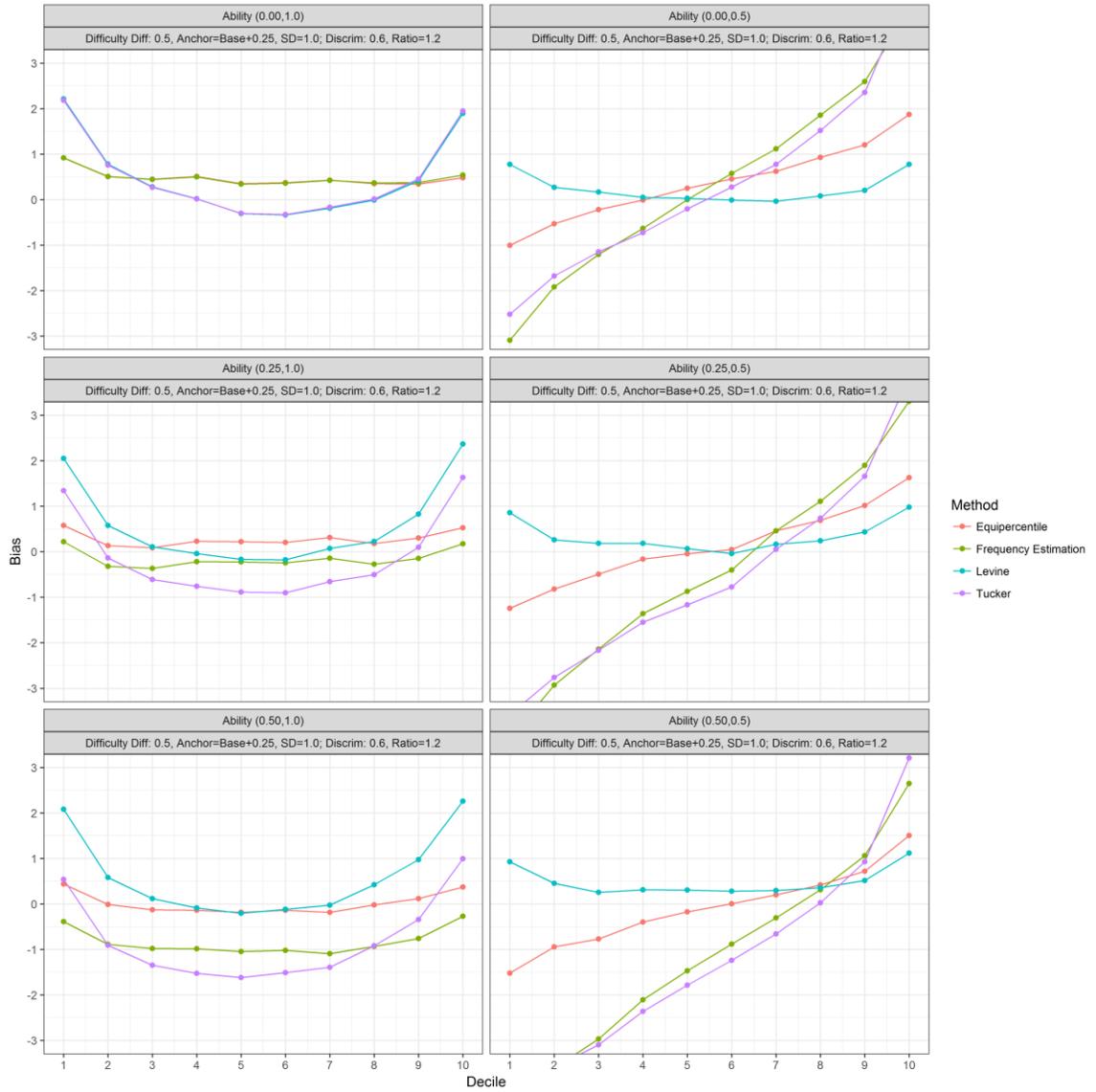


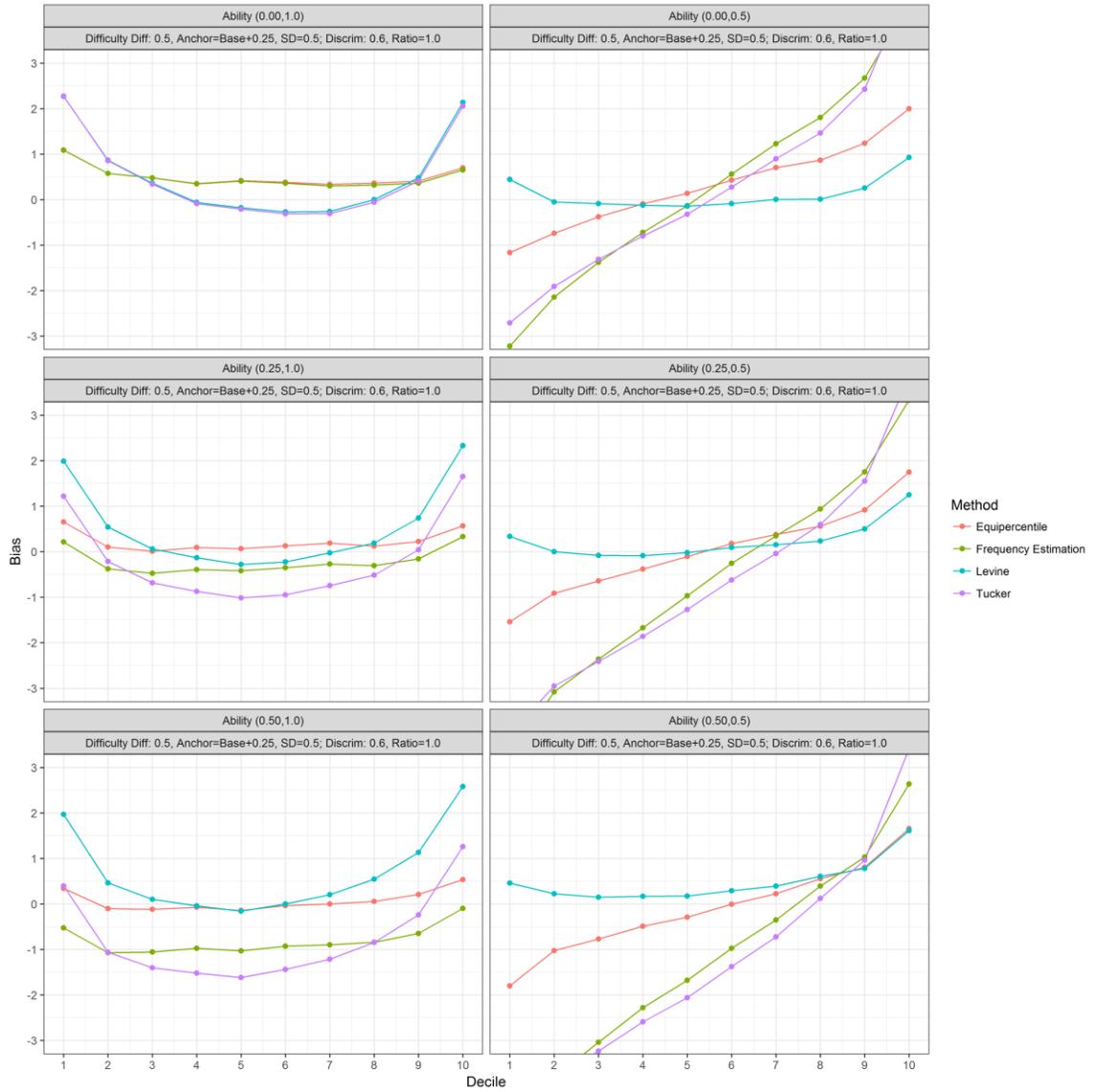


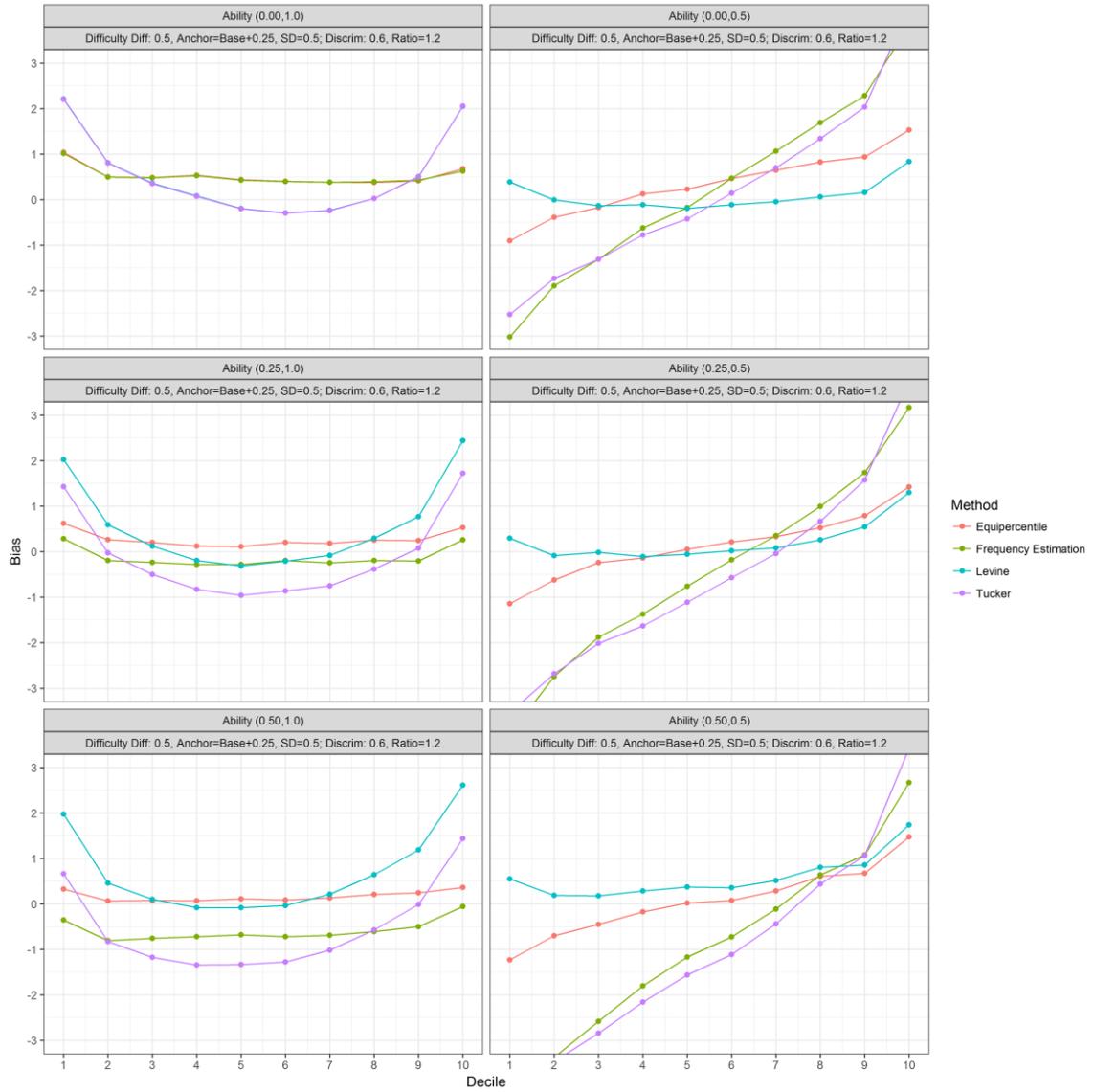


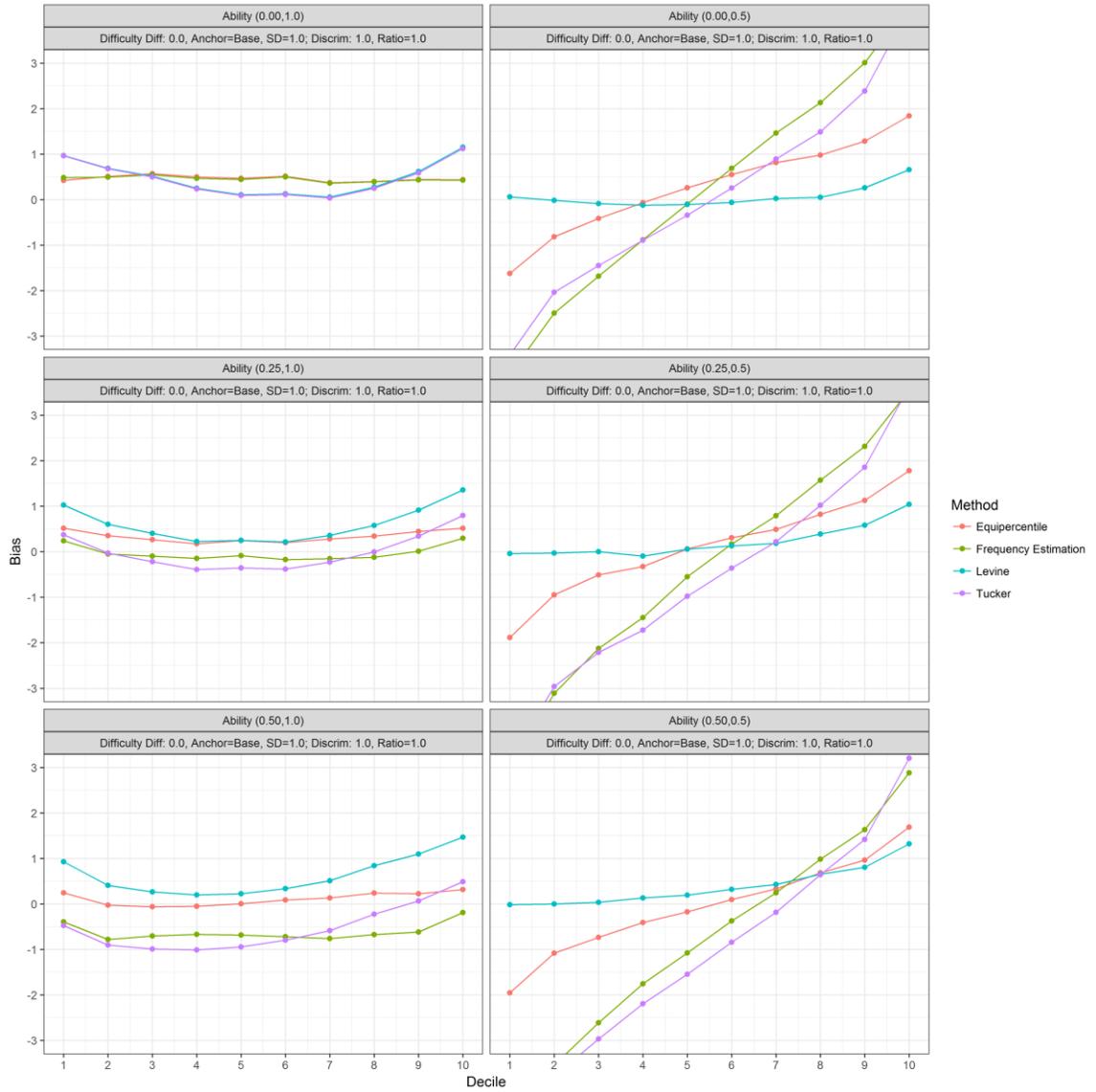


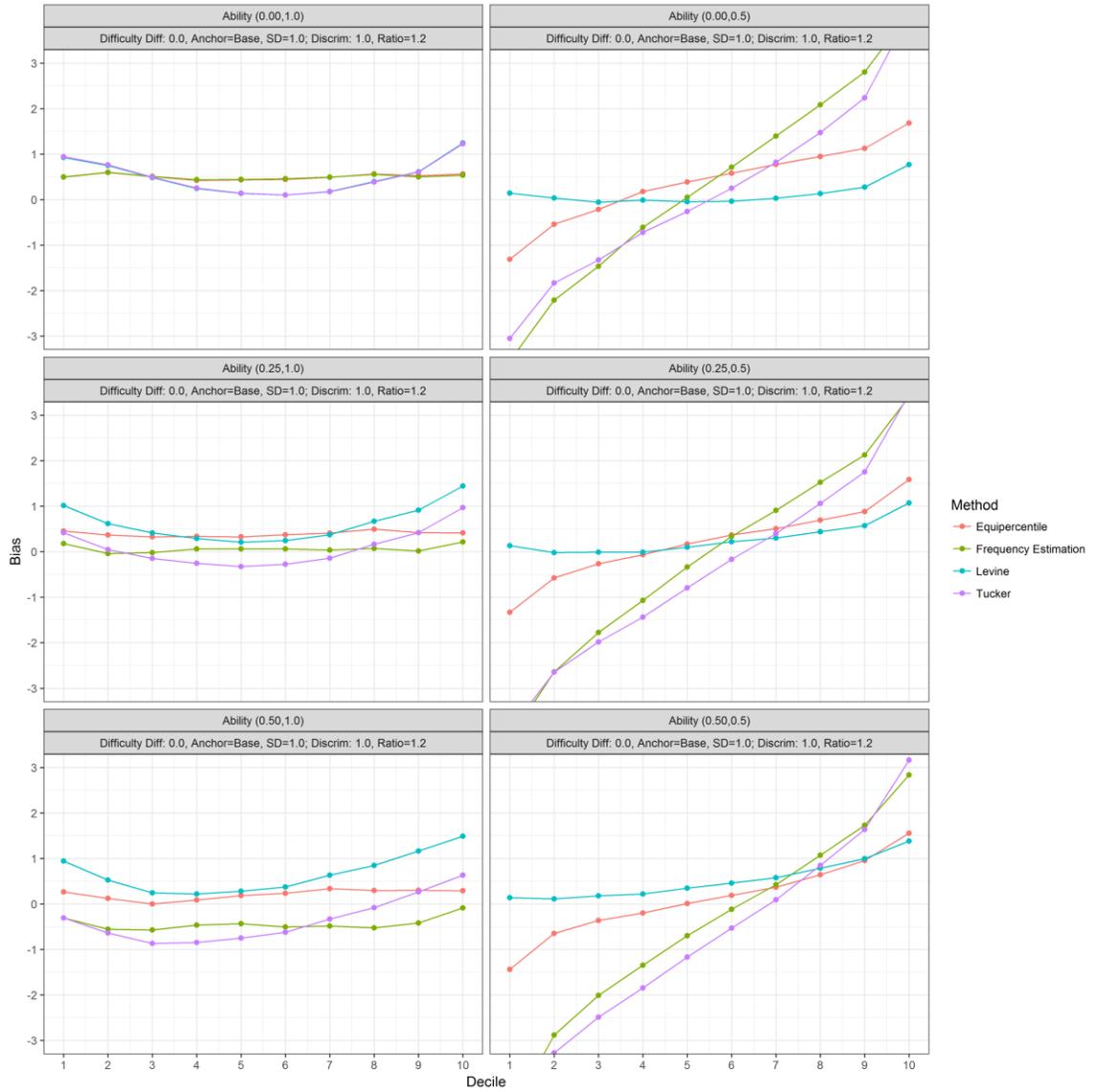


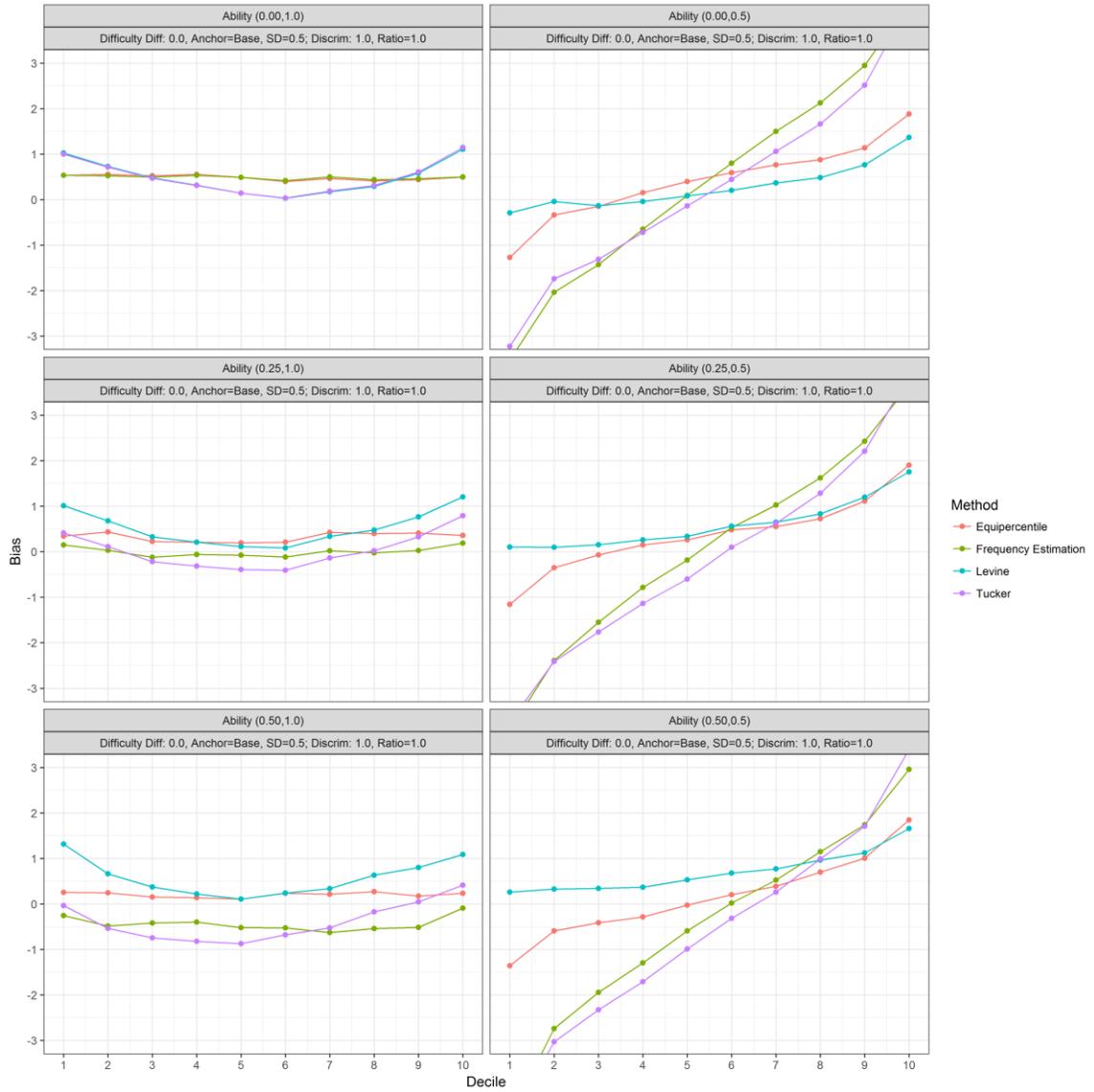


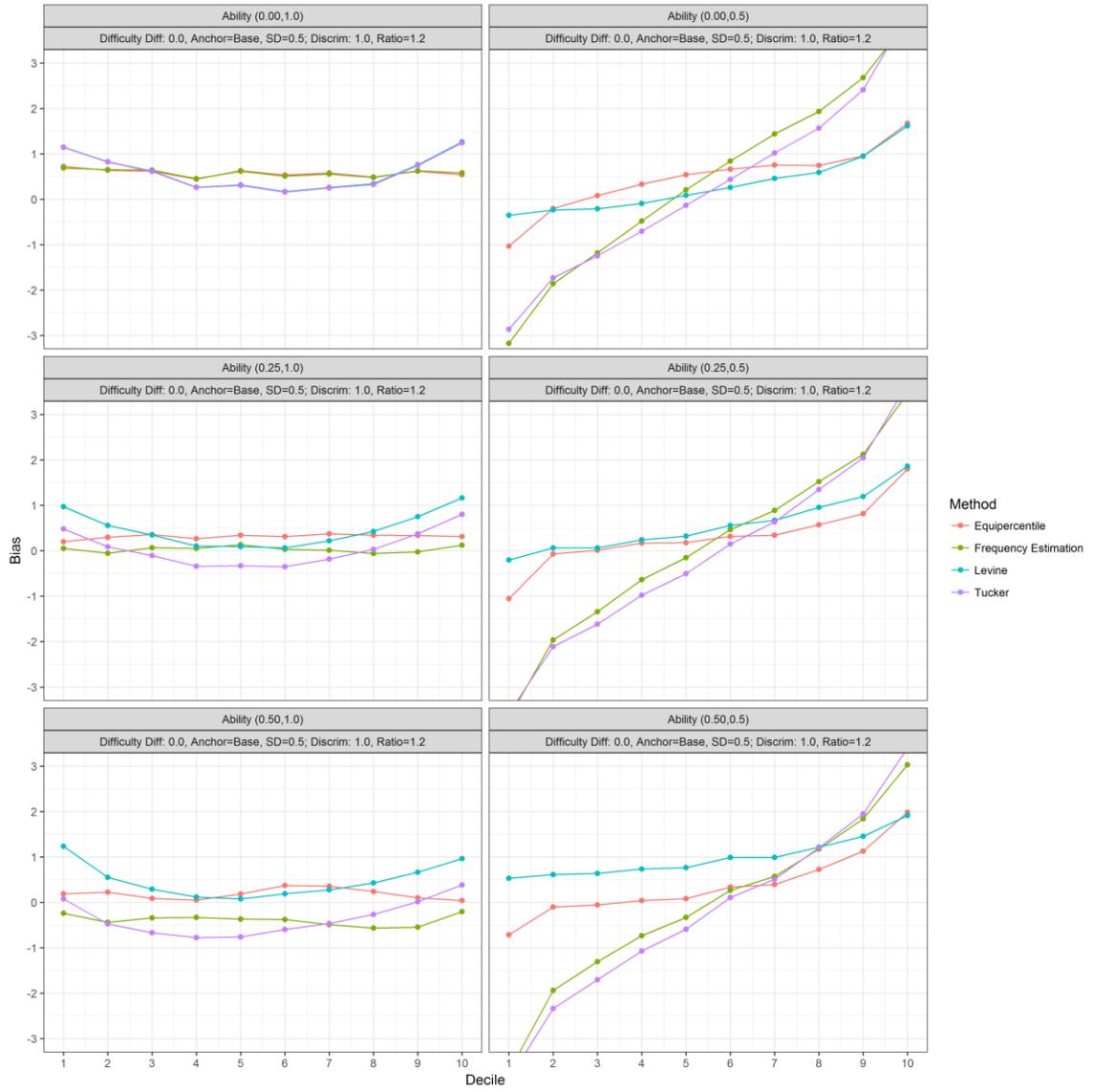


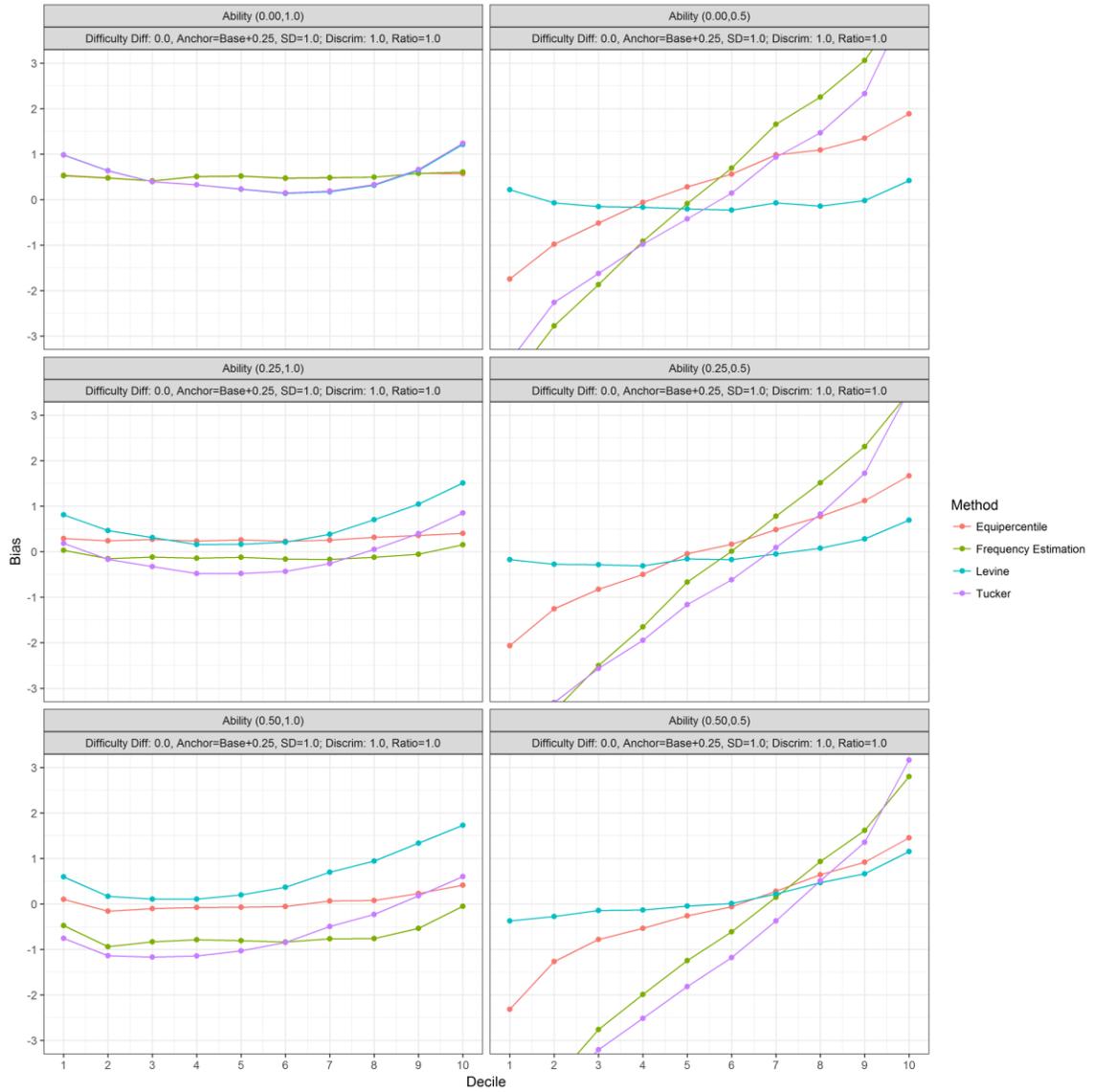


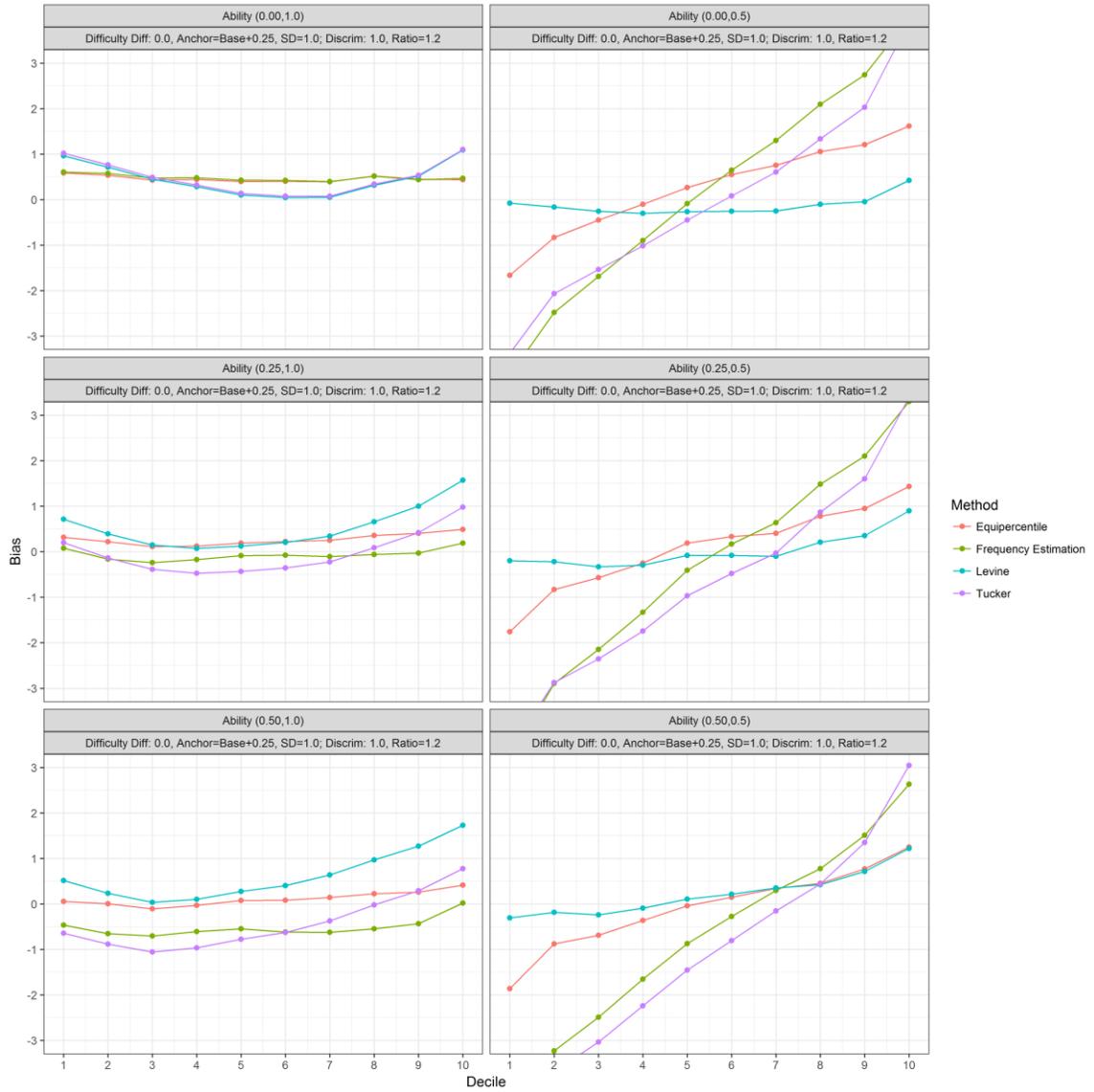


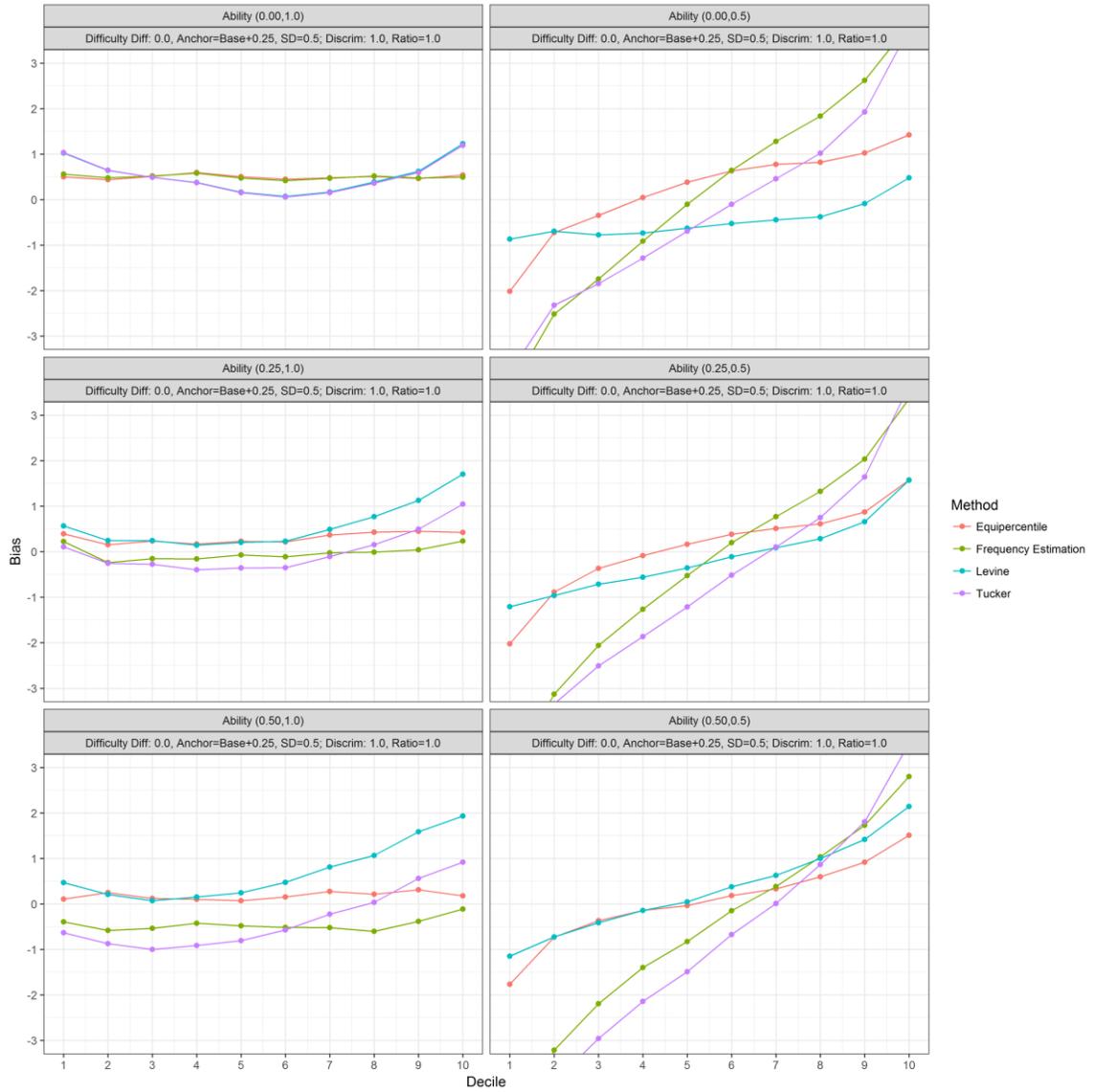


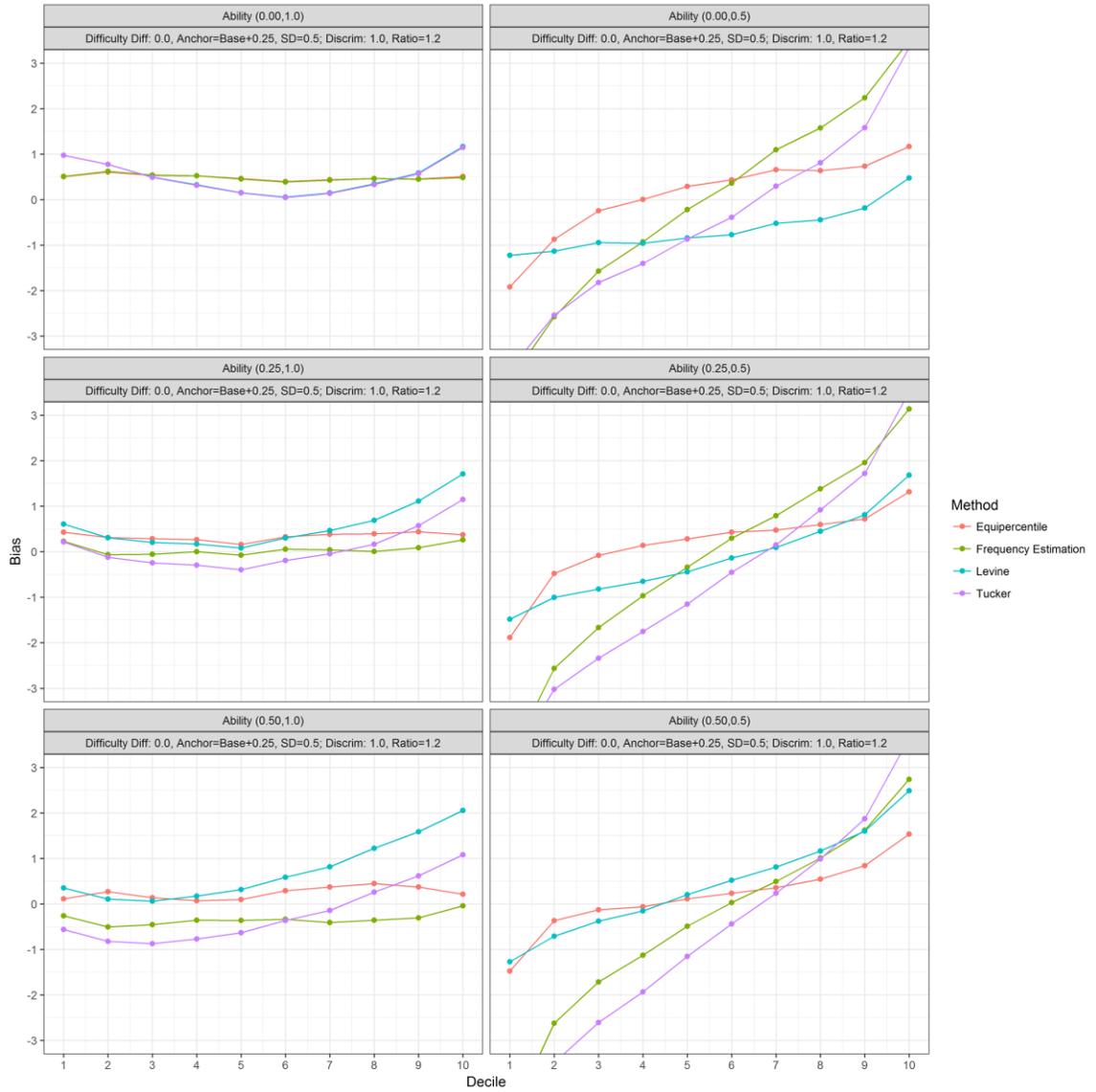


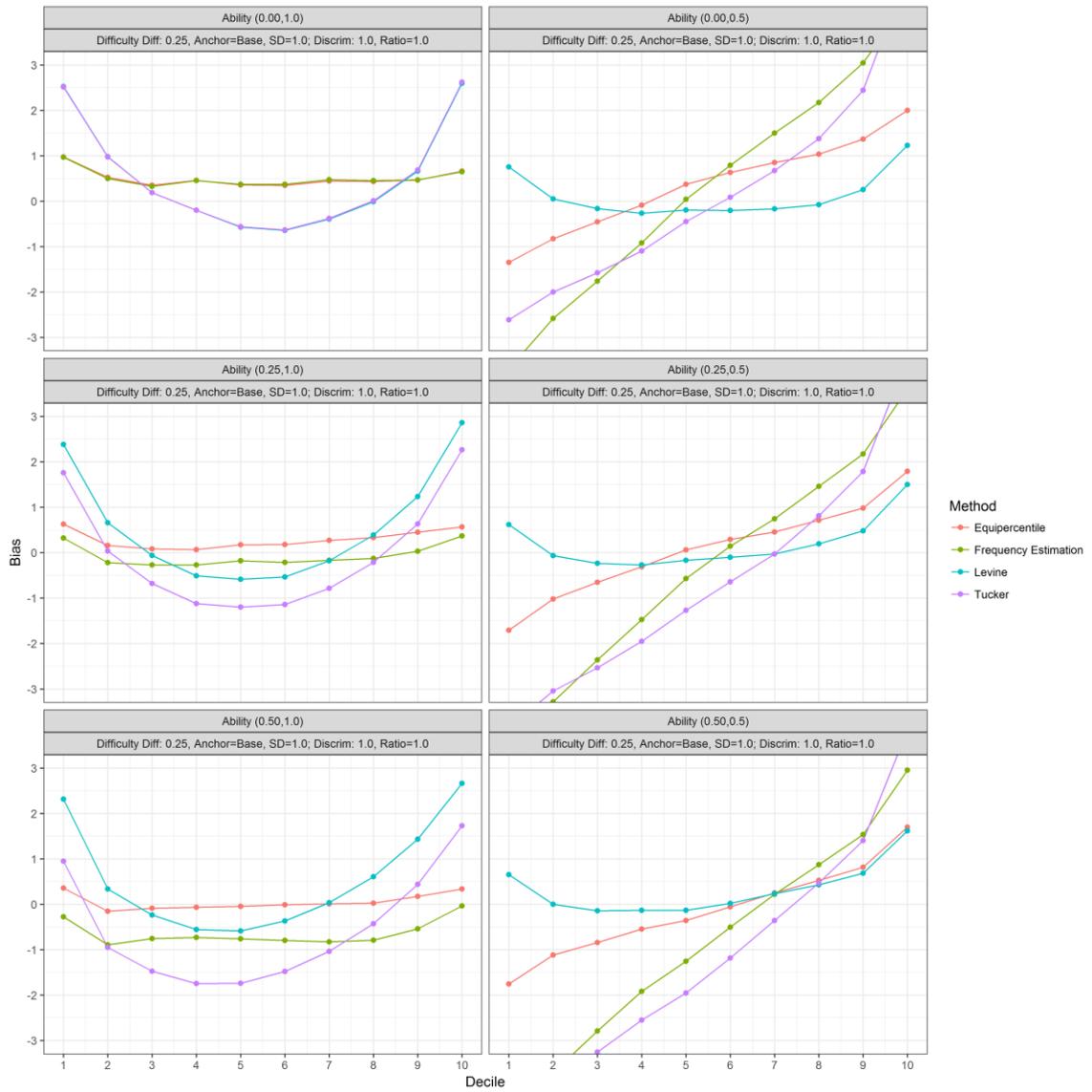


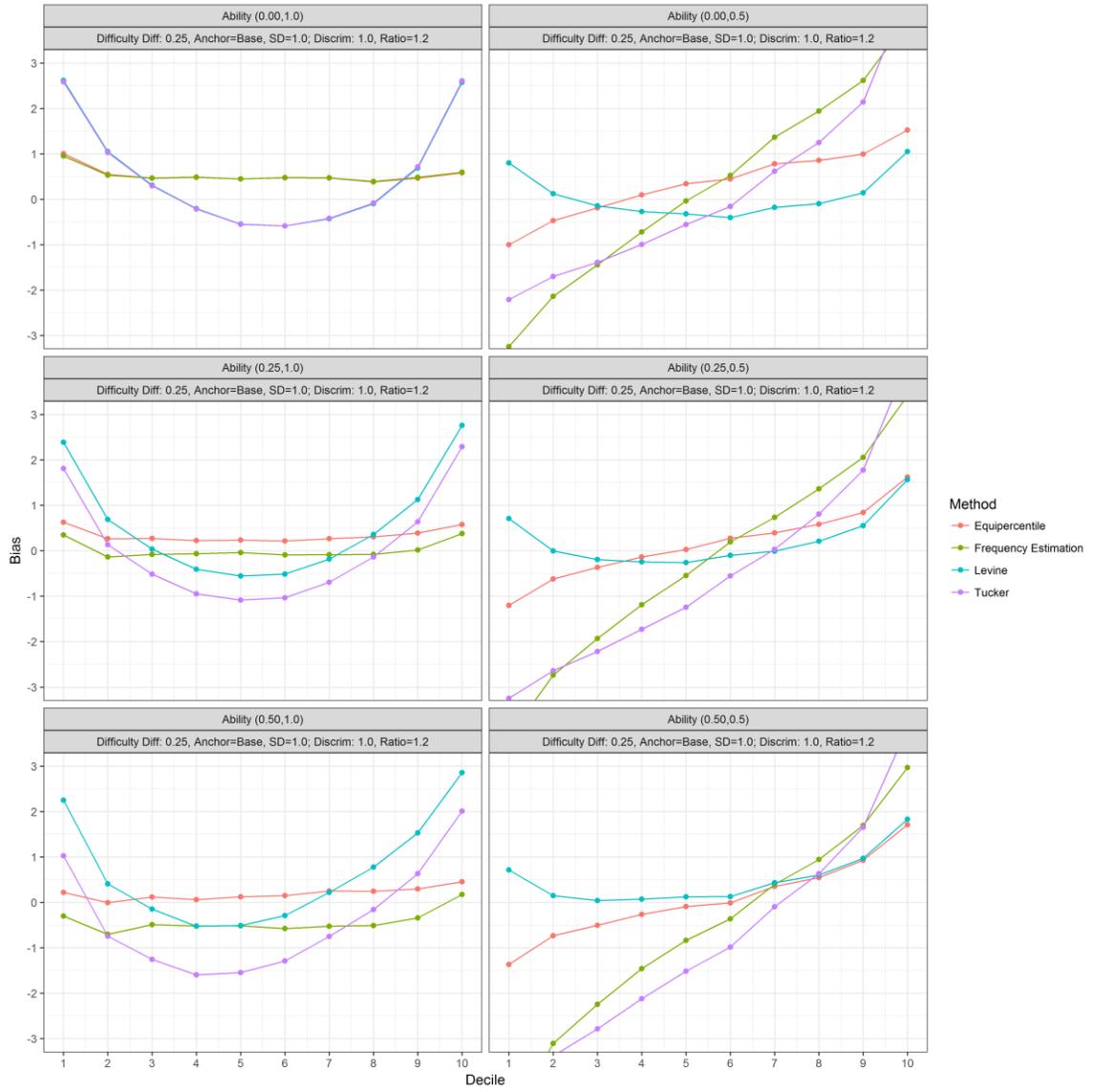


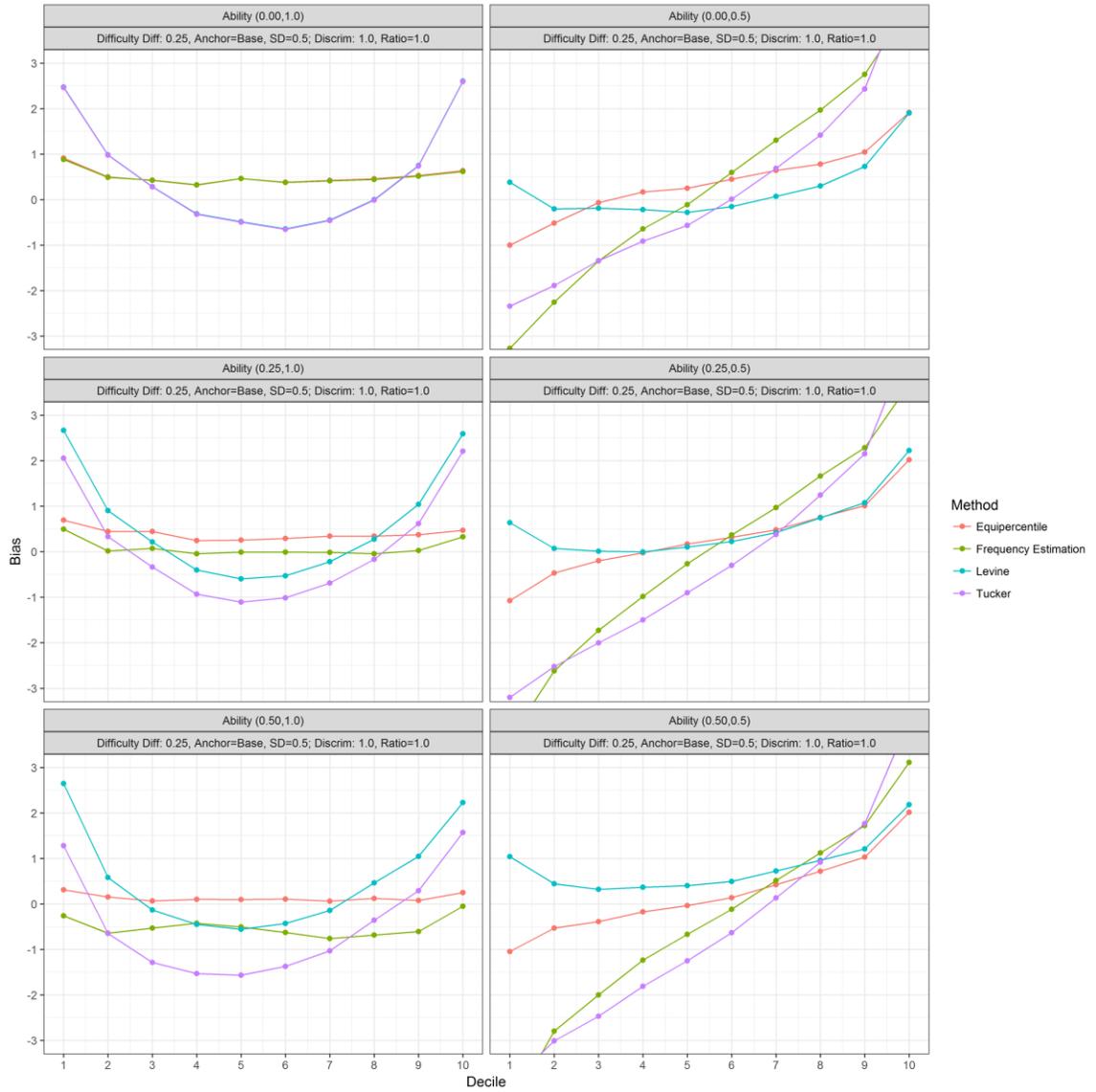


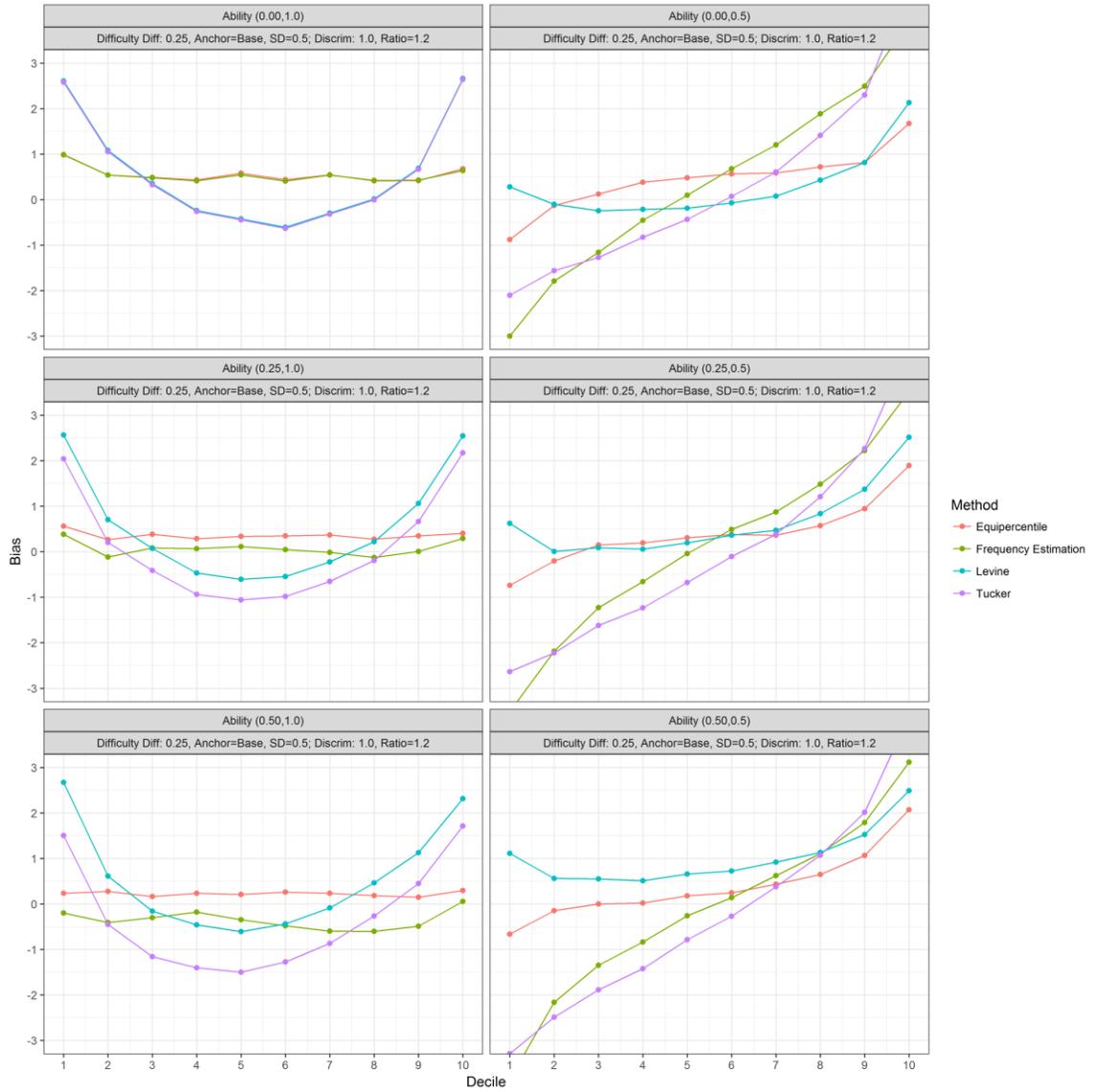


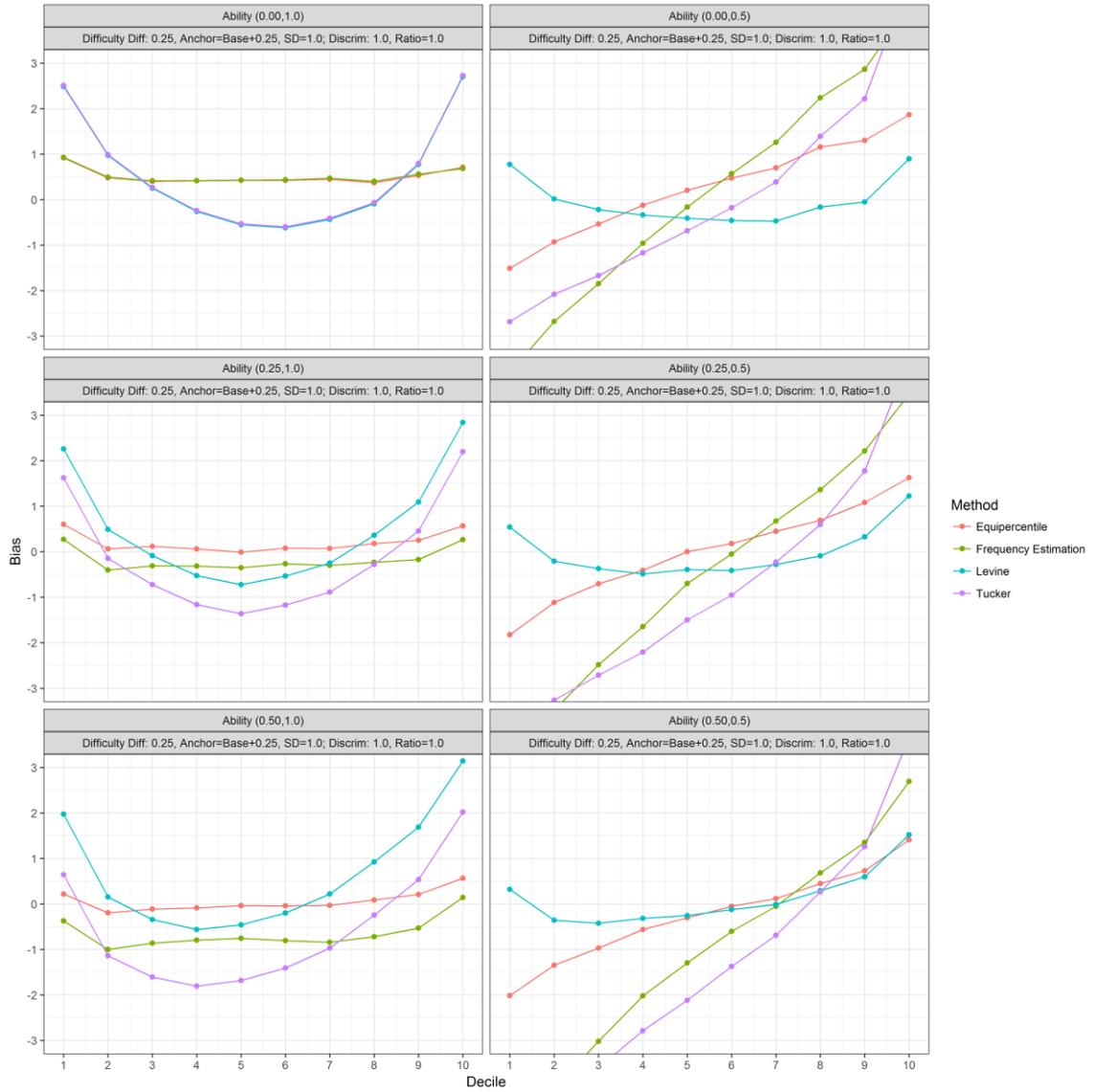


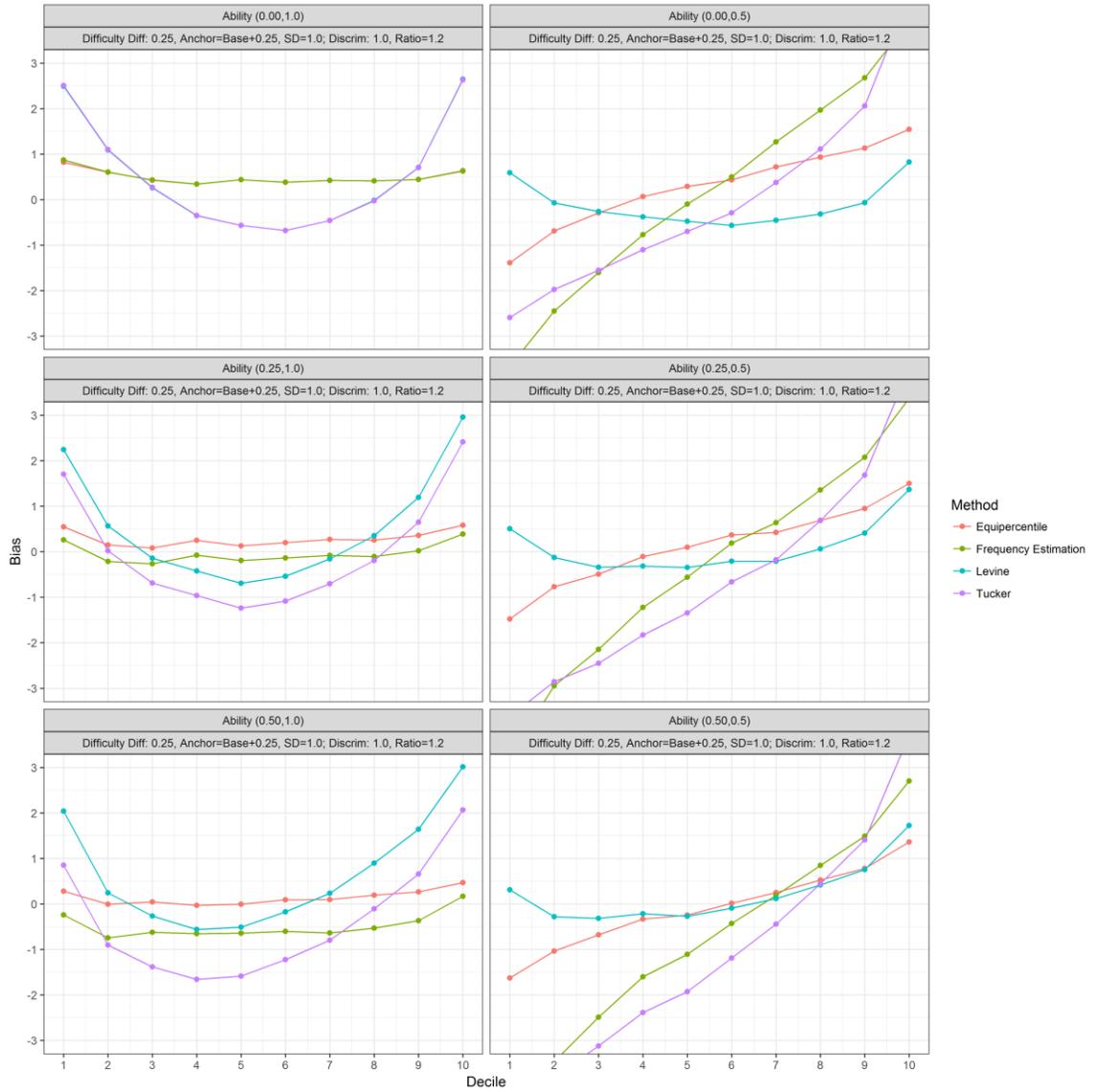


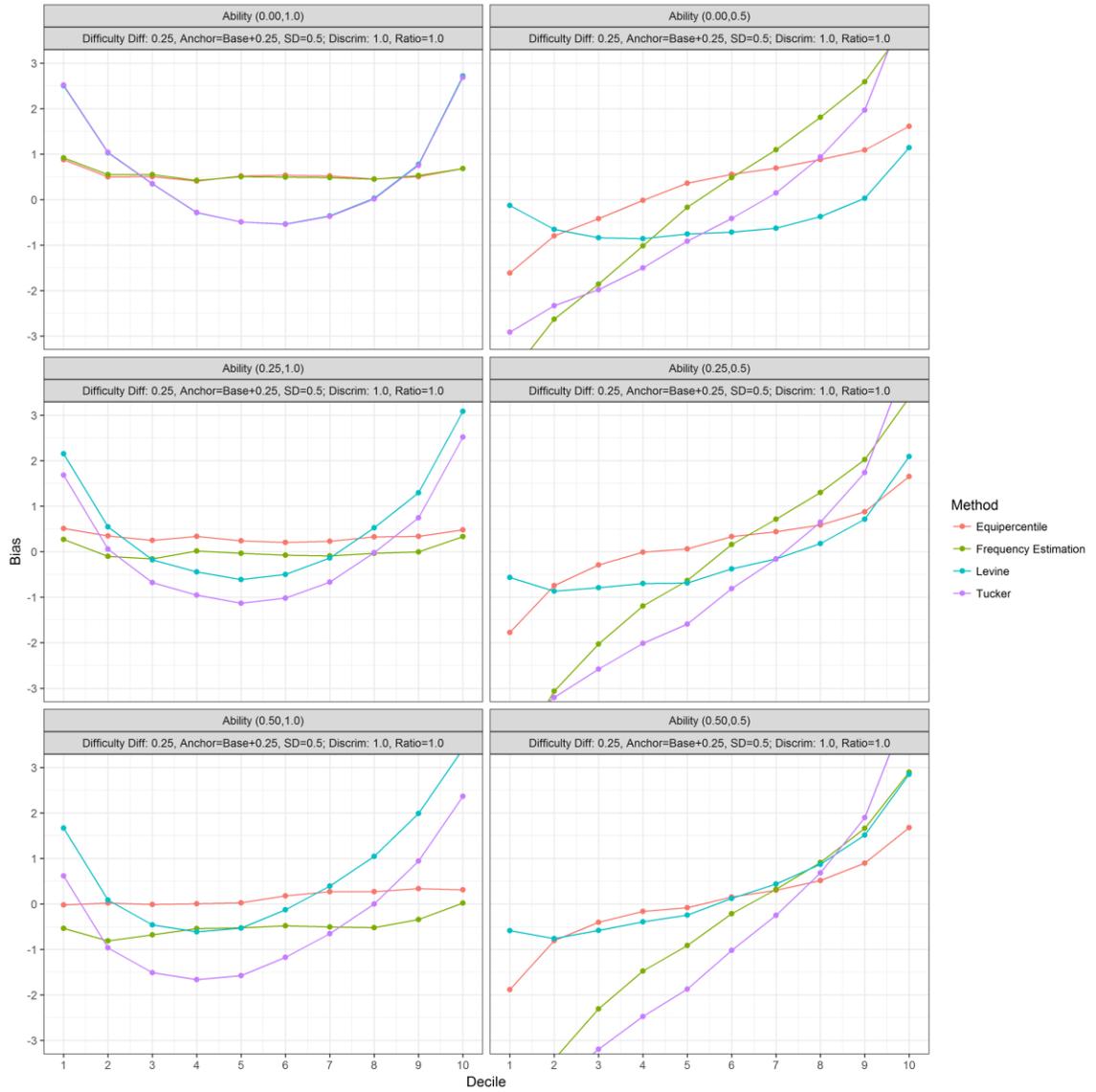


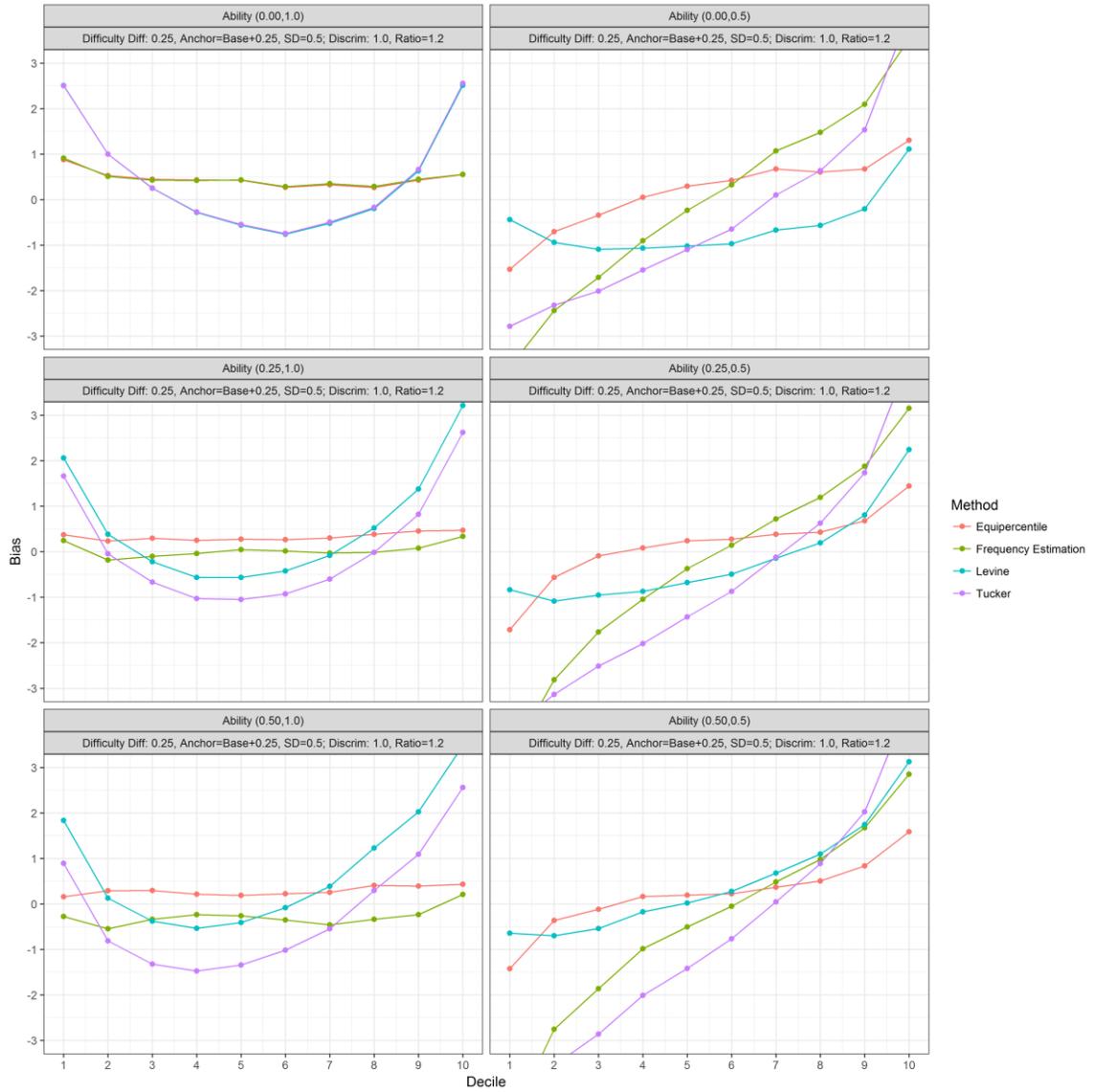


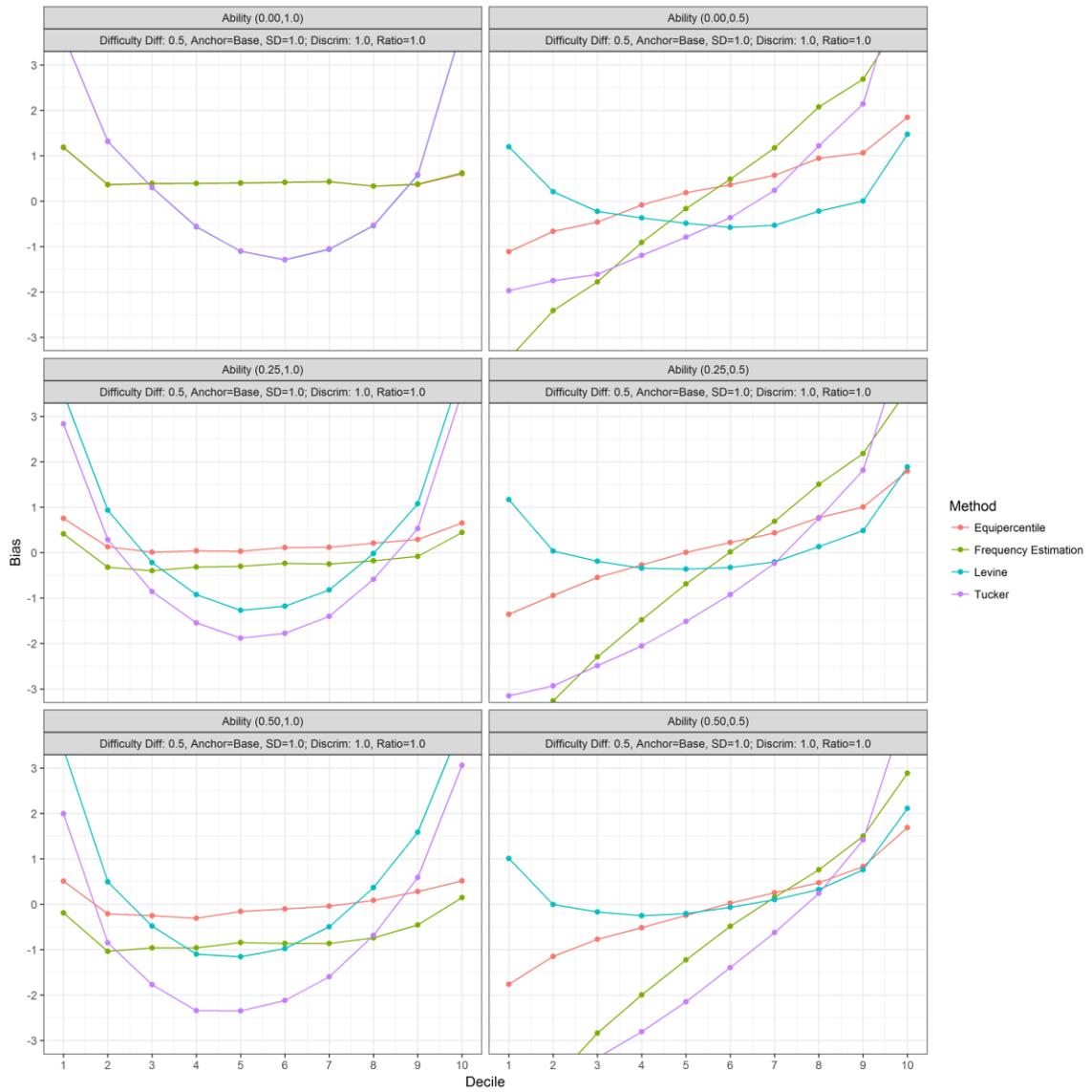


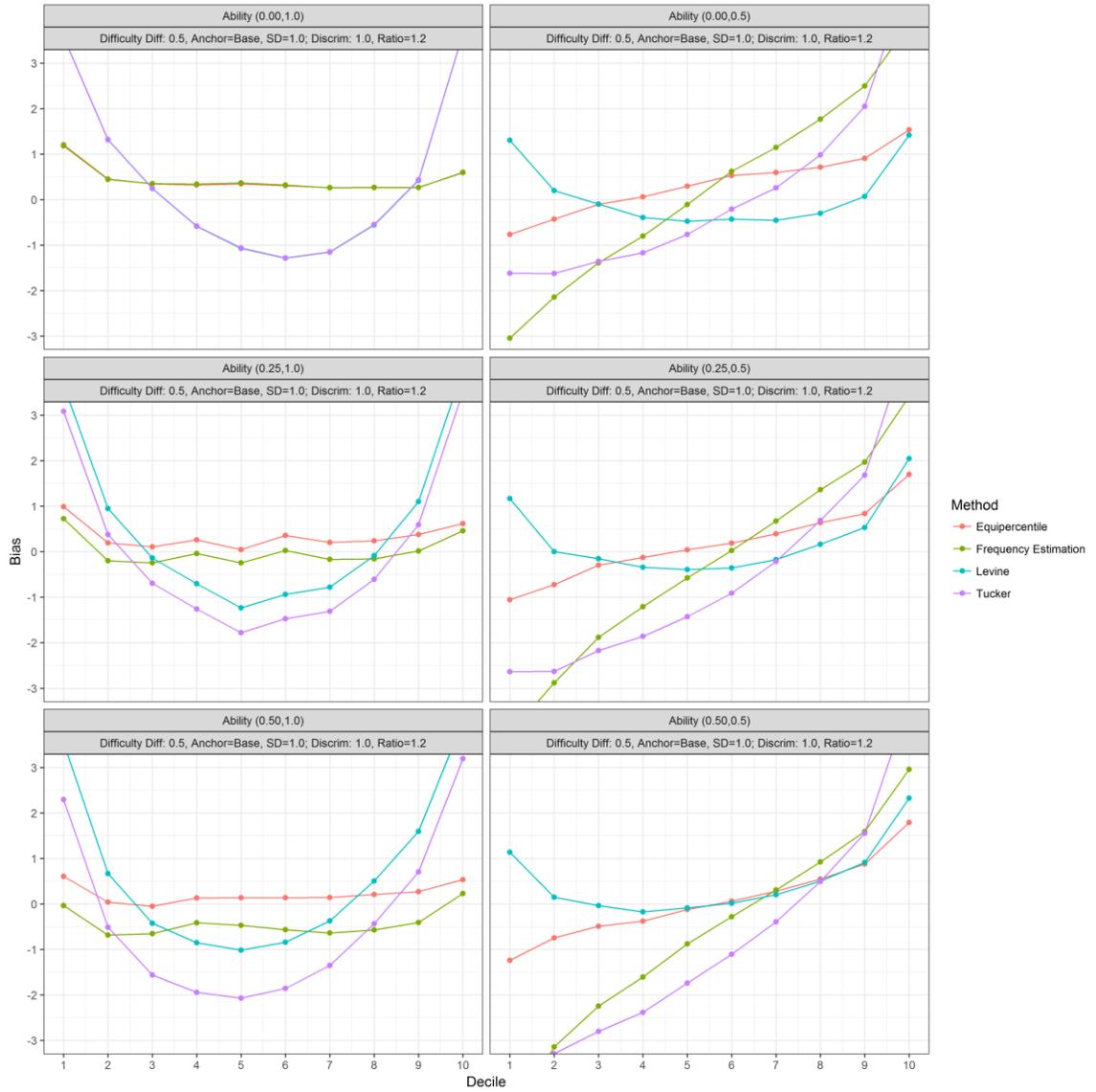


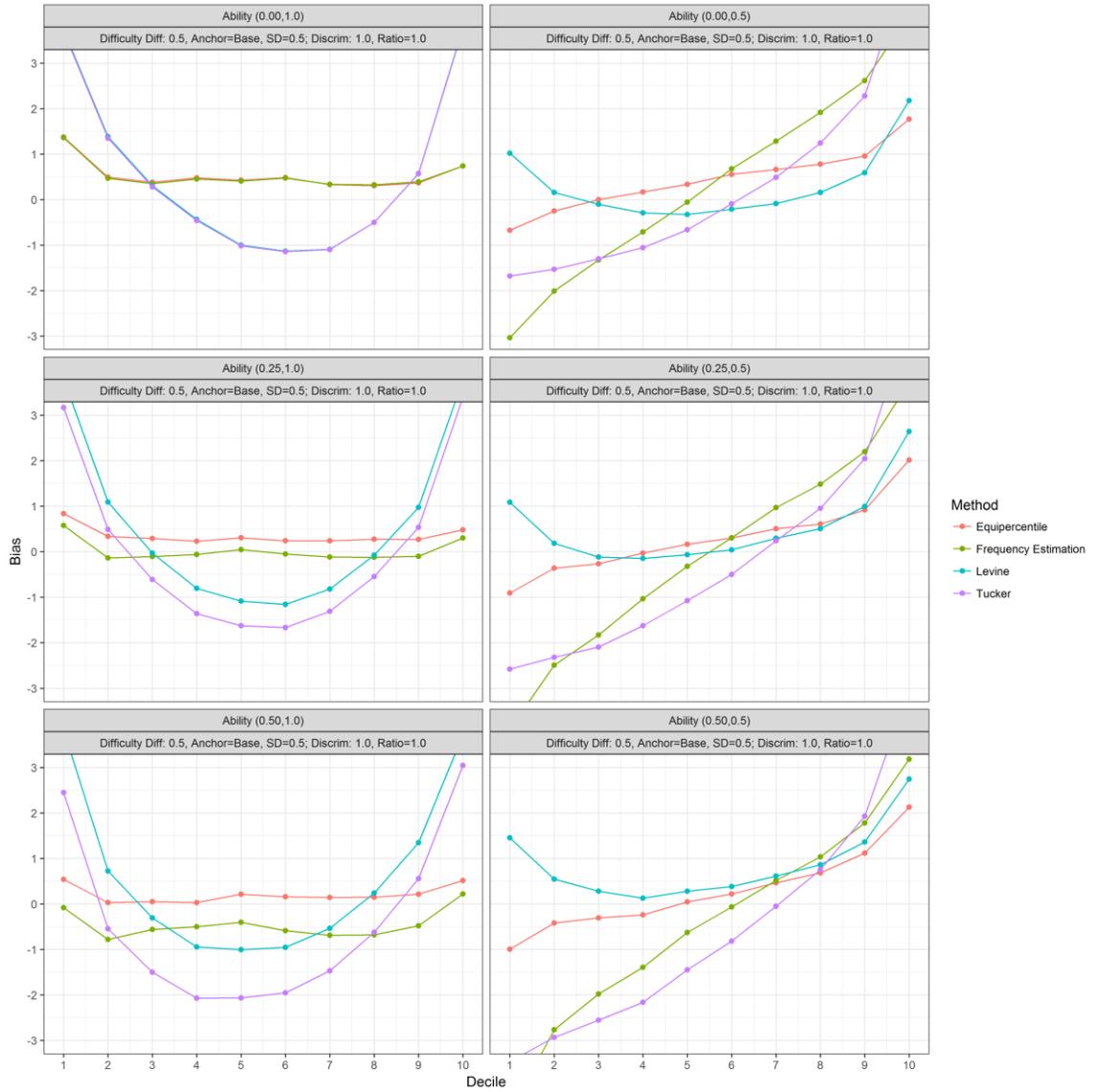


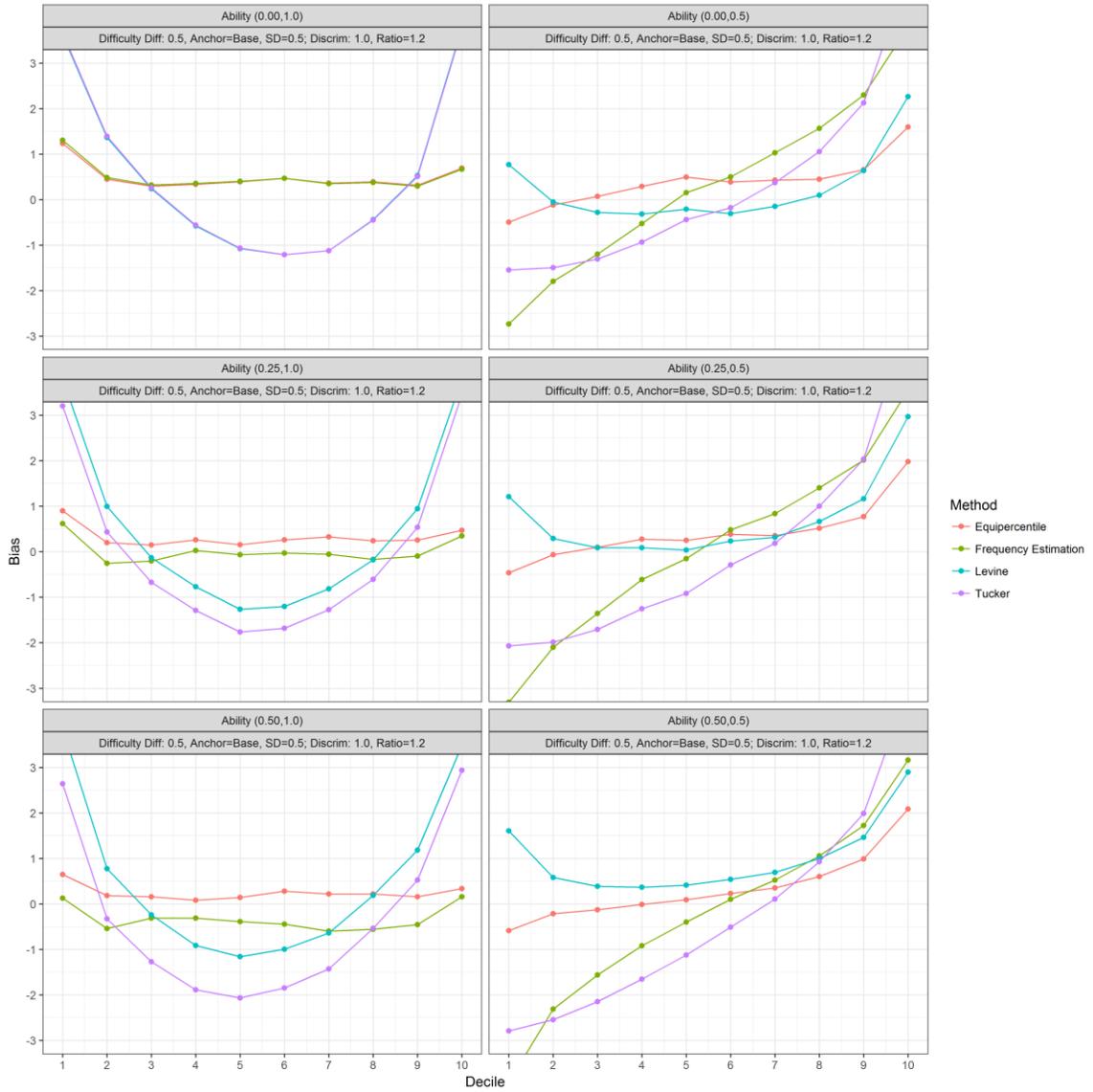


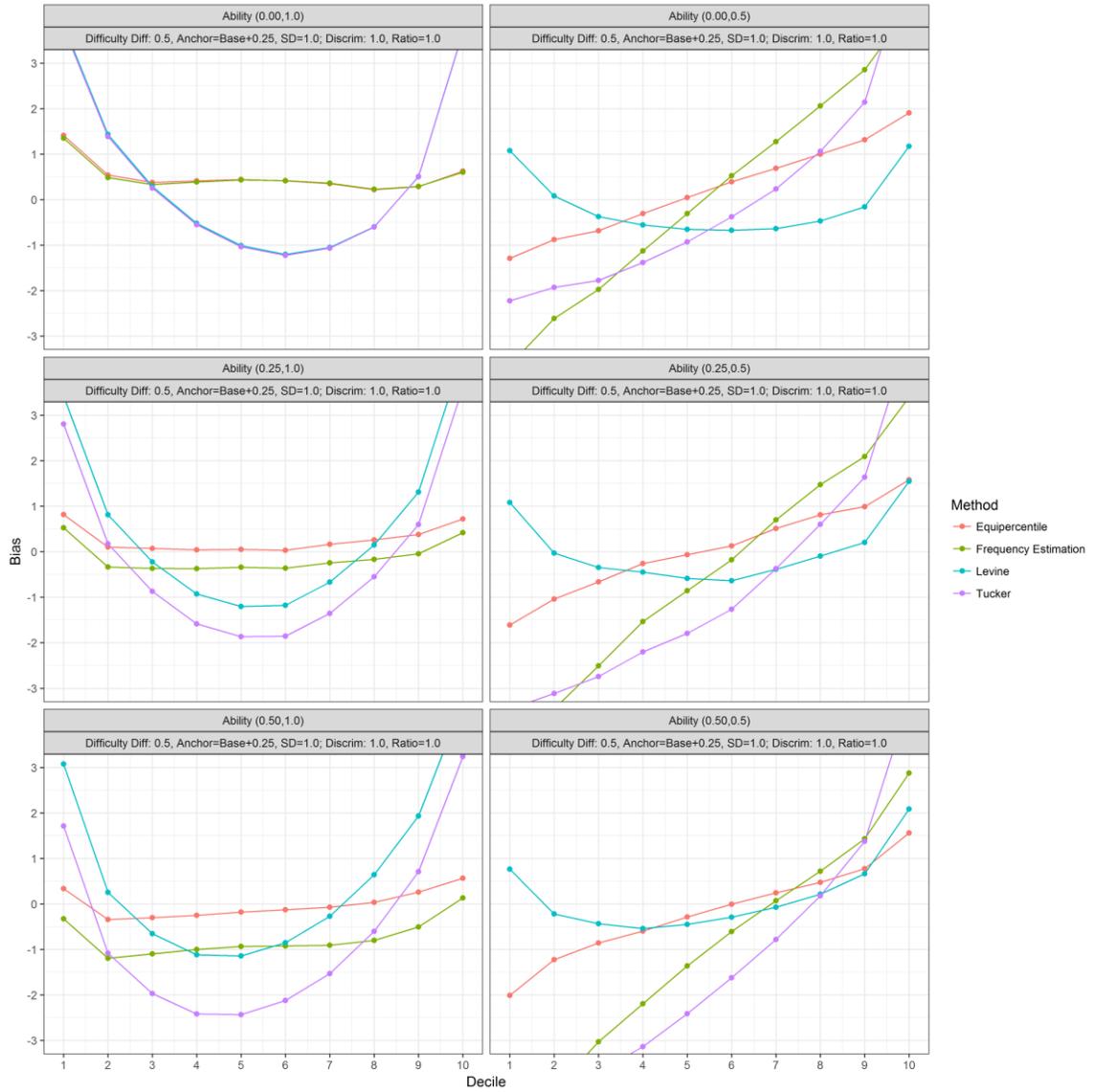


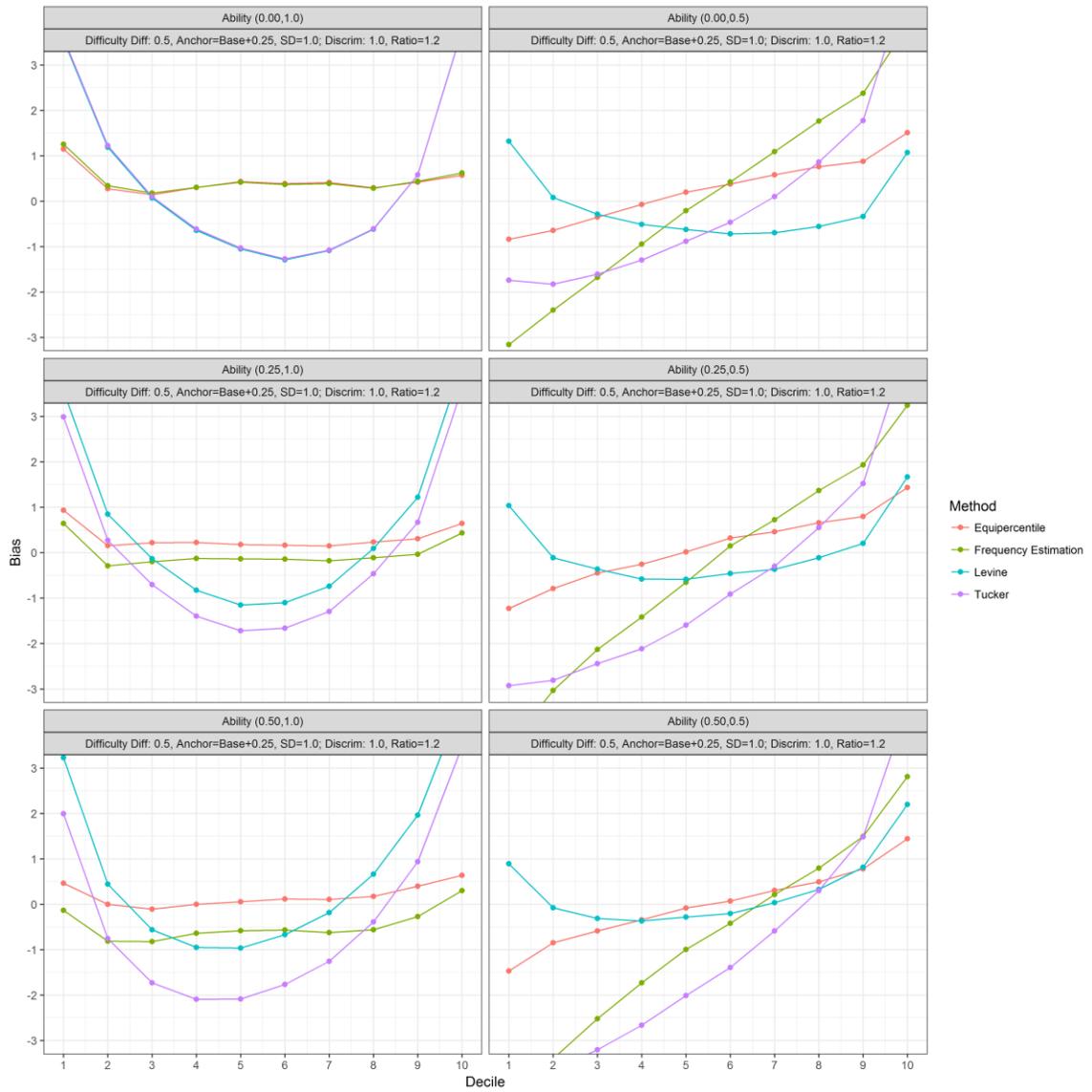


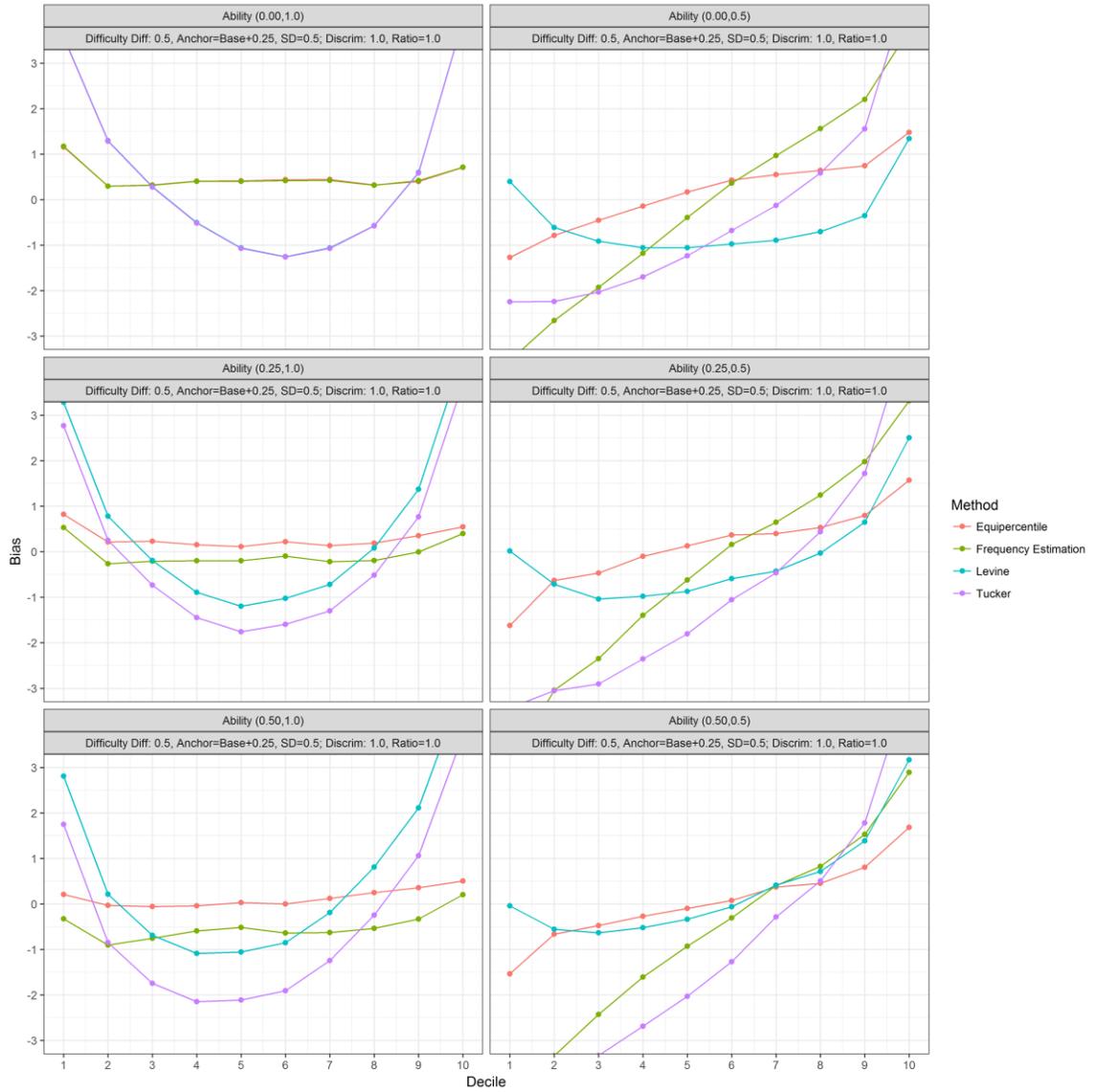


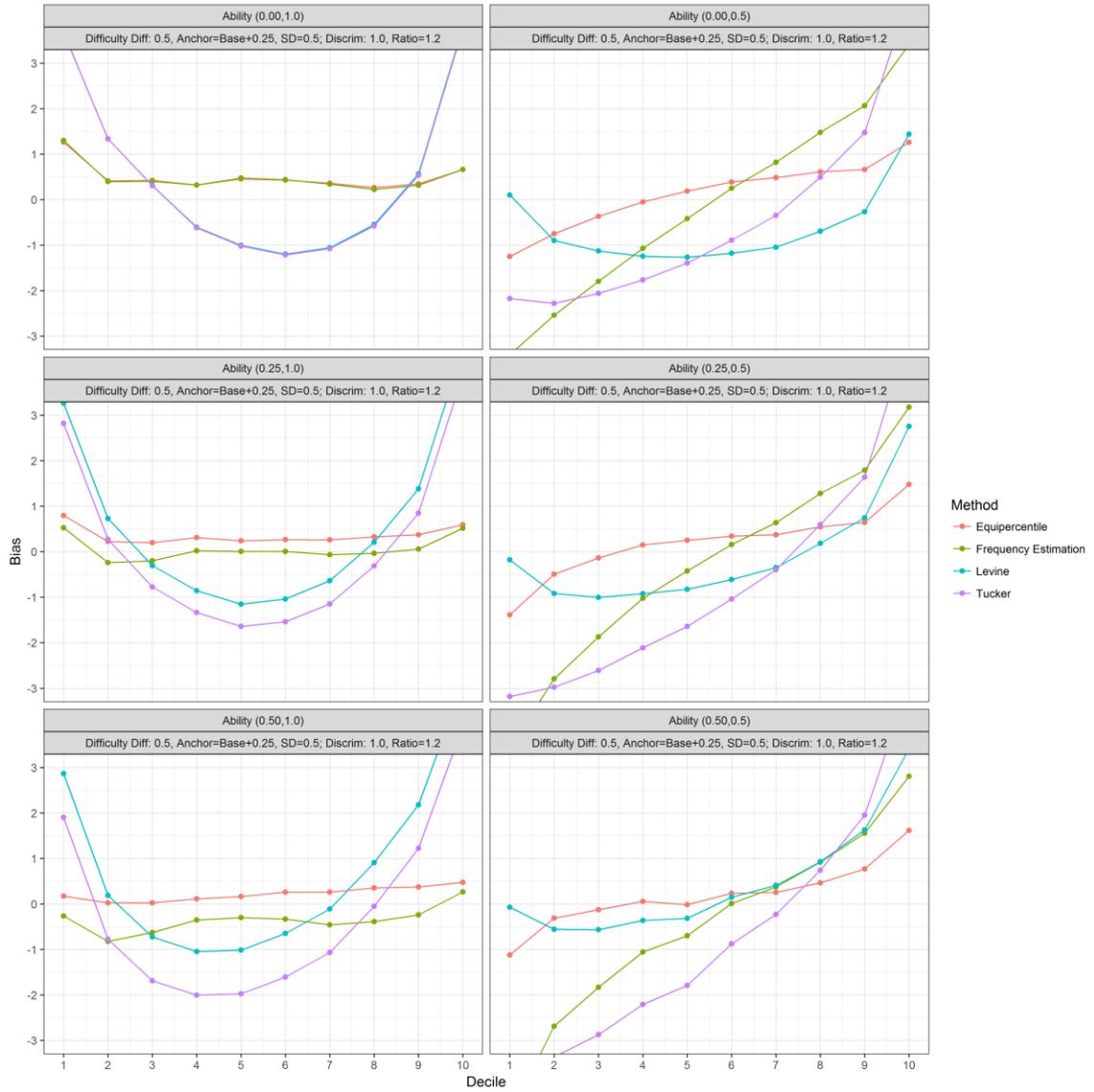




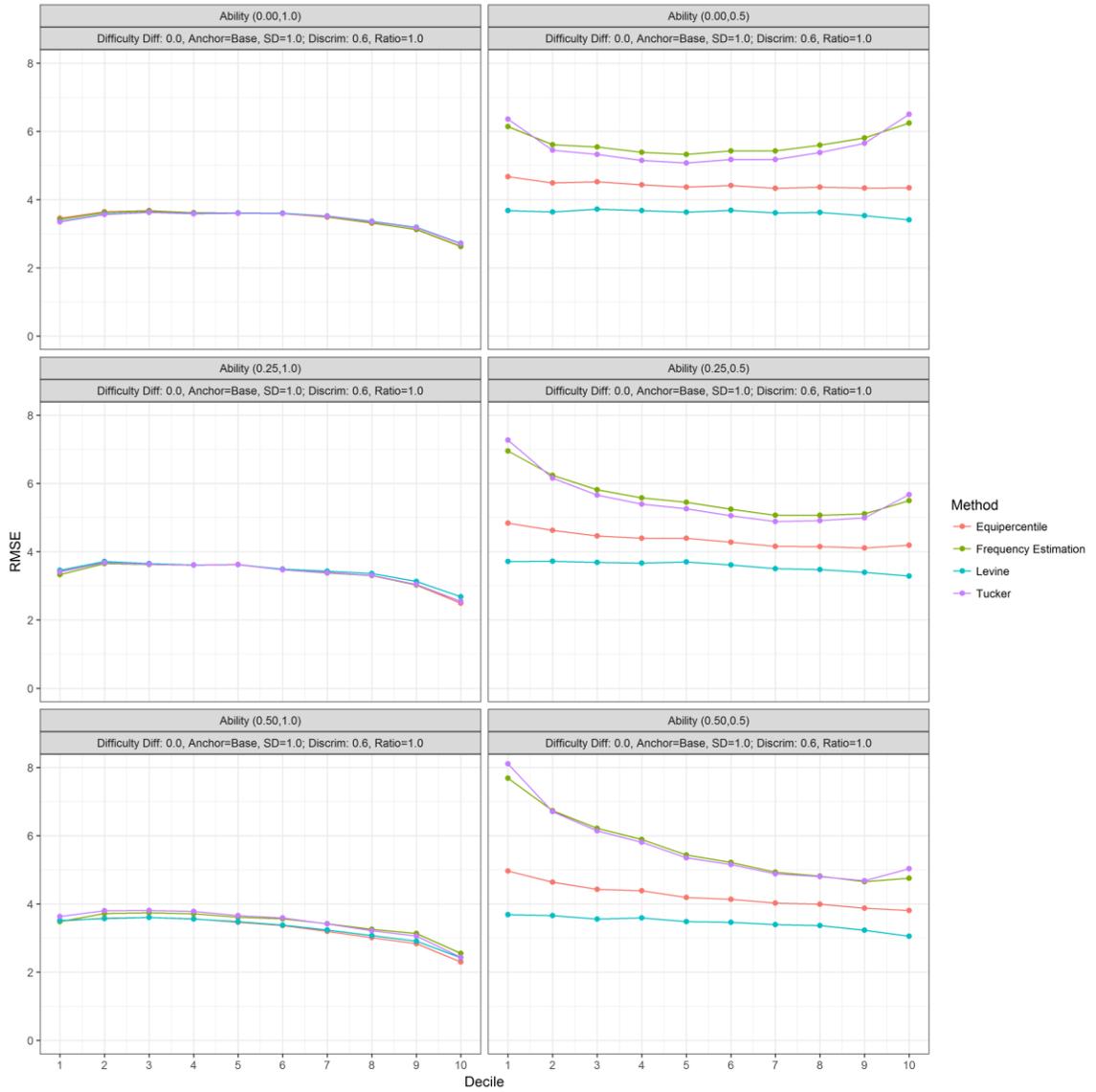


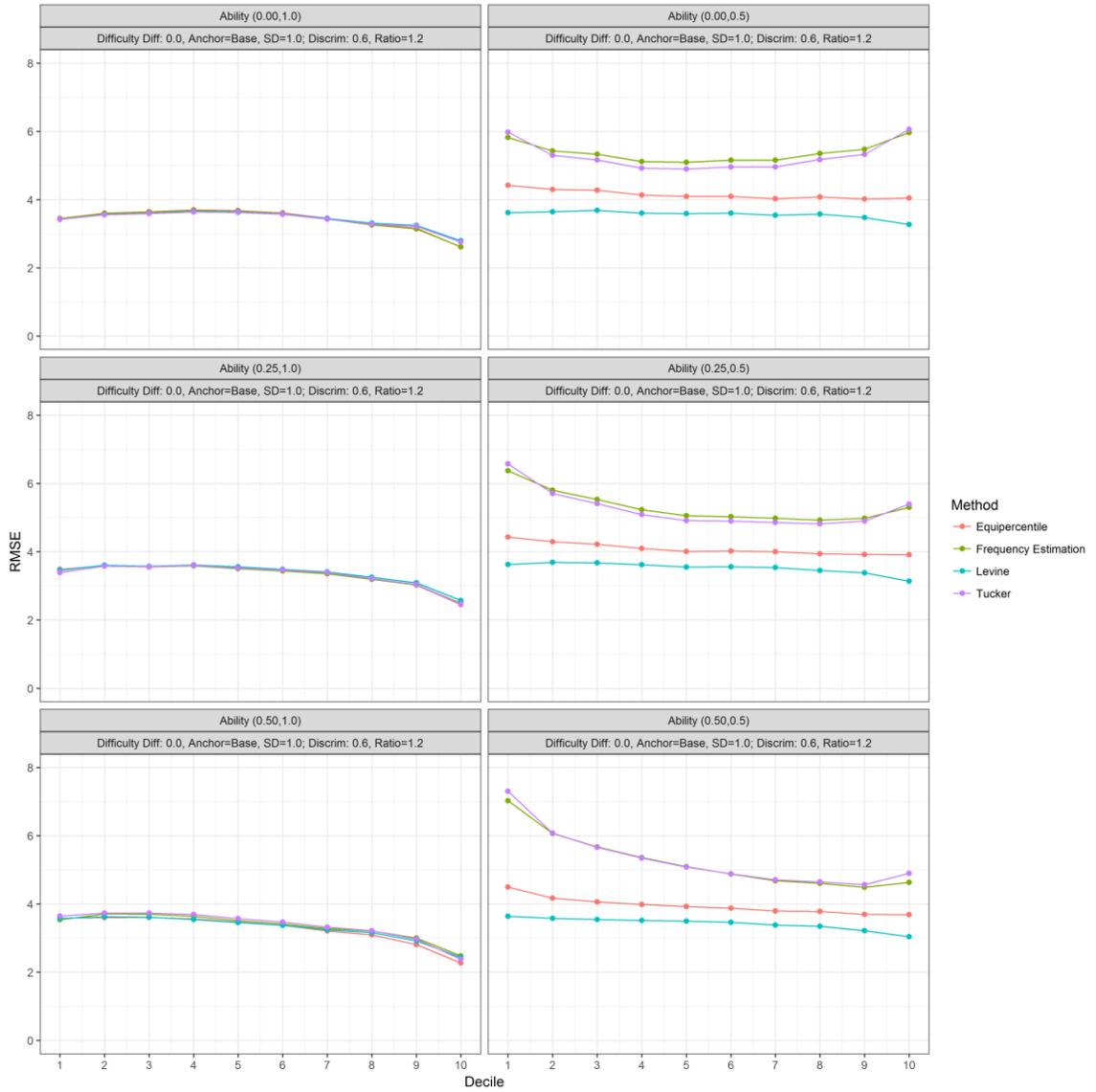


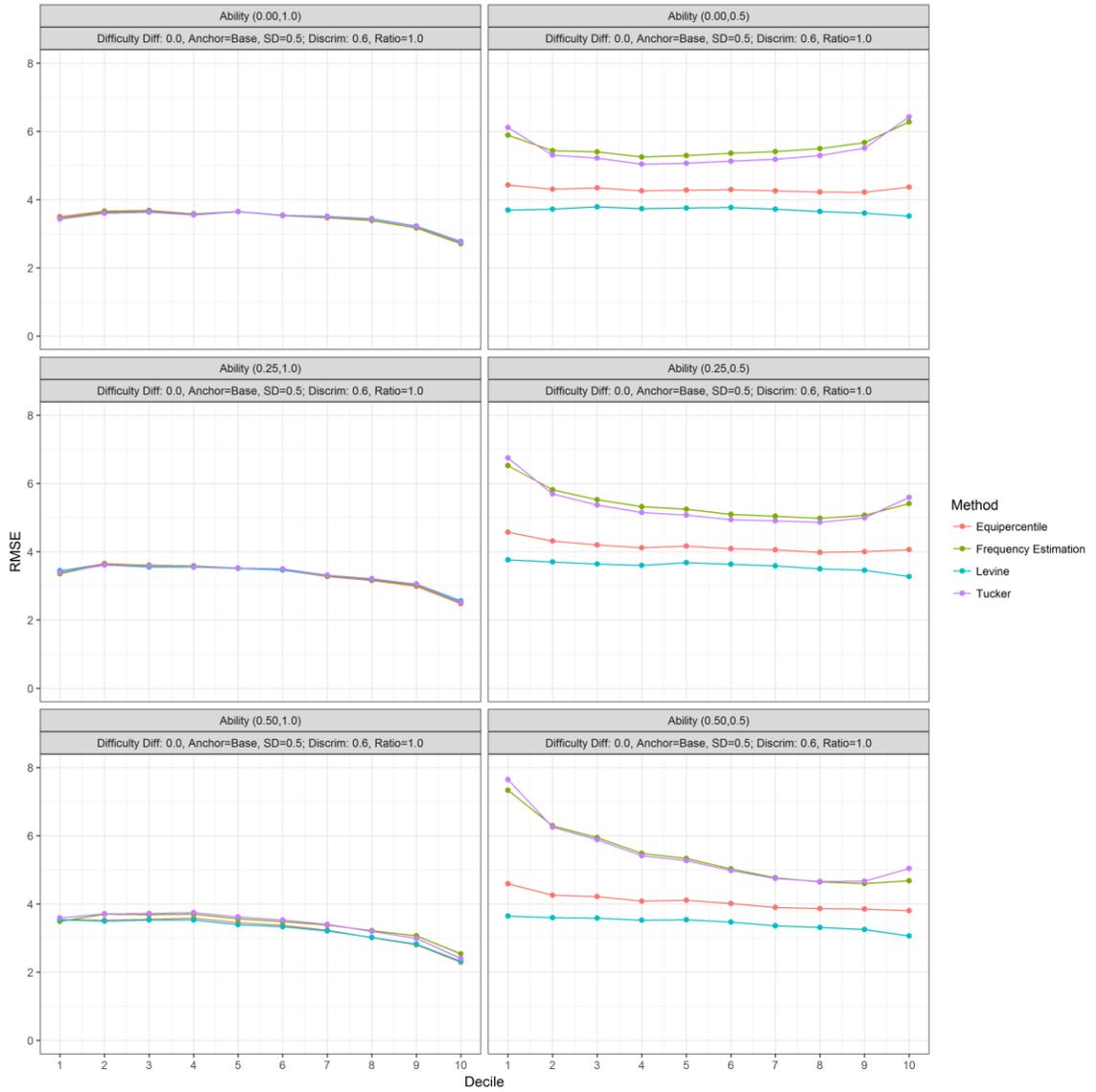


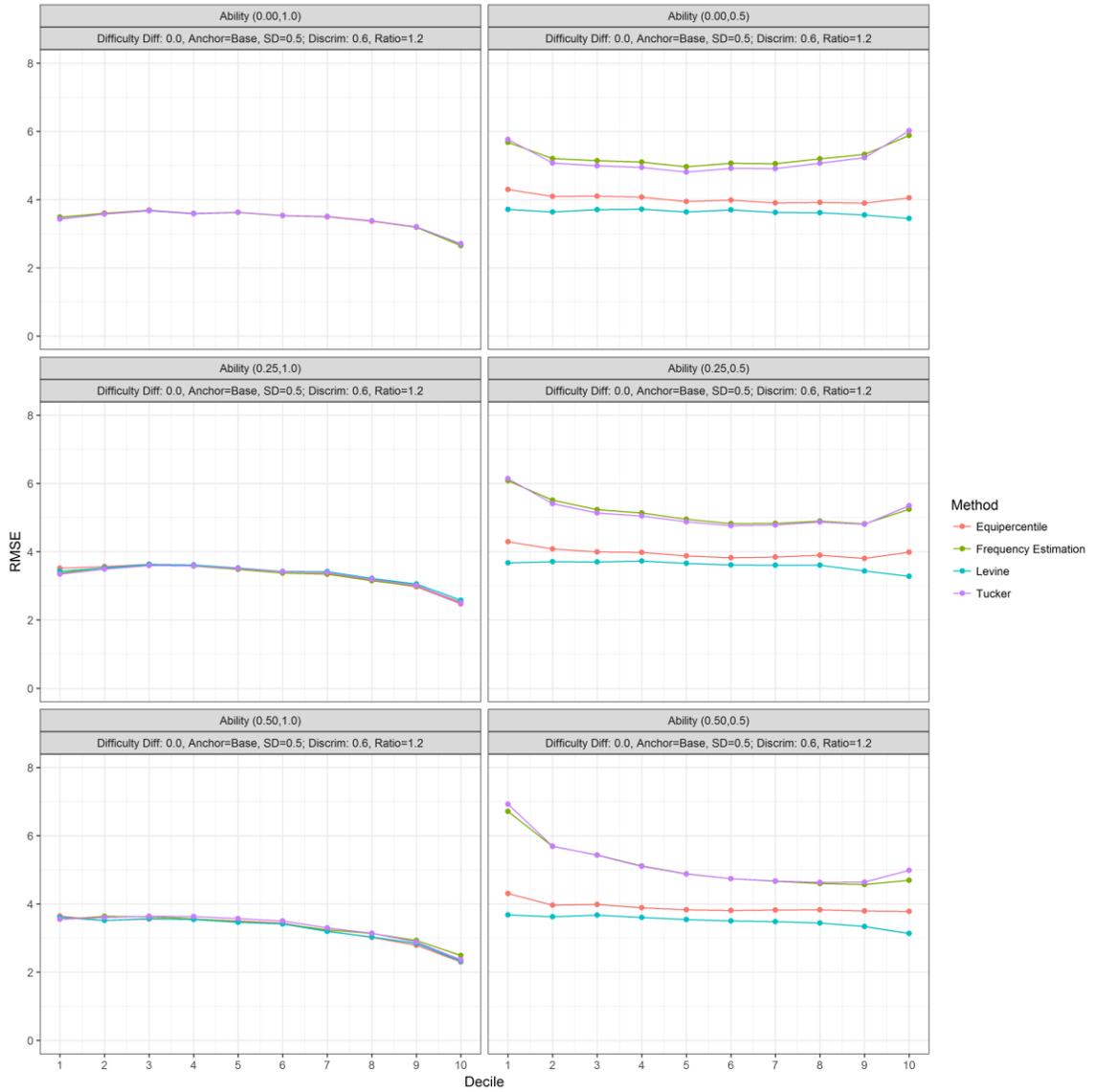


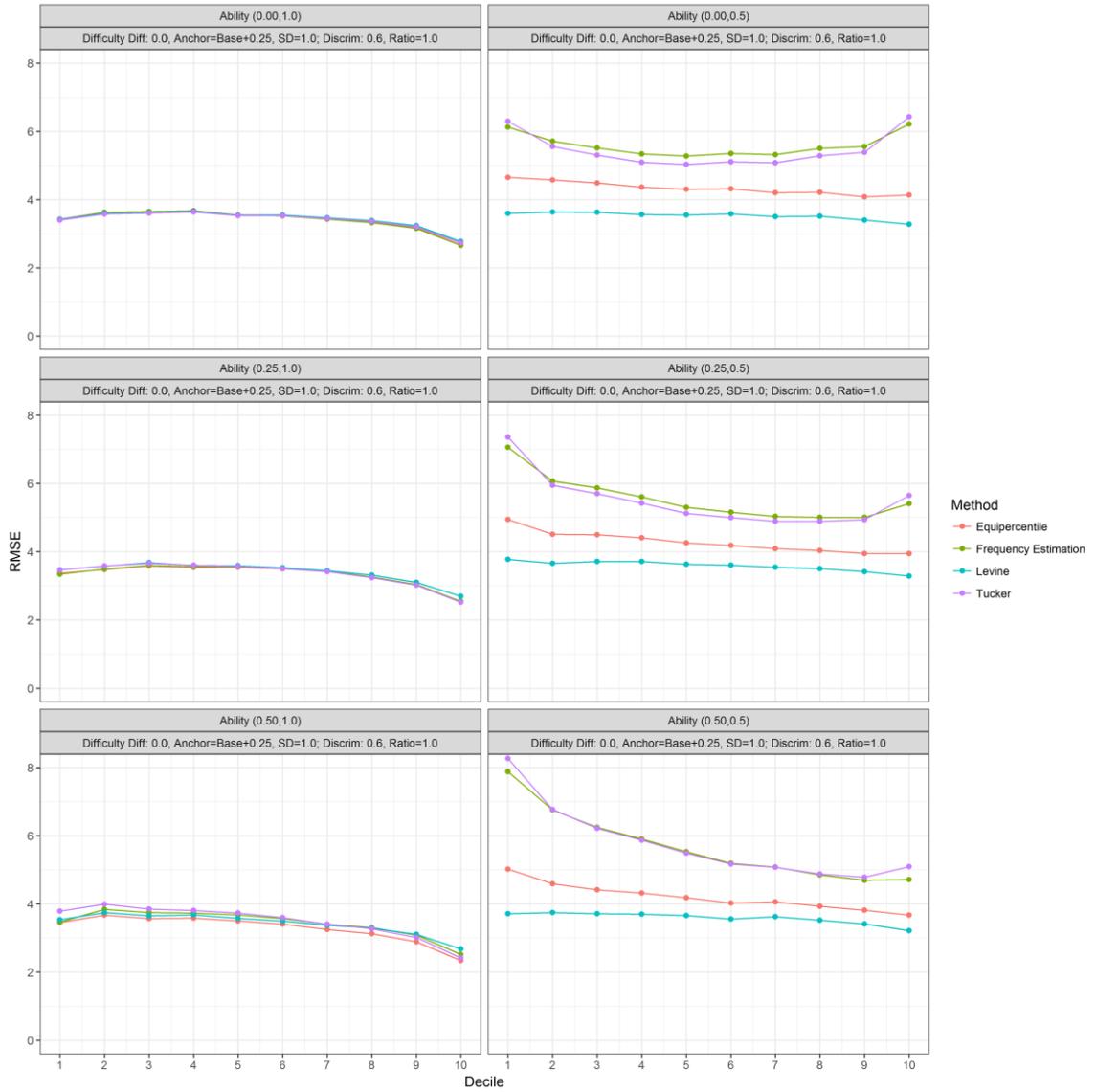
APPENDIX B  
ACHIEVEMENT RMSE RESULTS

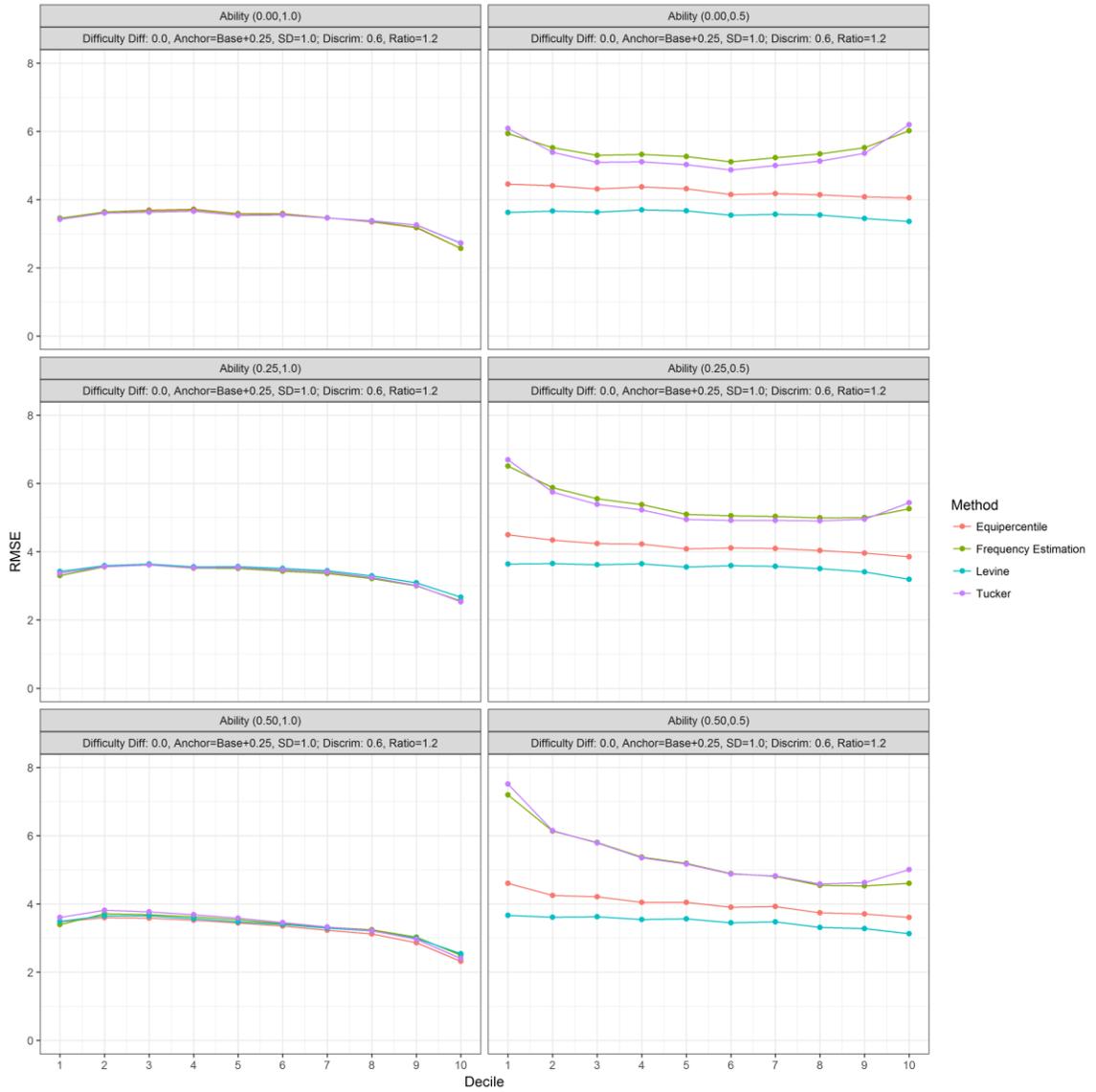


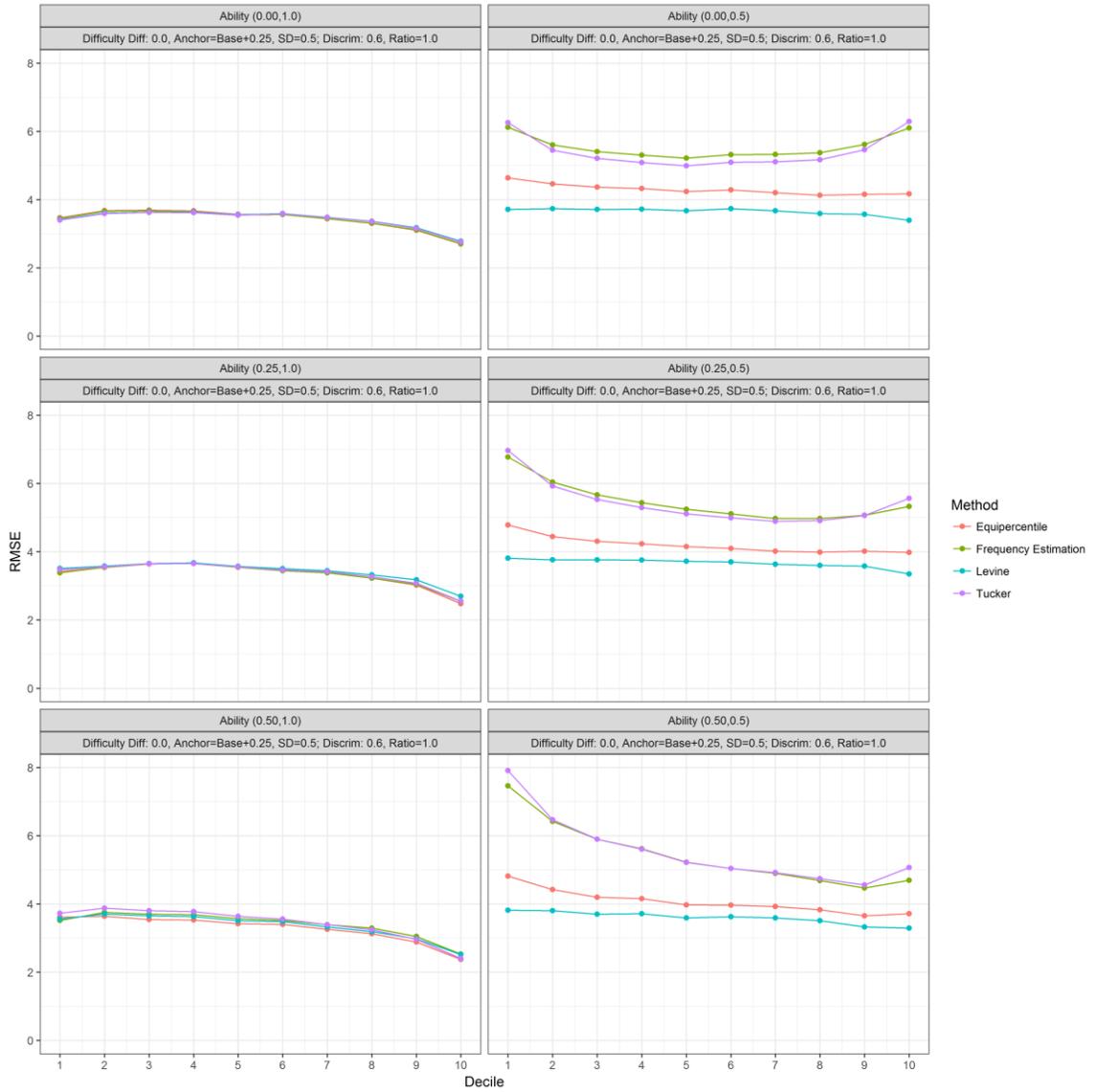


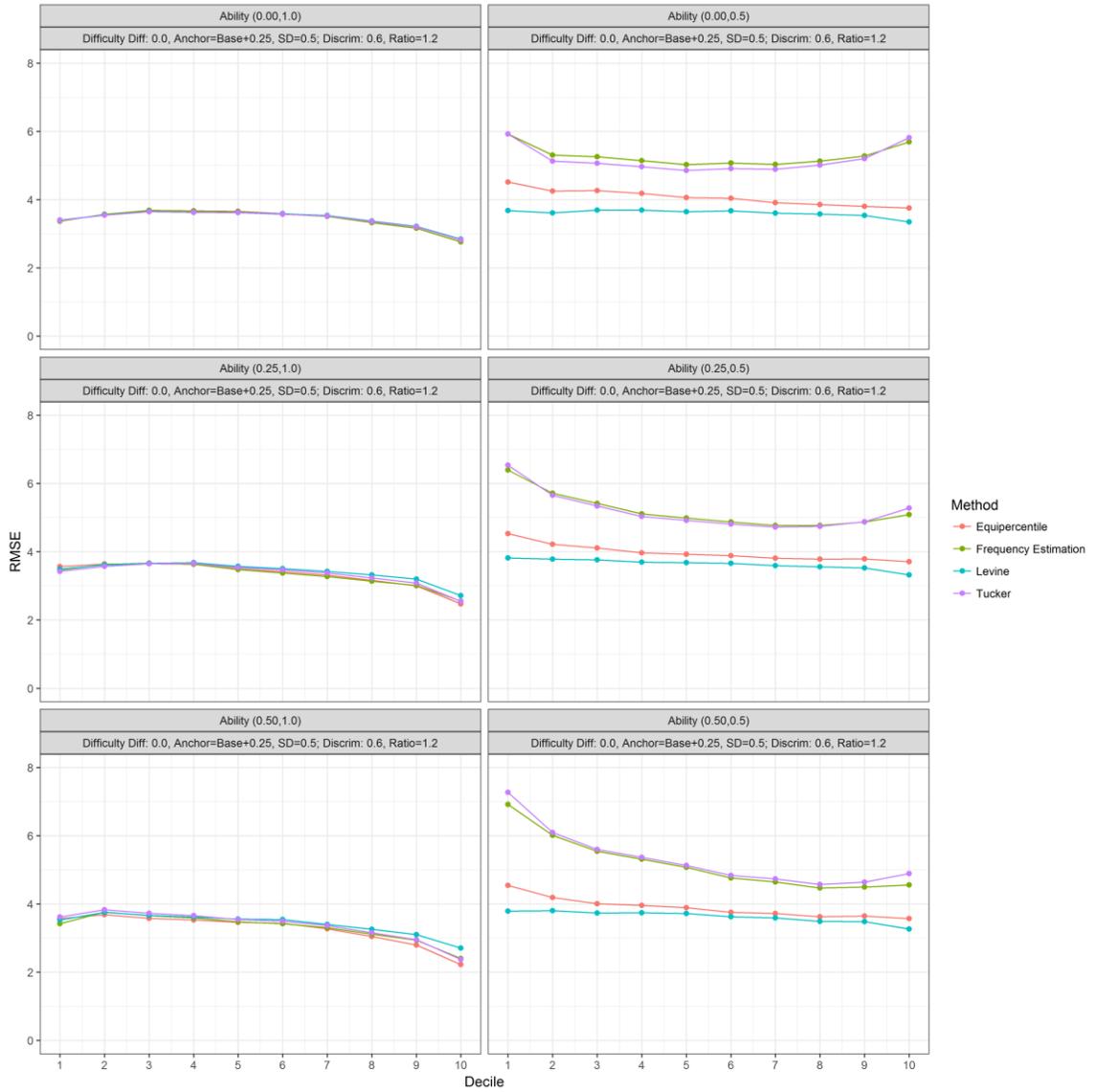


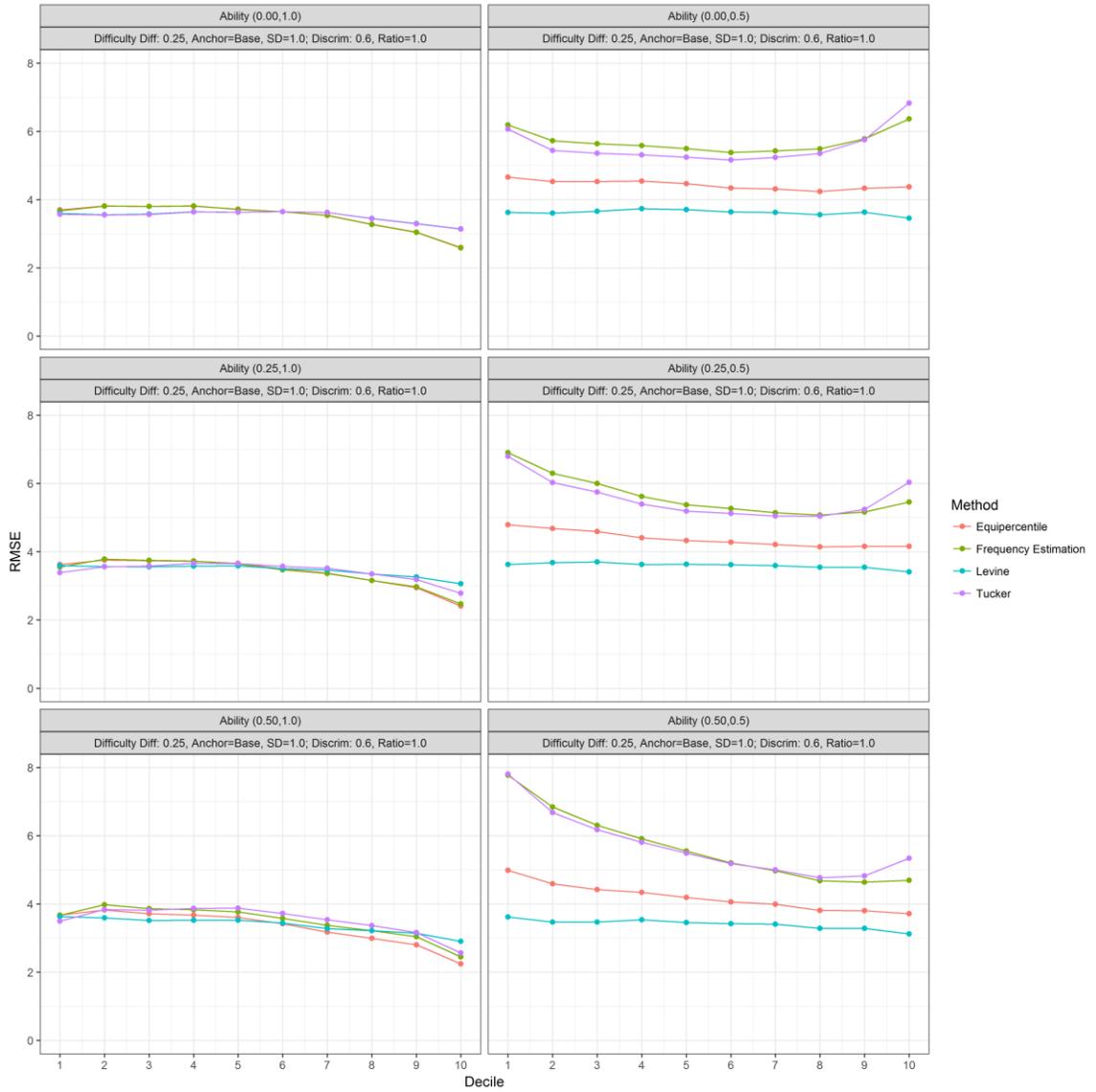


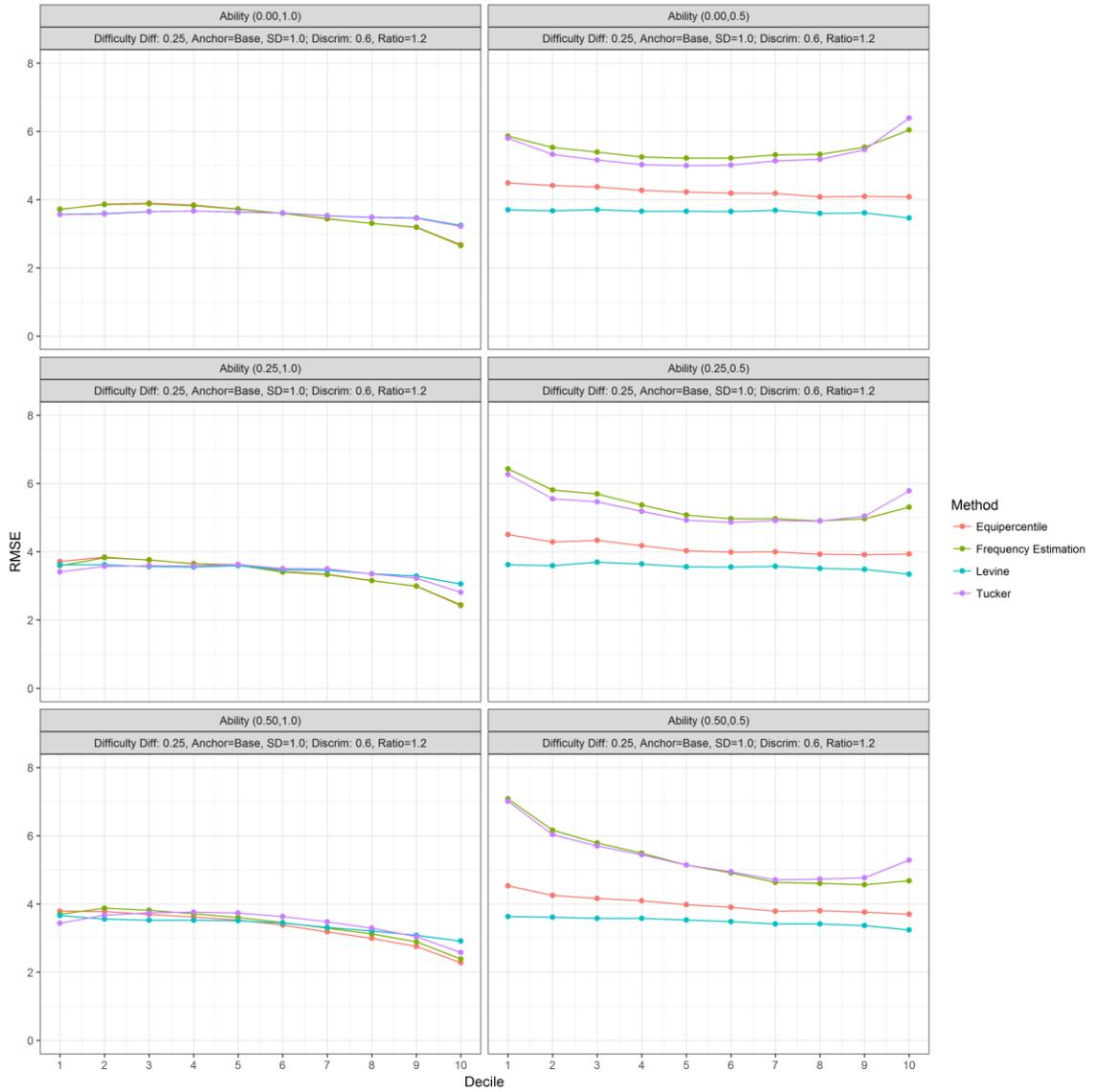


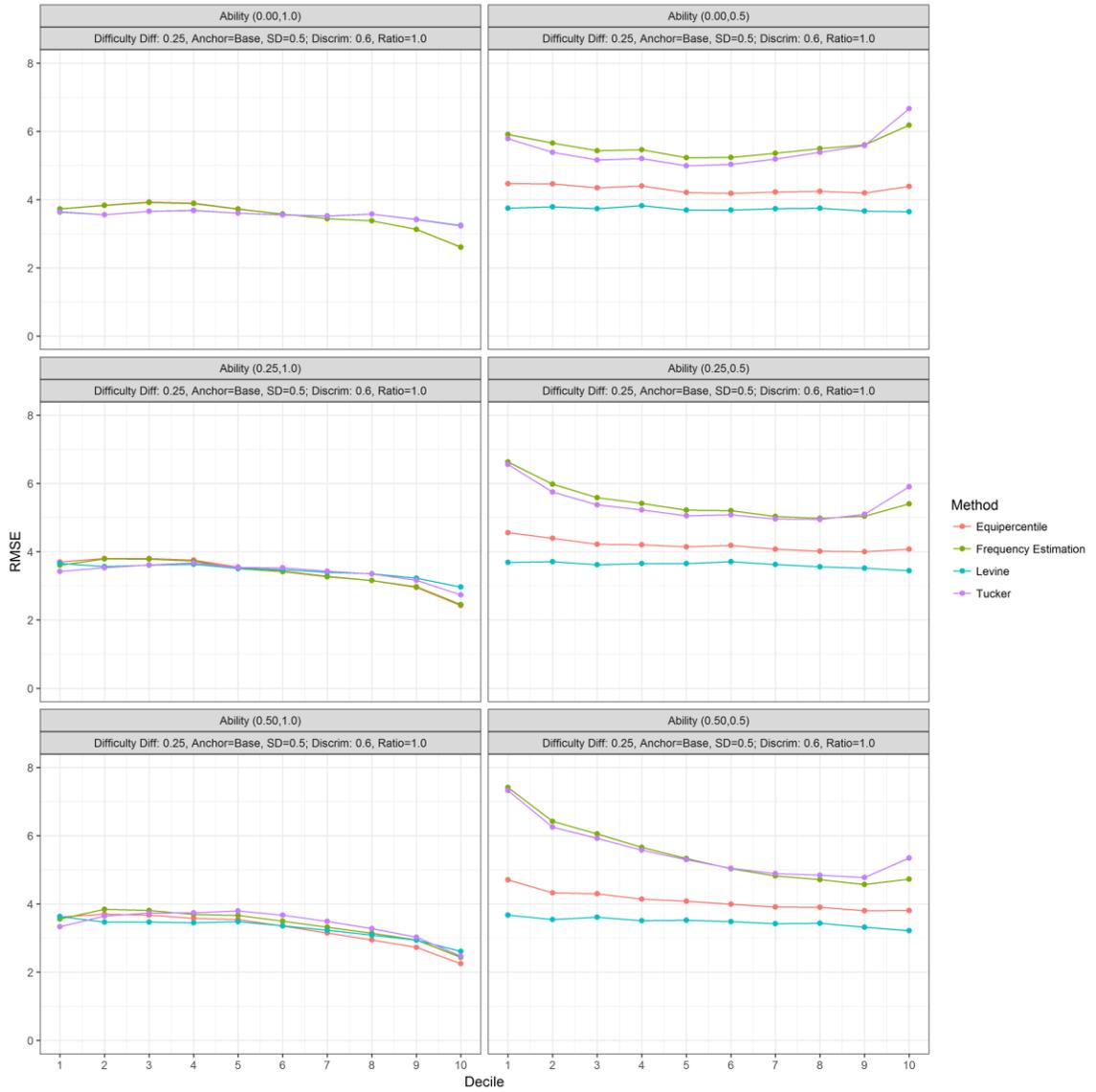


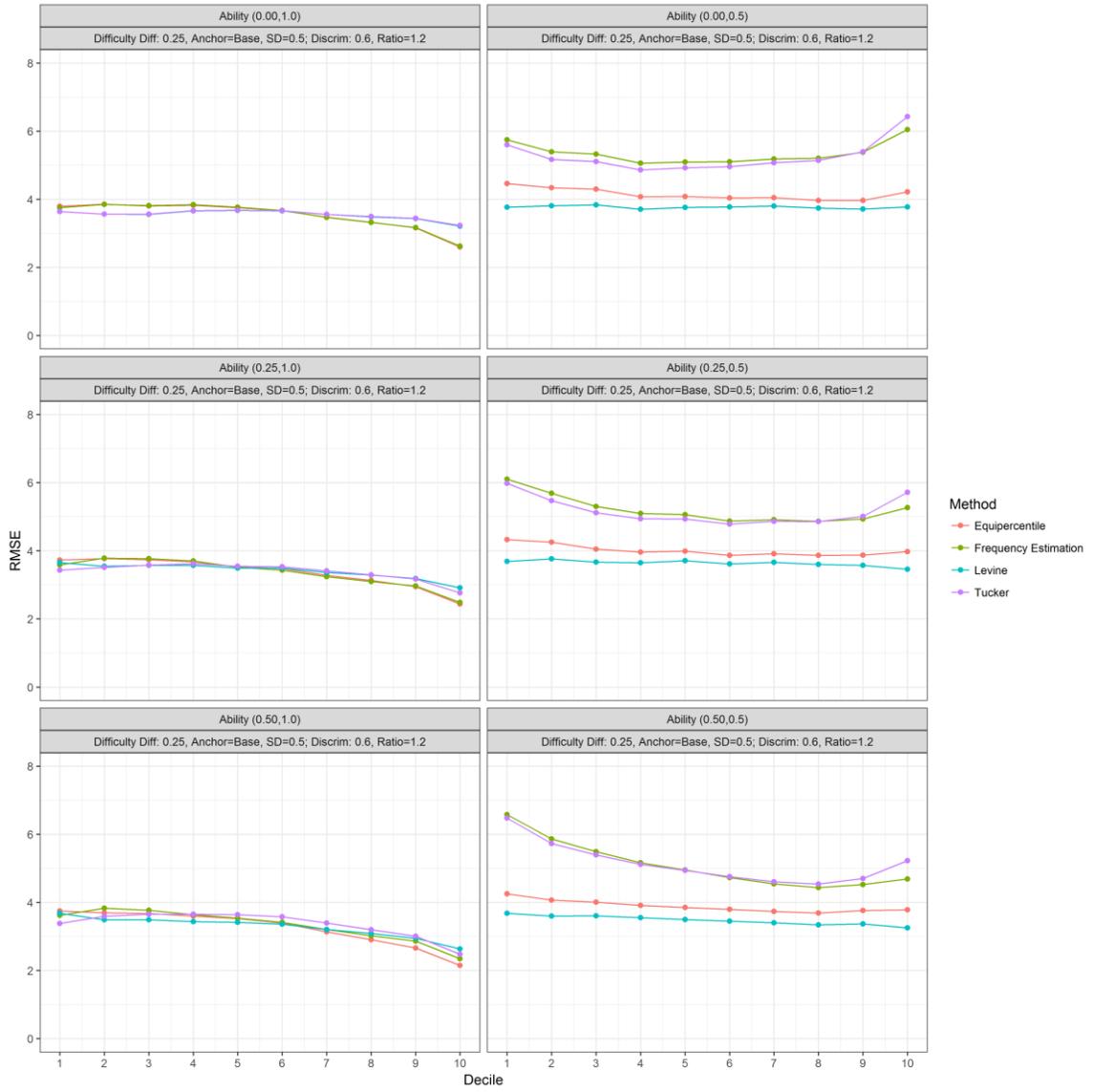


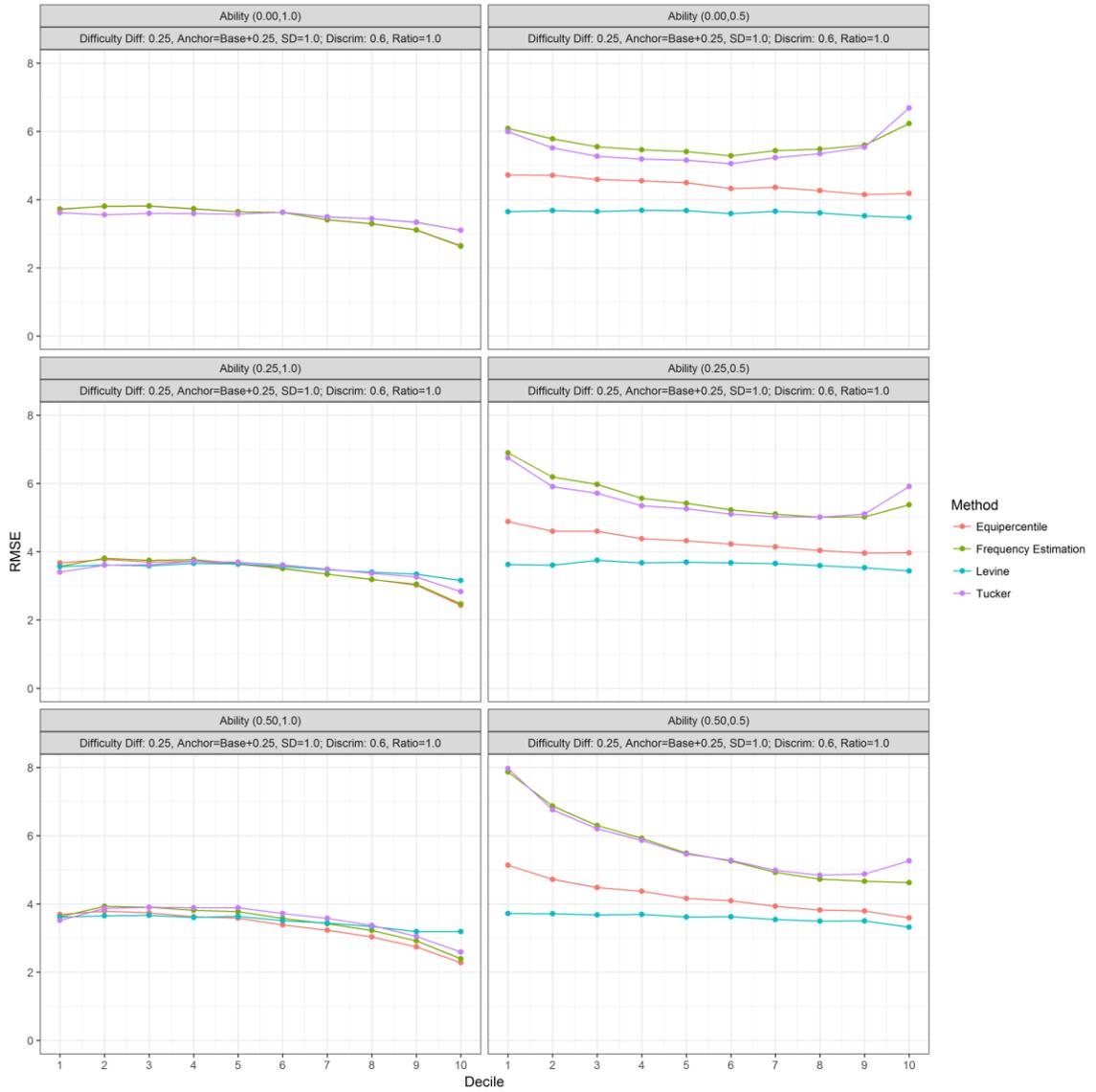


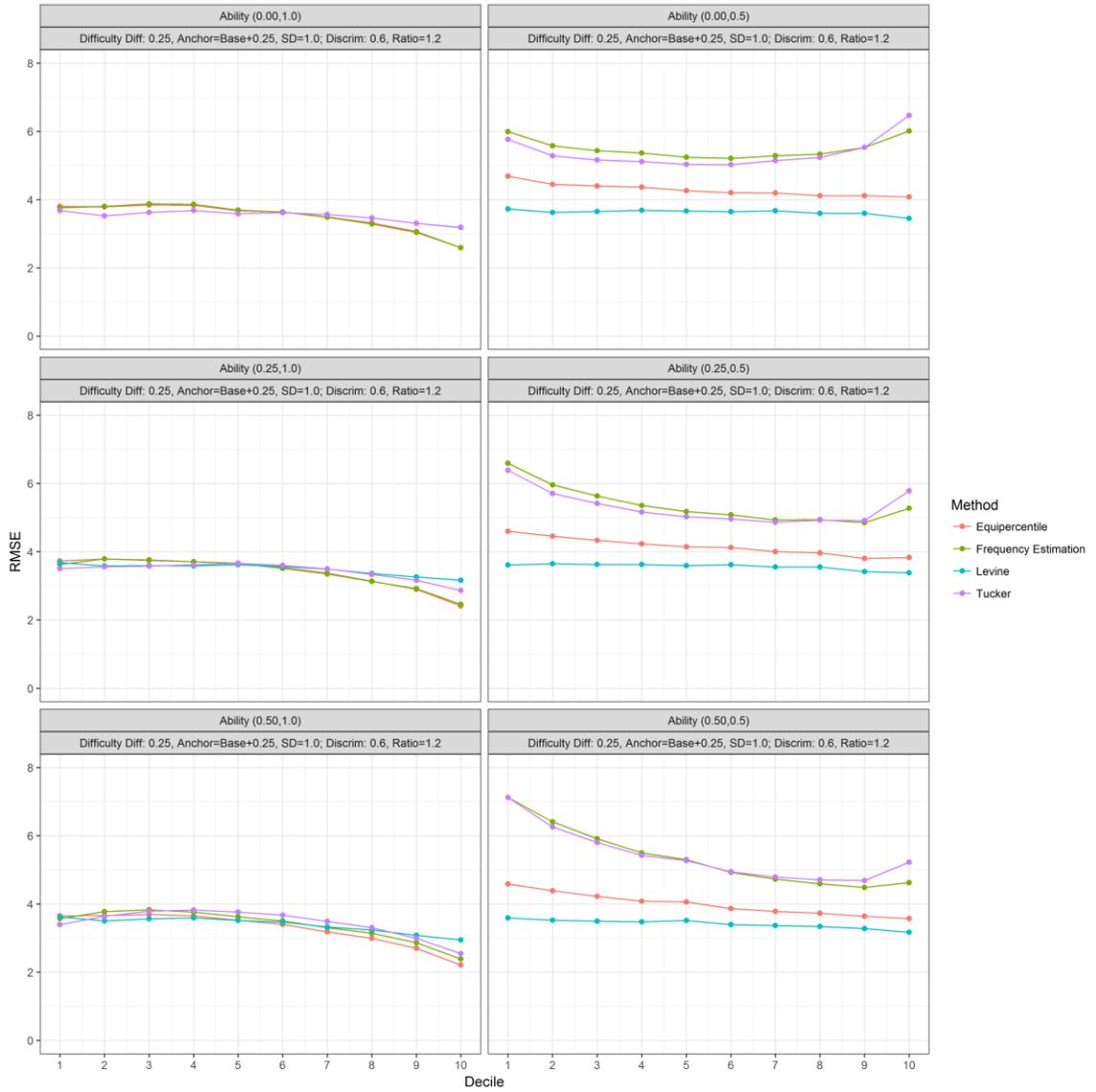


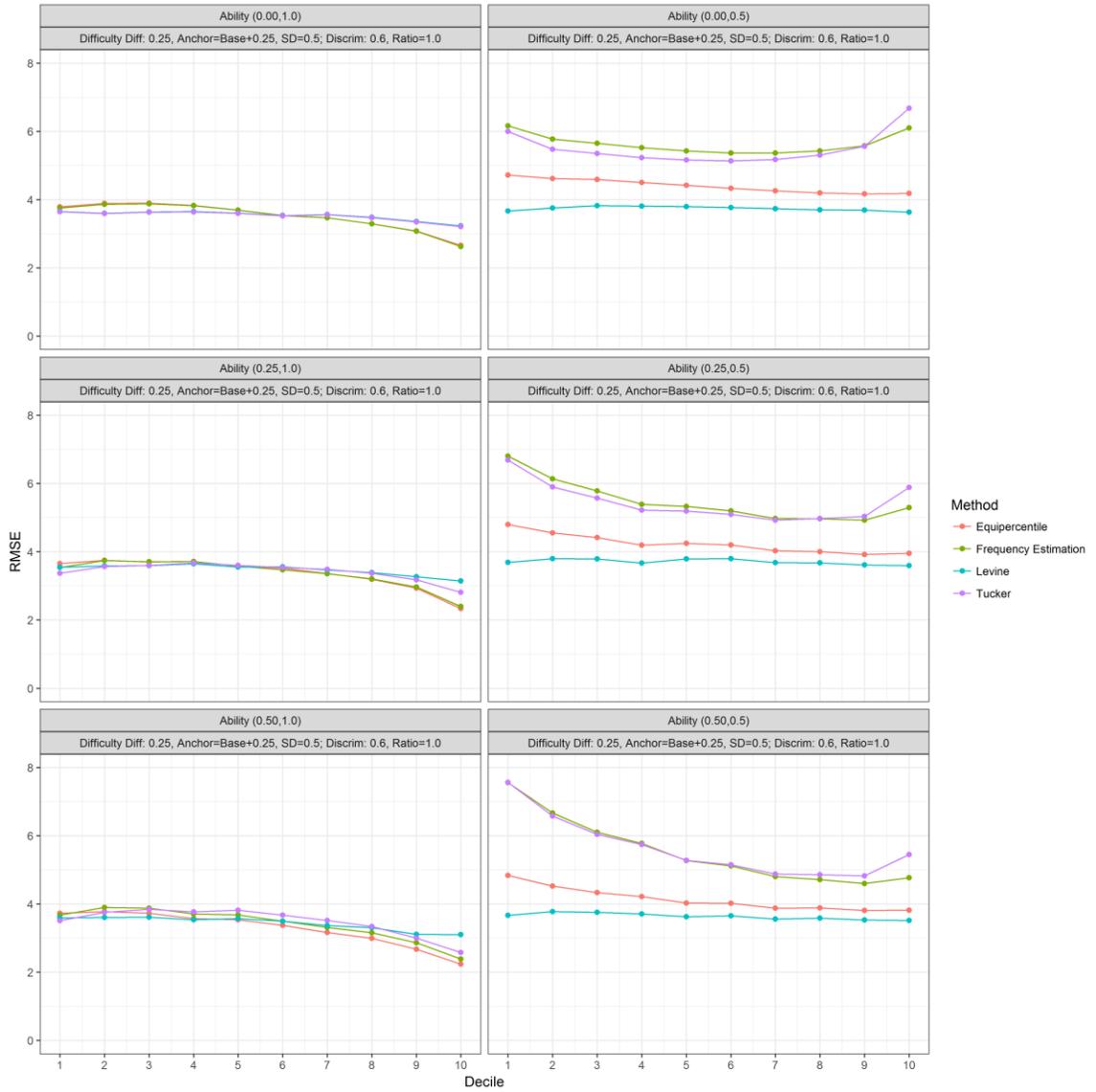


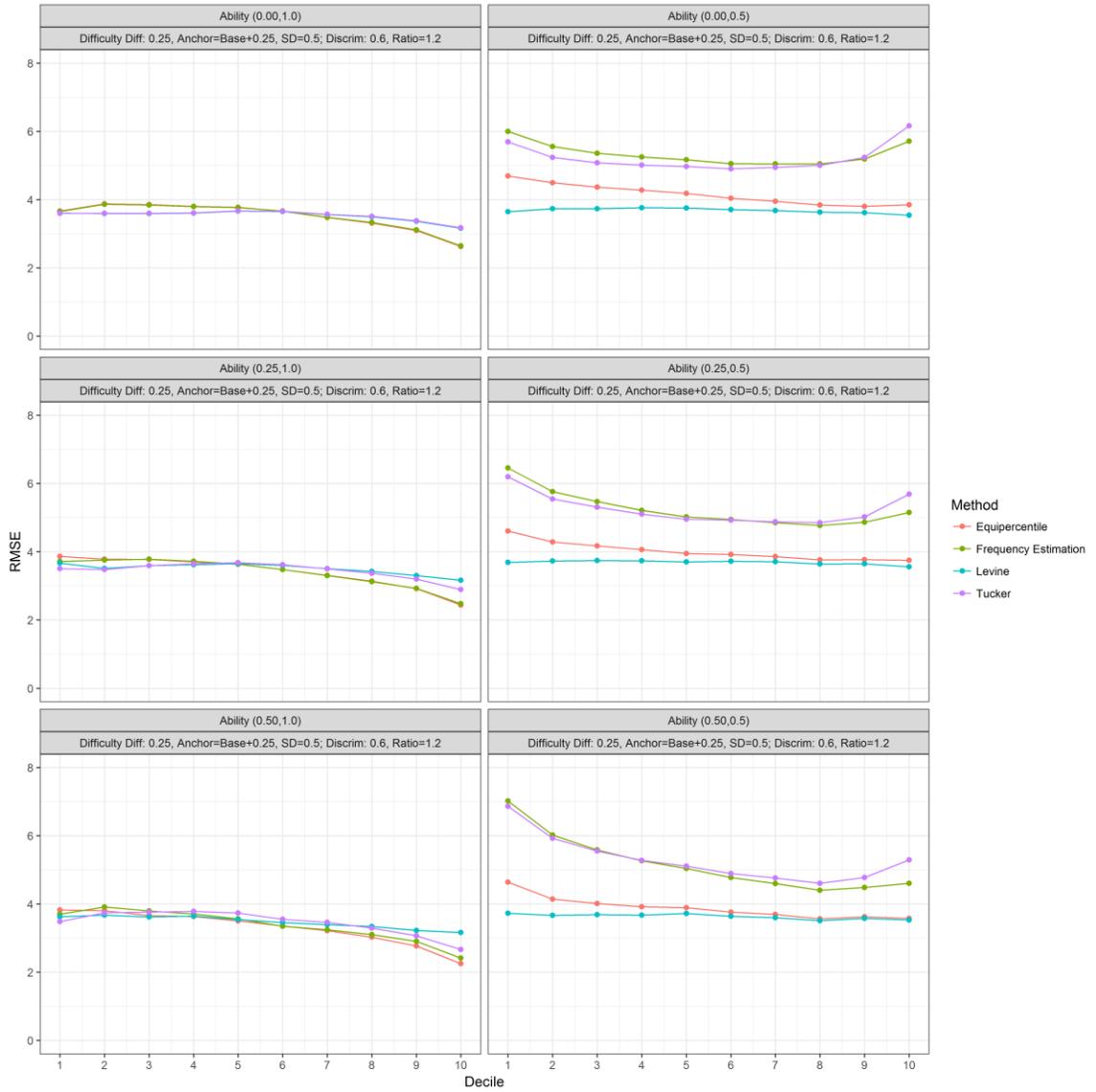


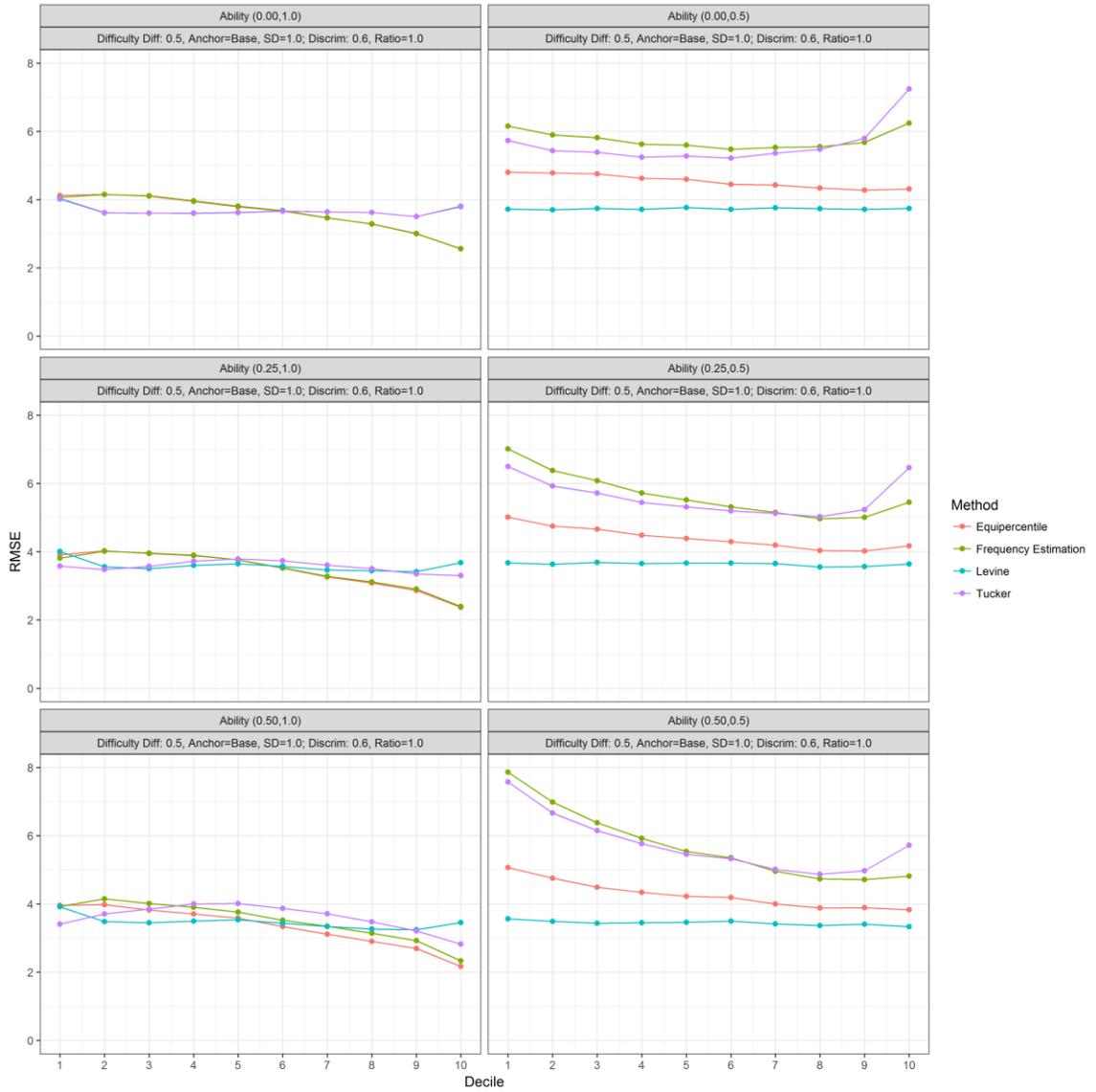


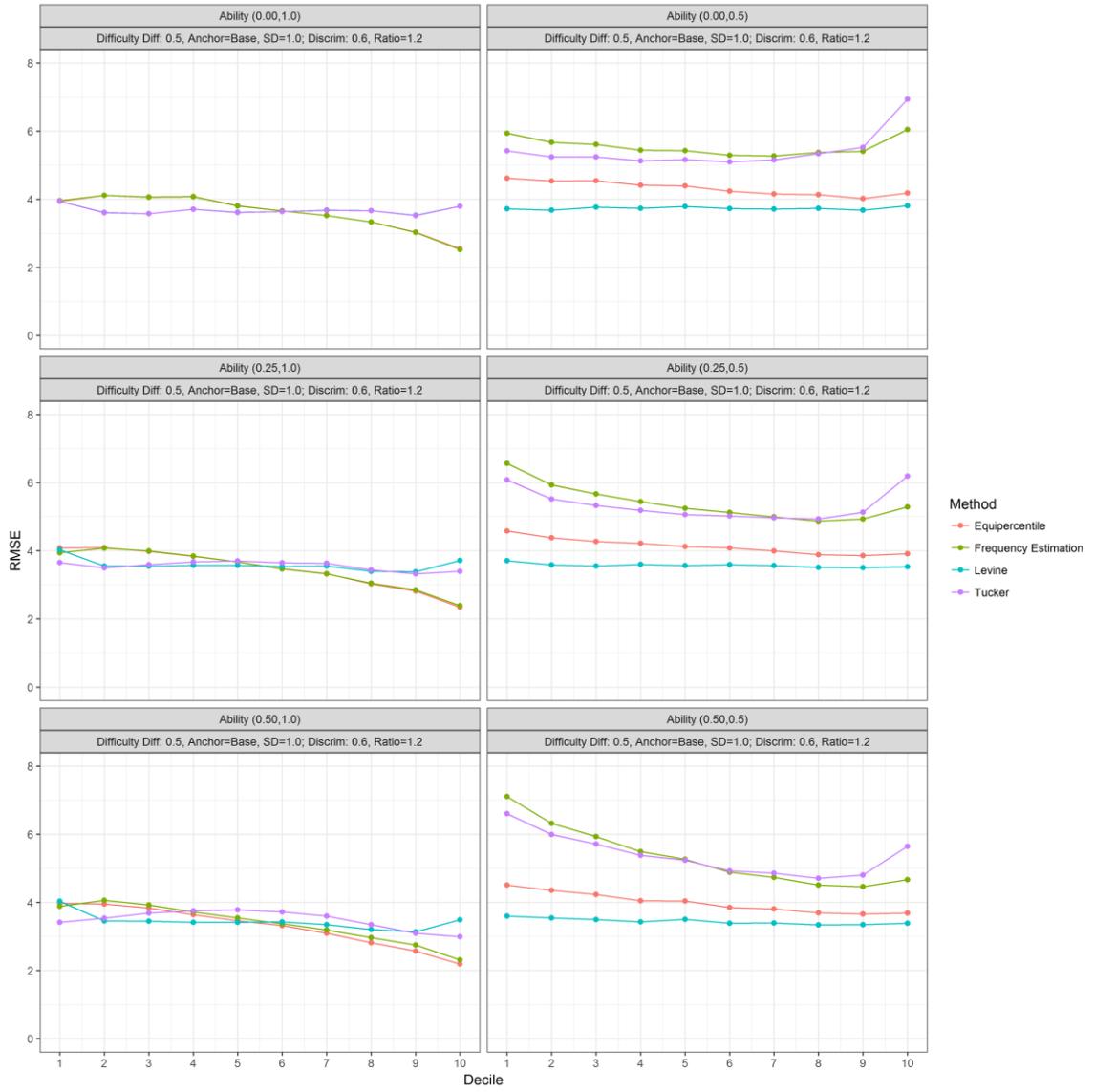


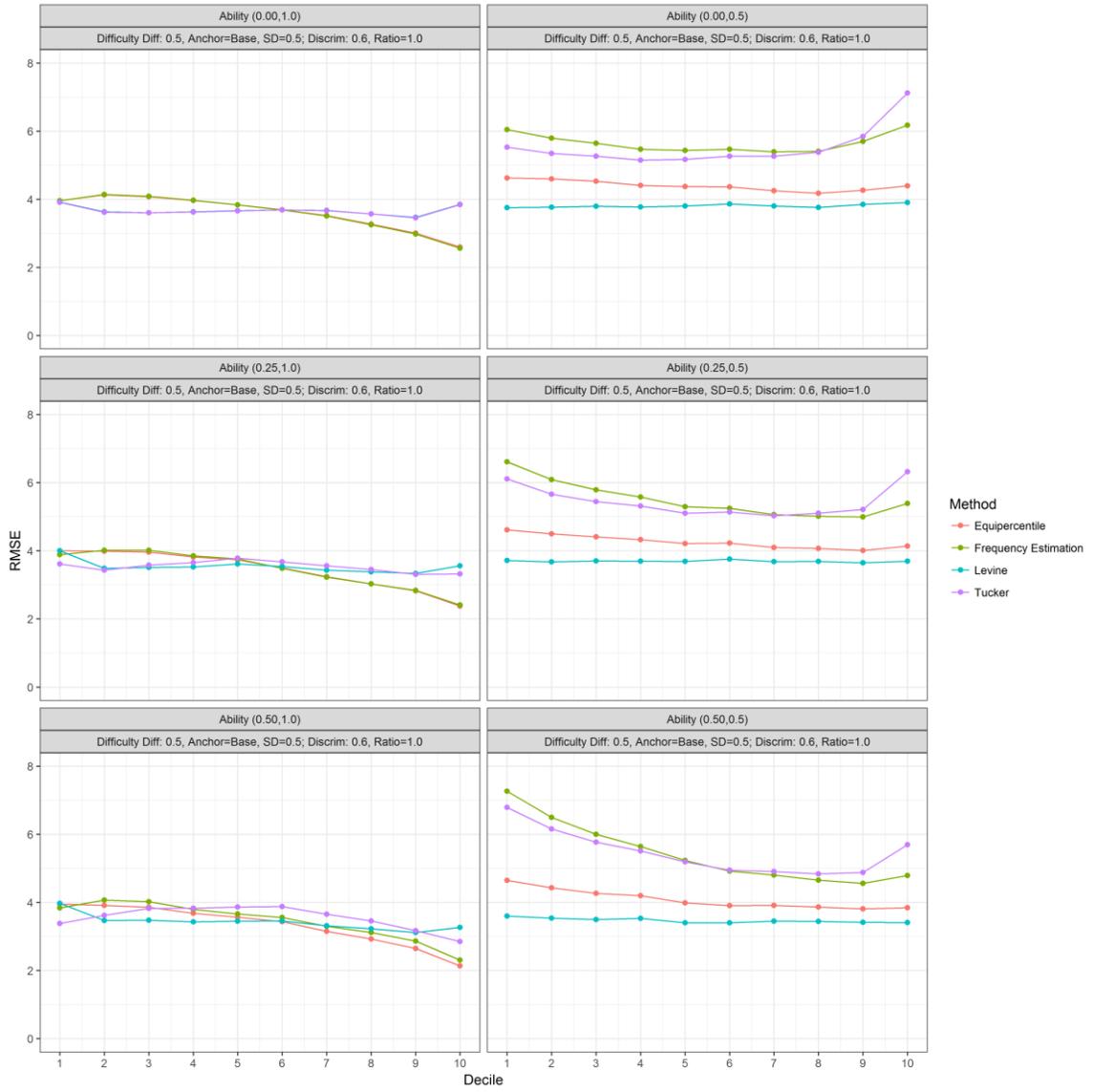


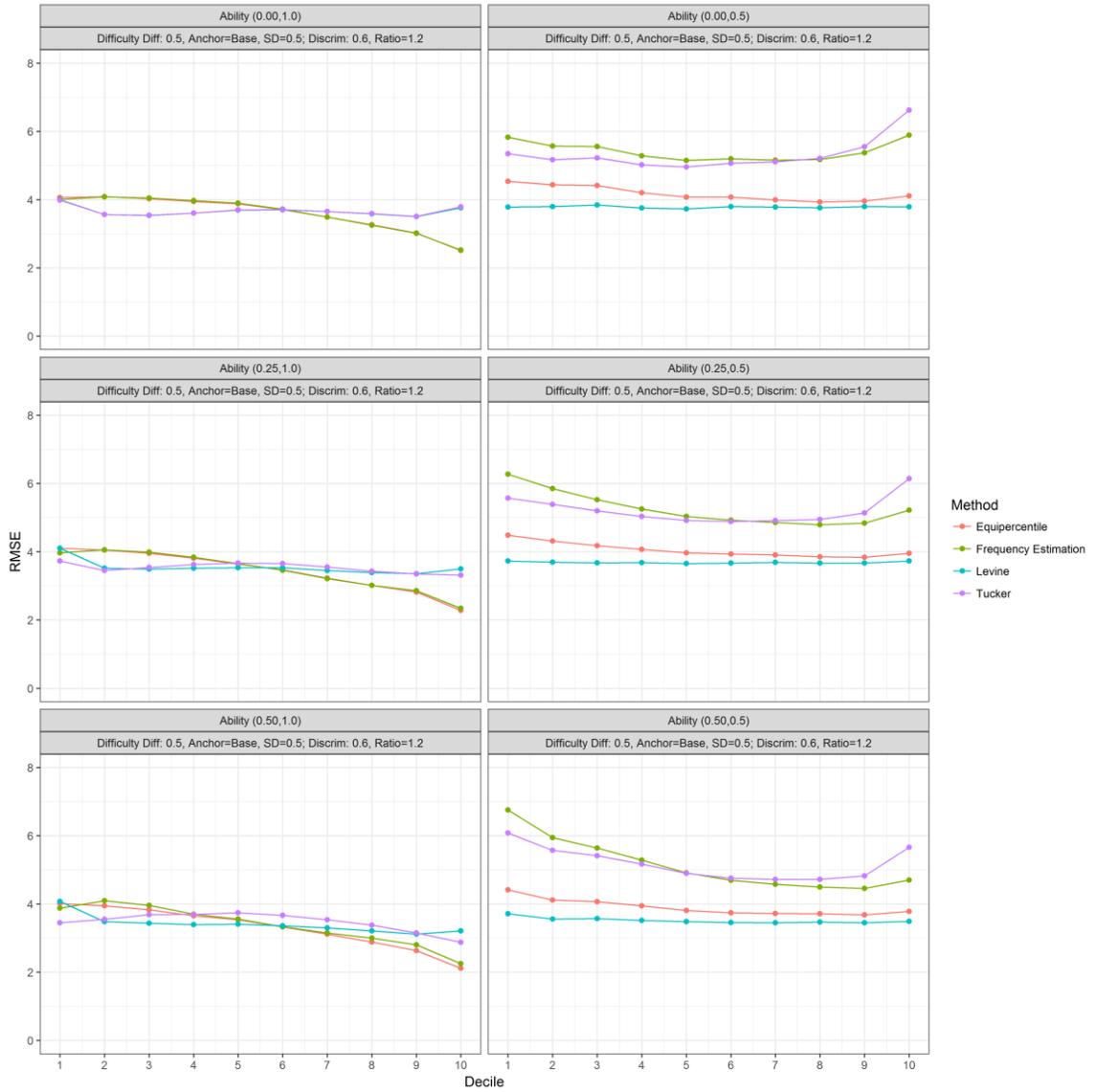


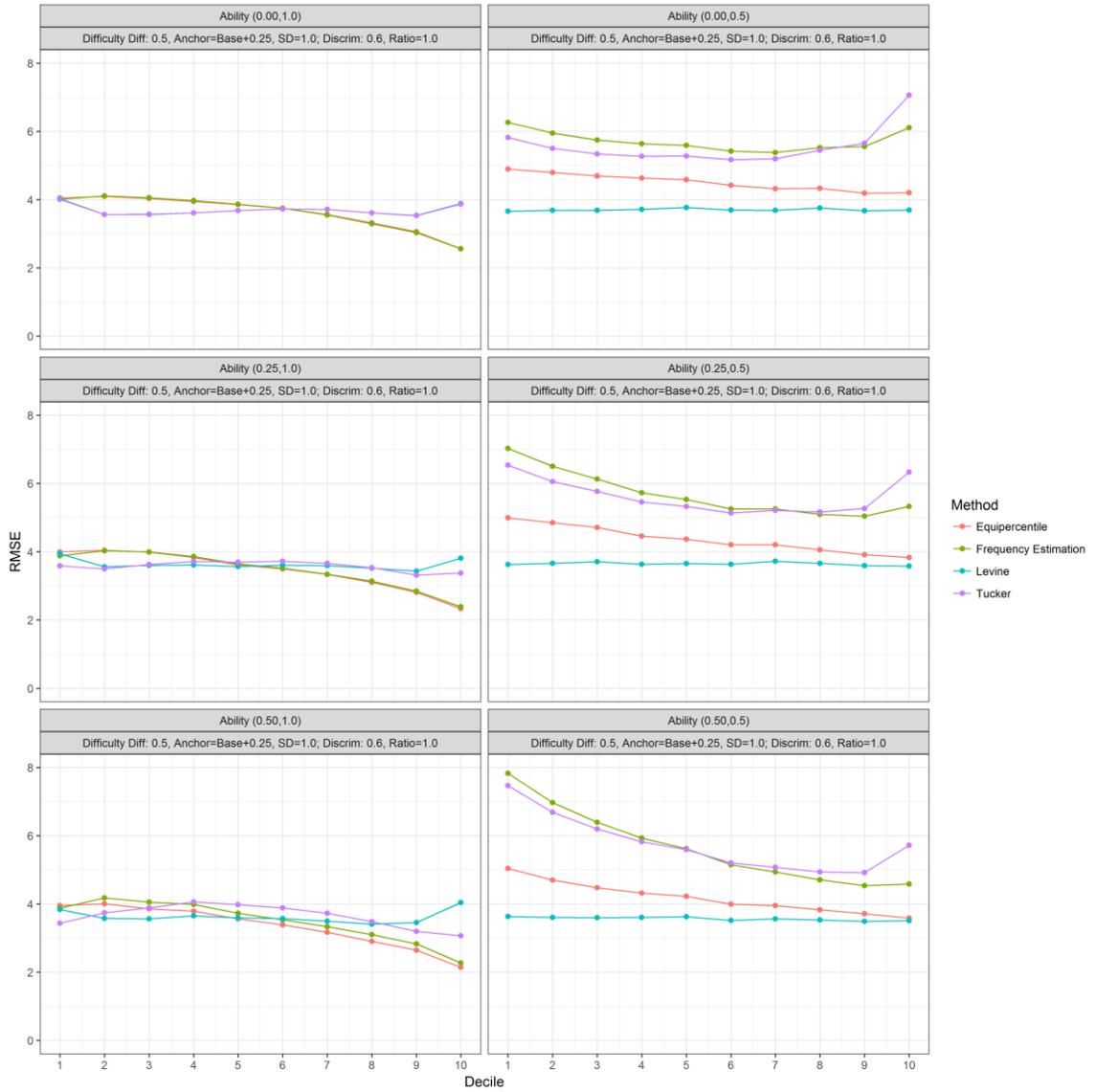


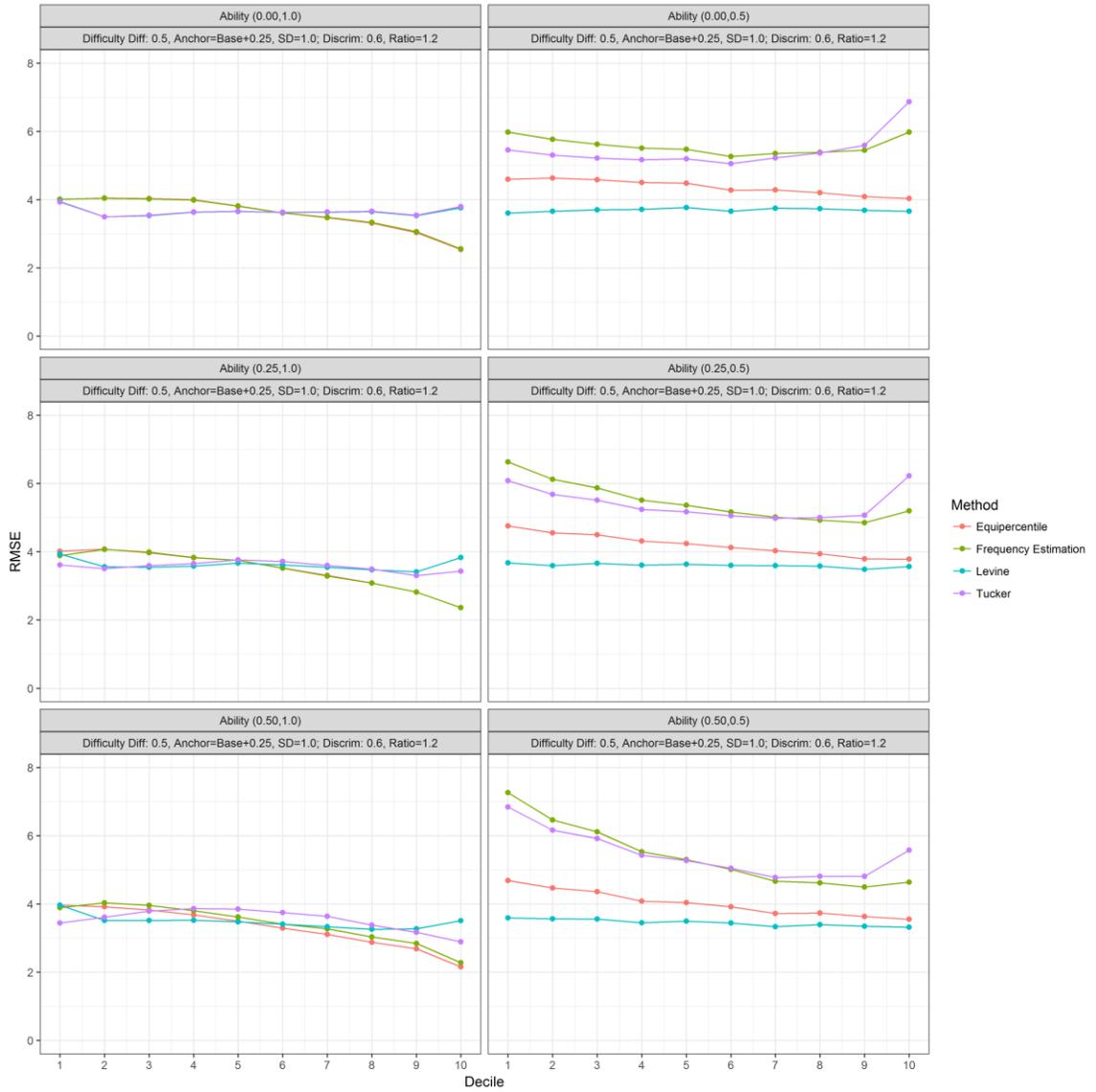


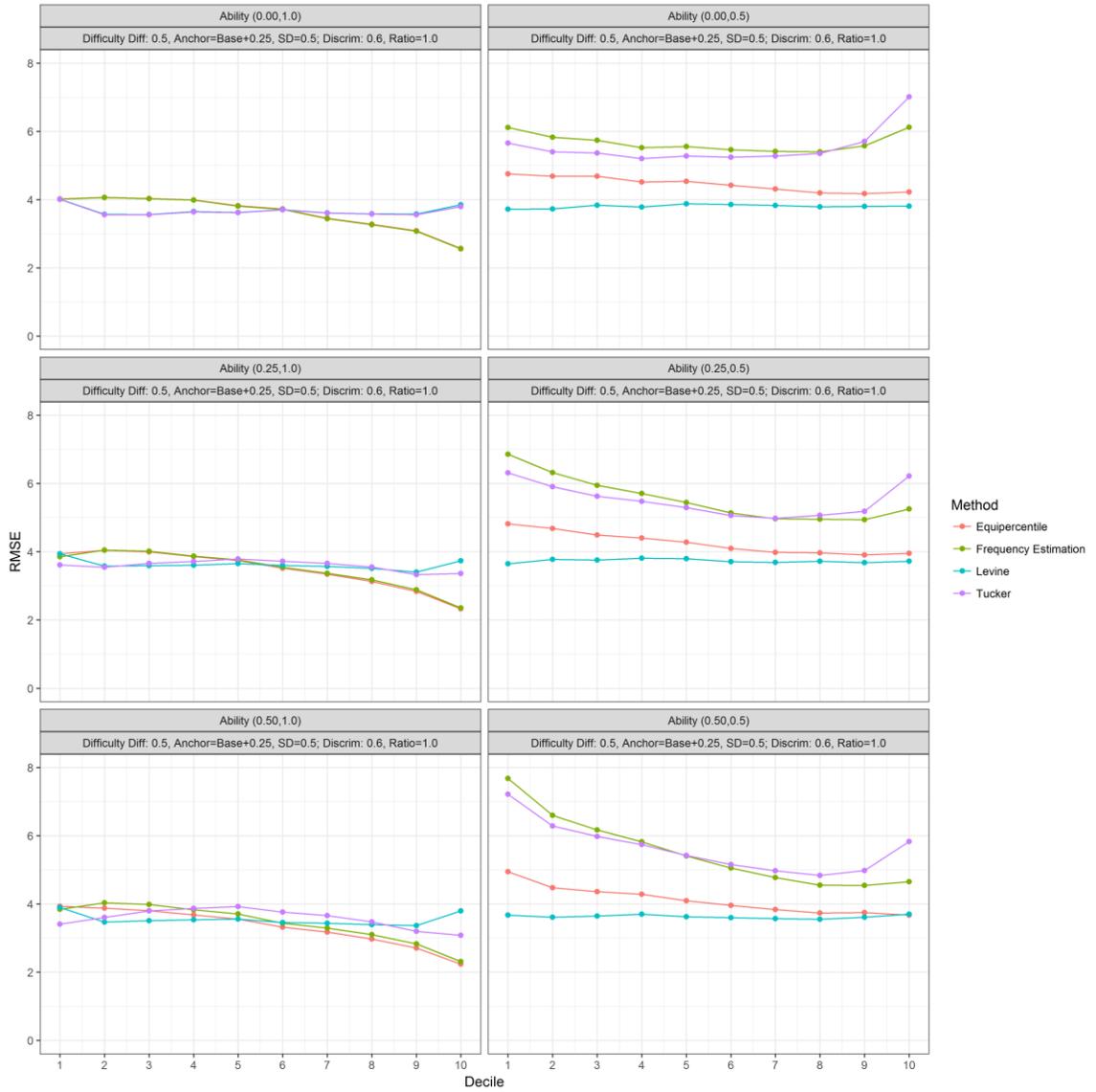


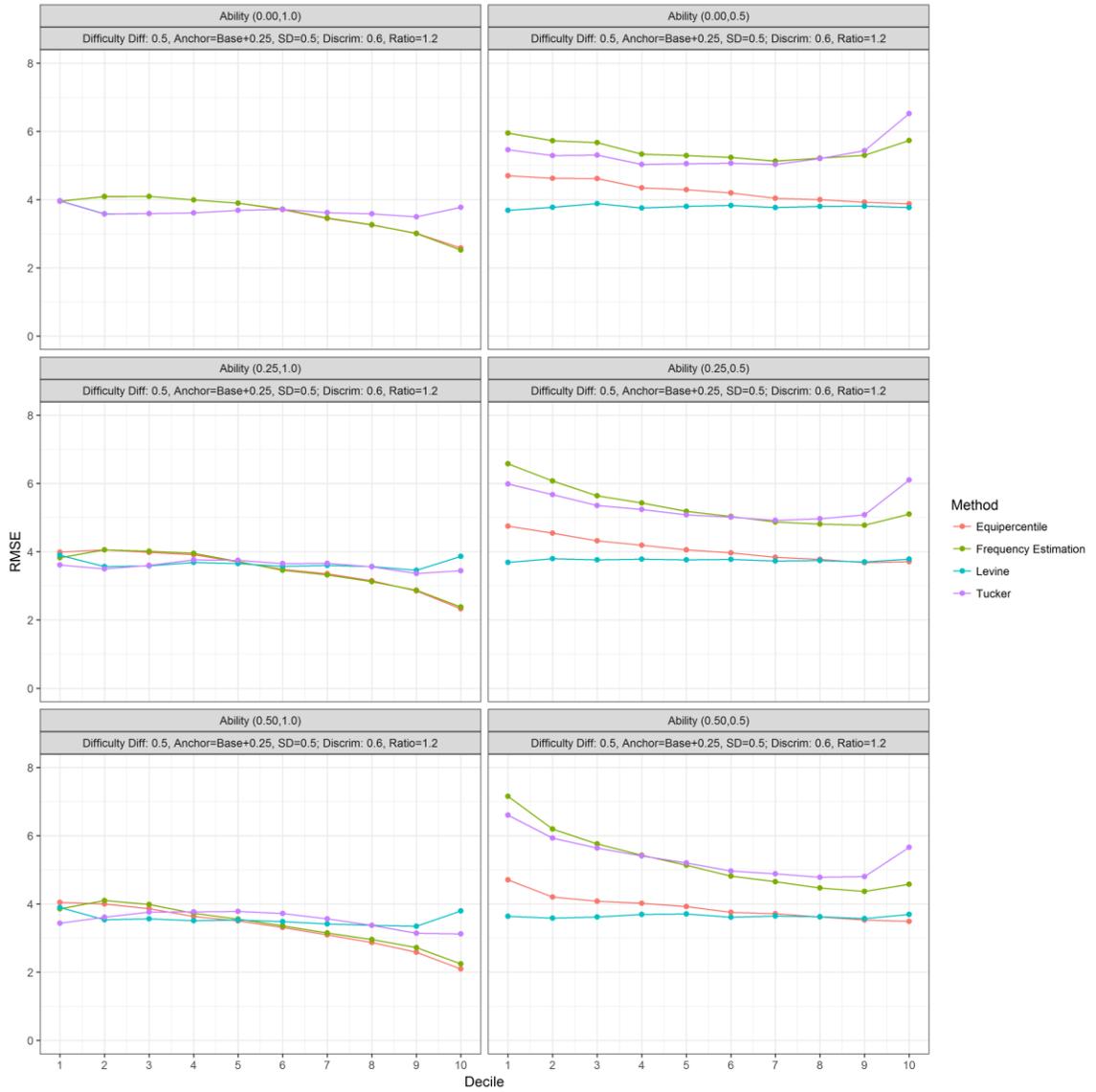


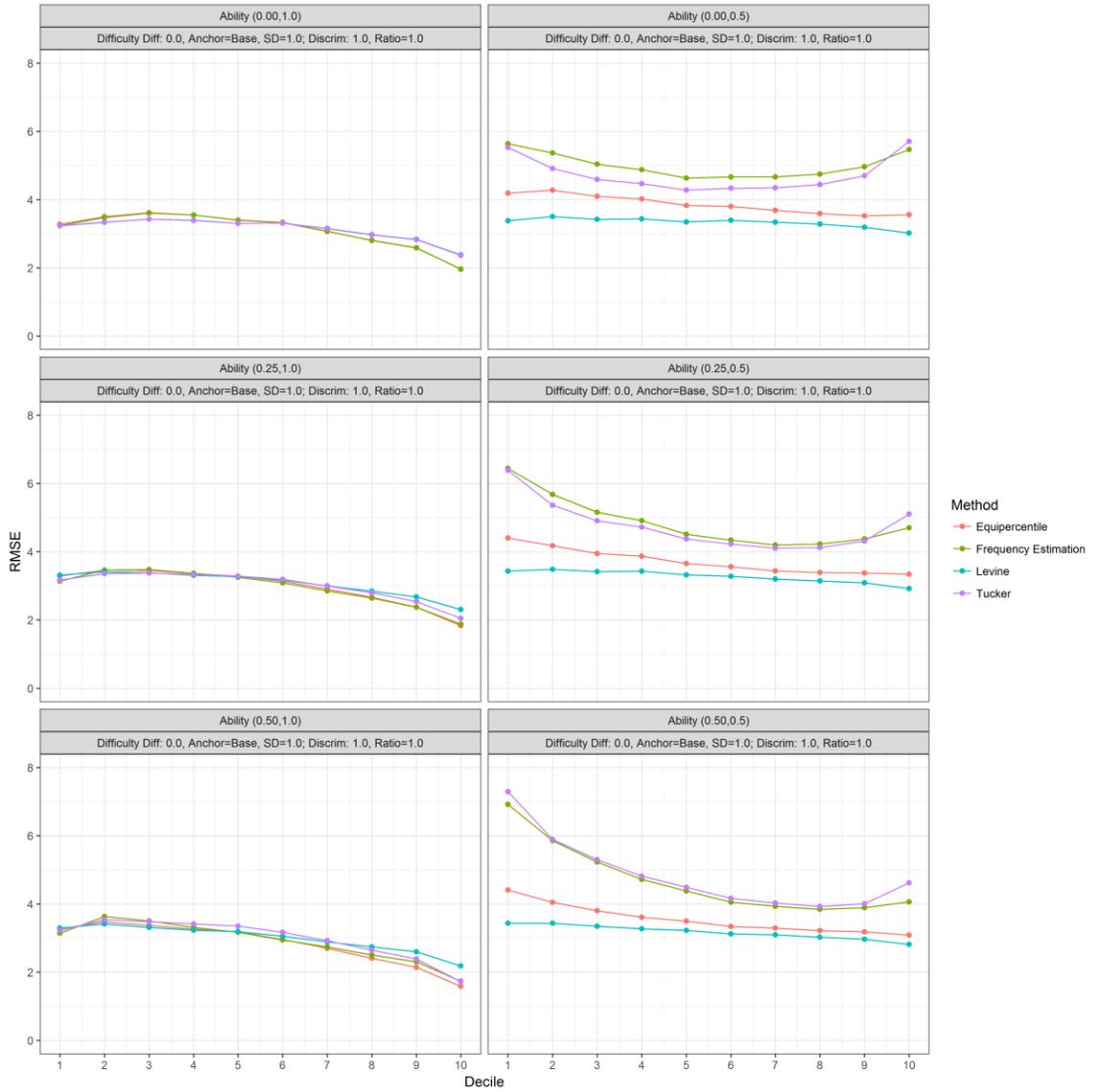


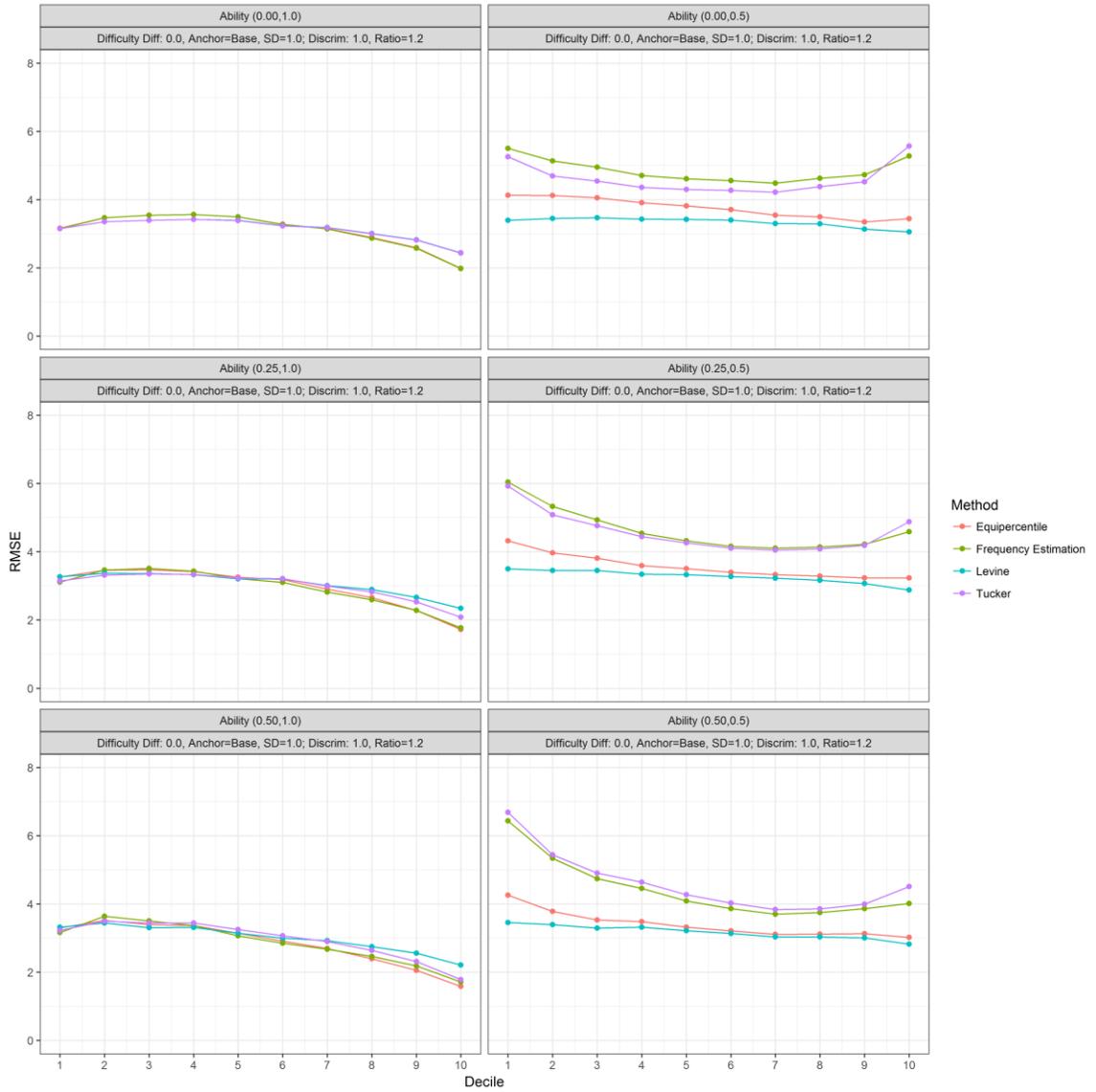


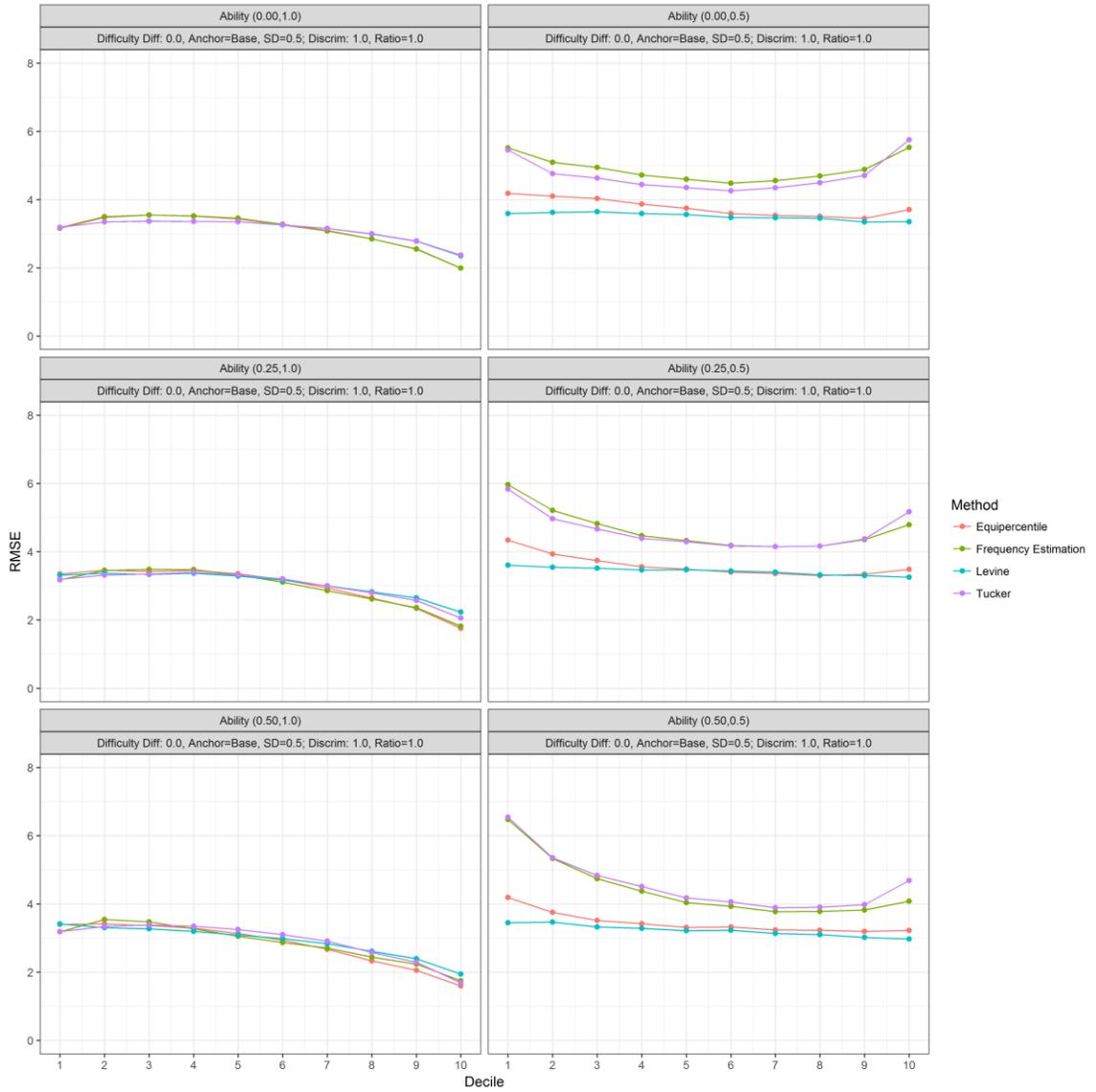


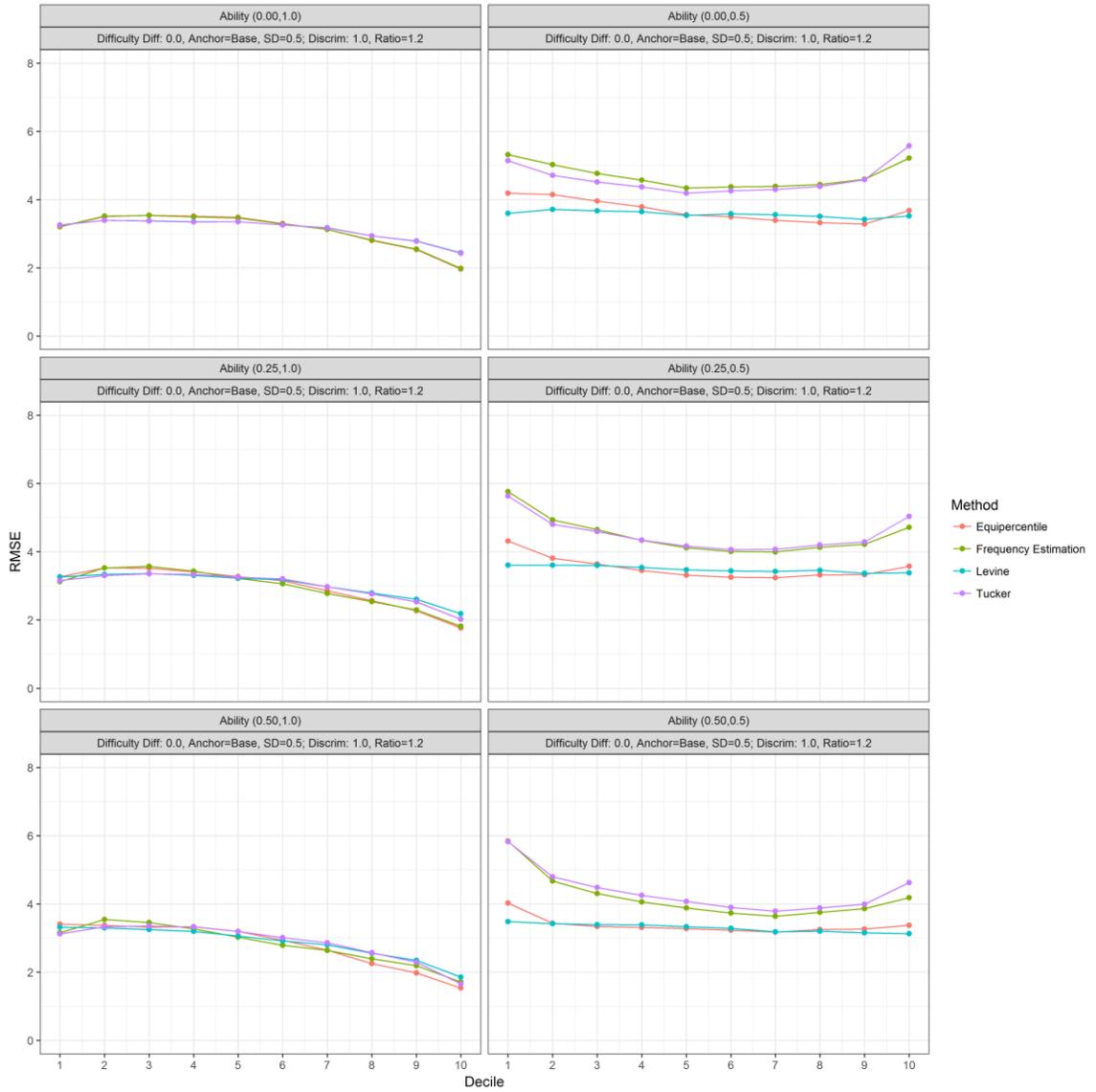


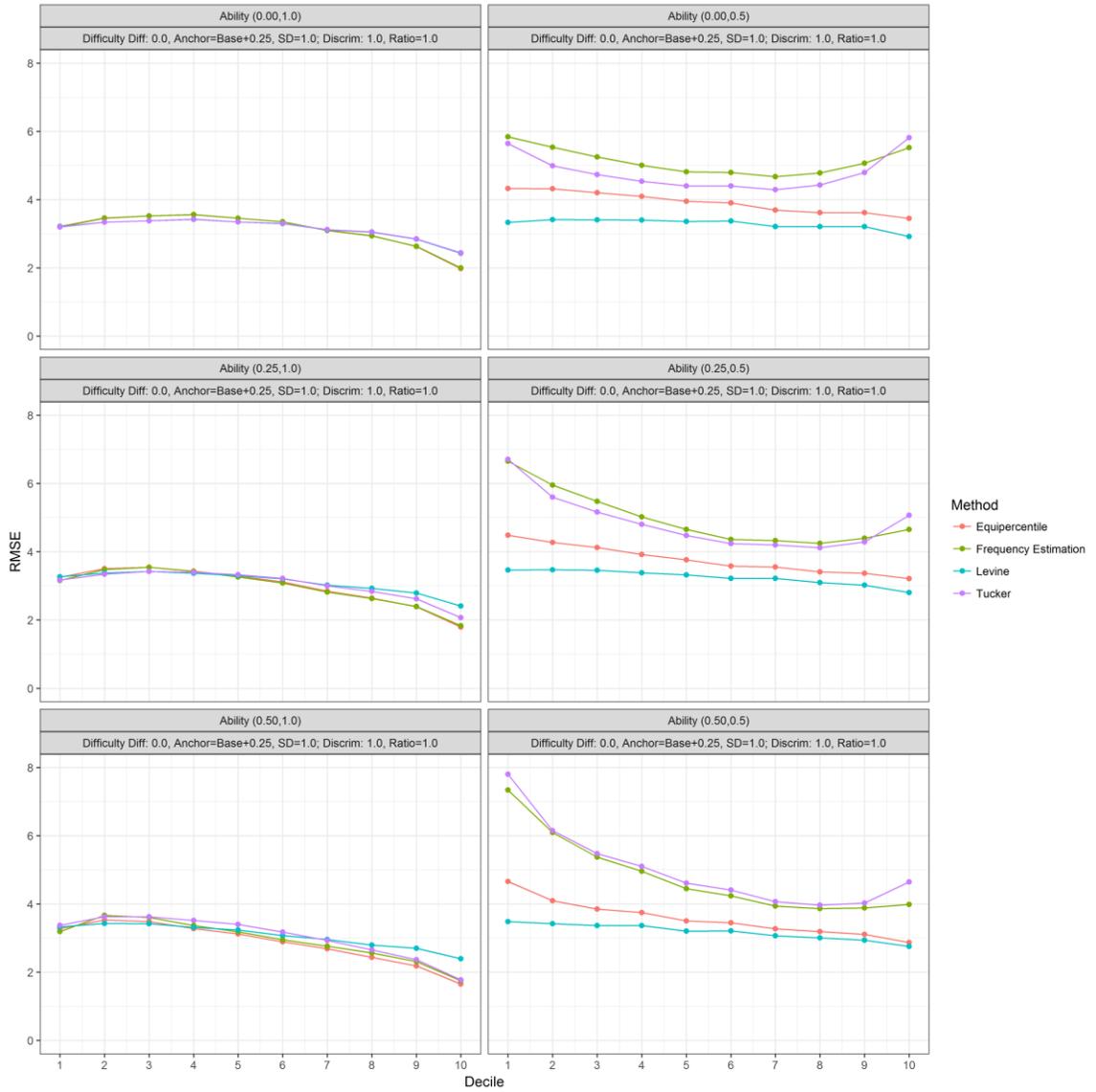


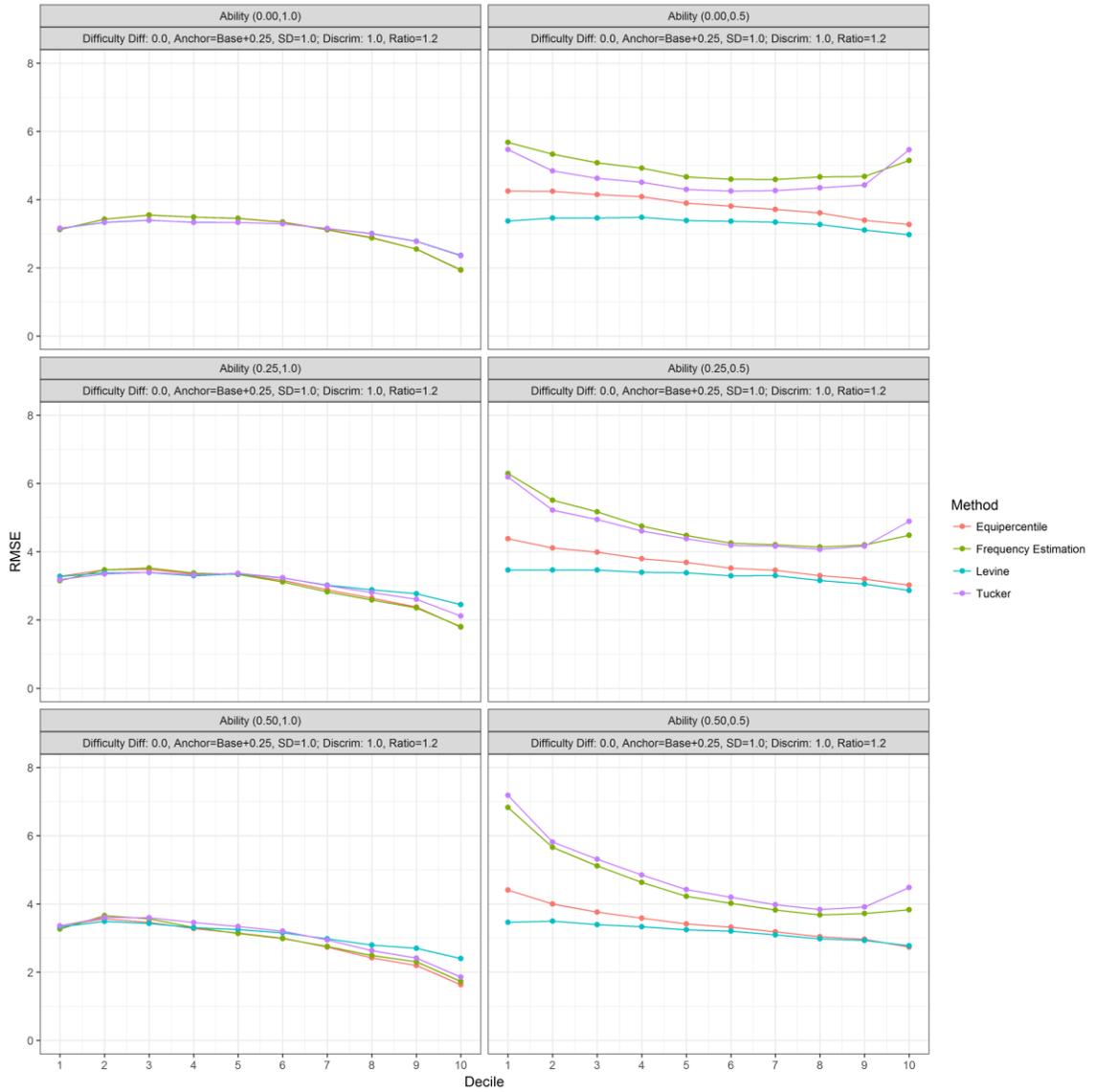


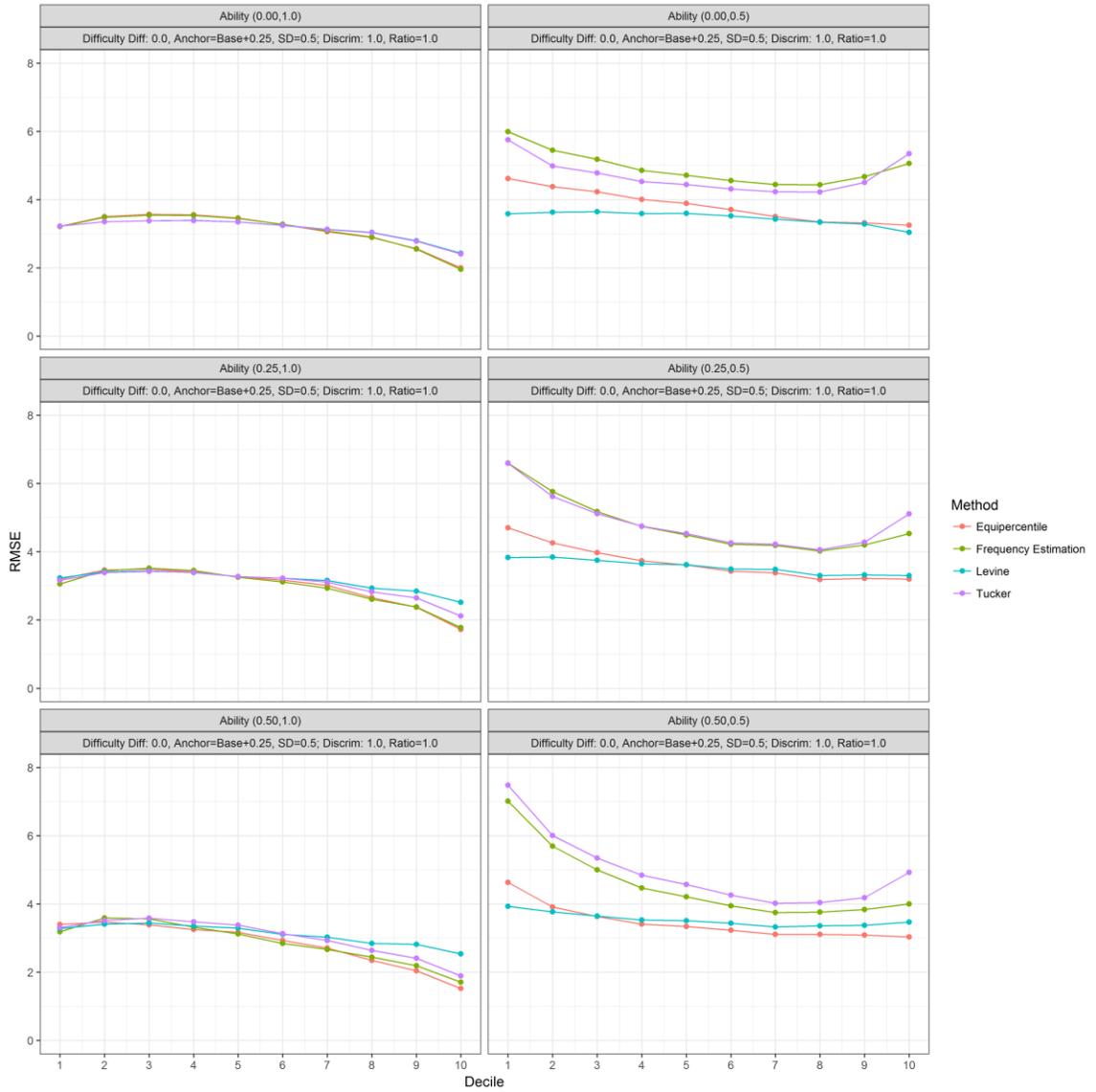


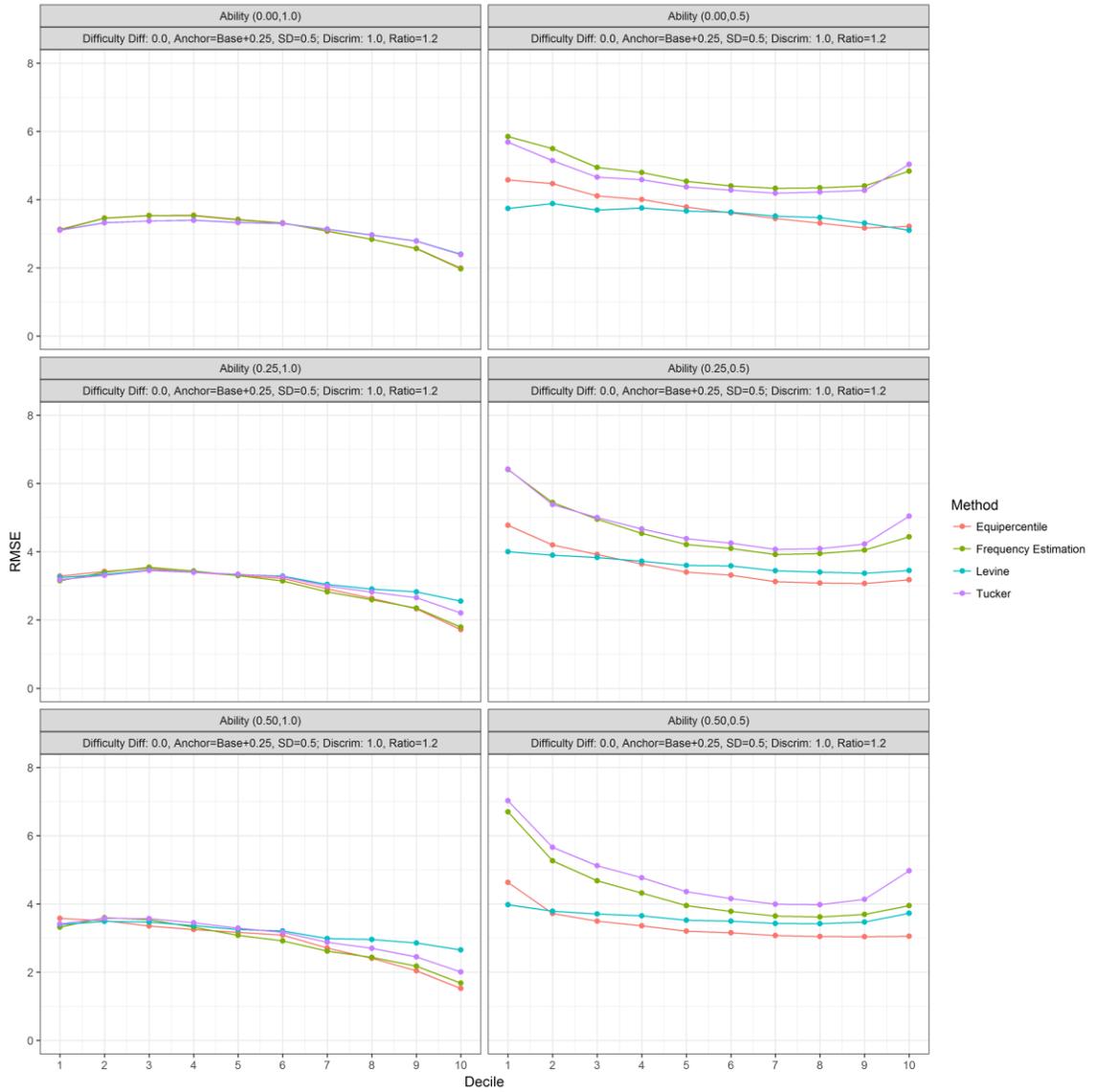


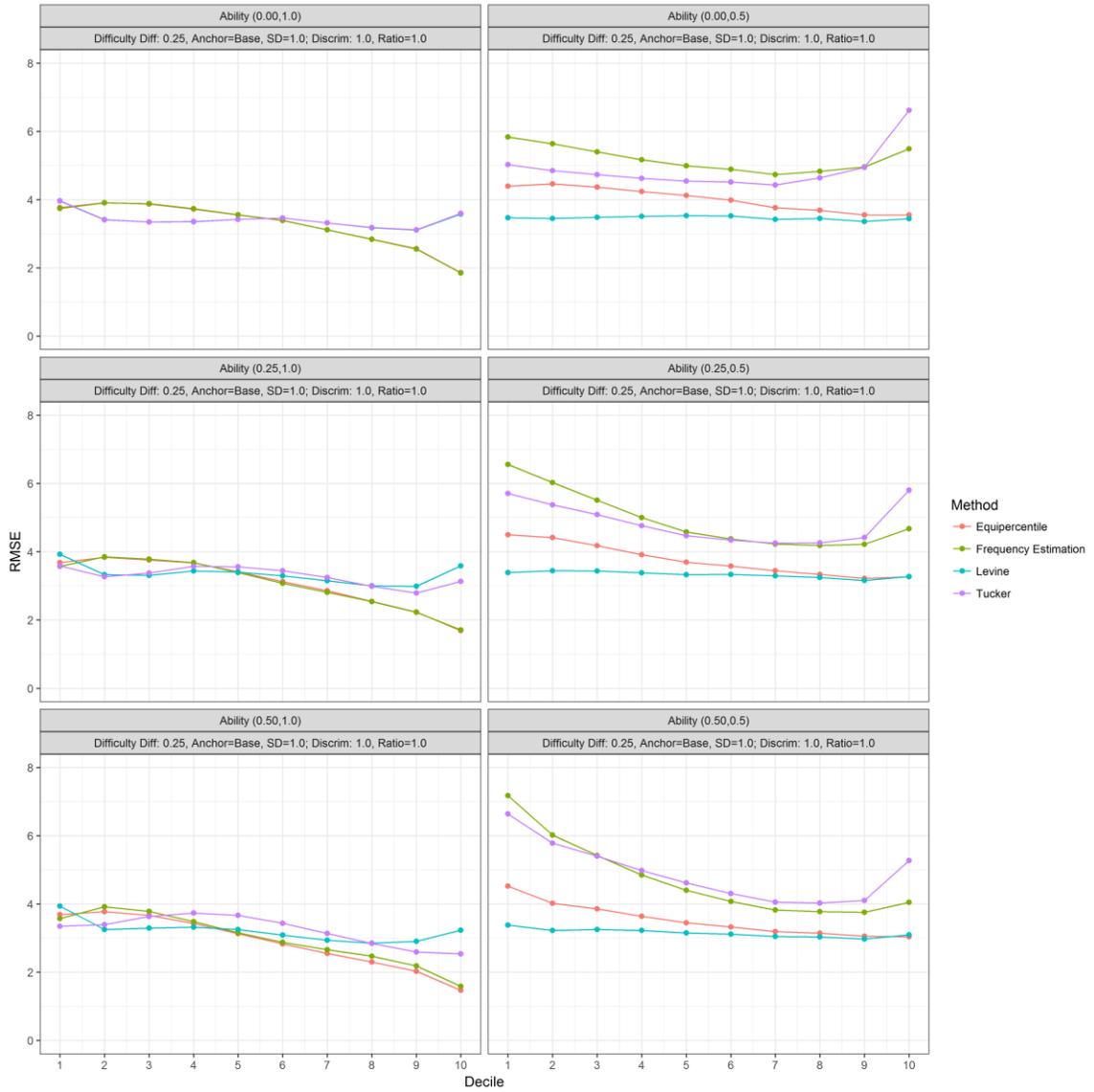


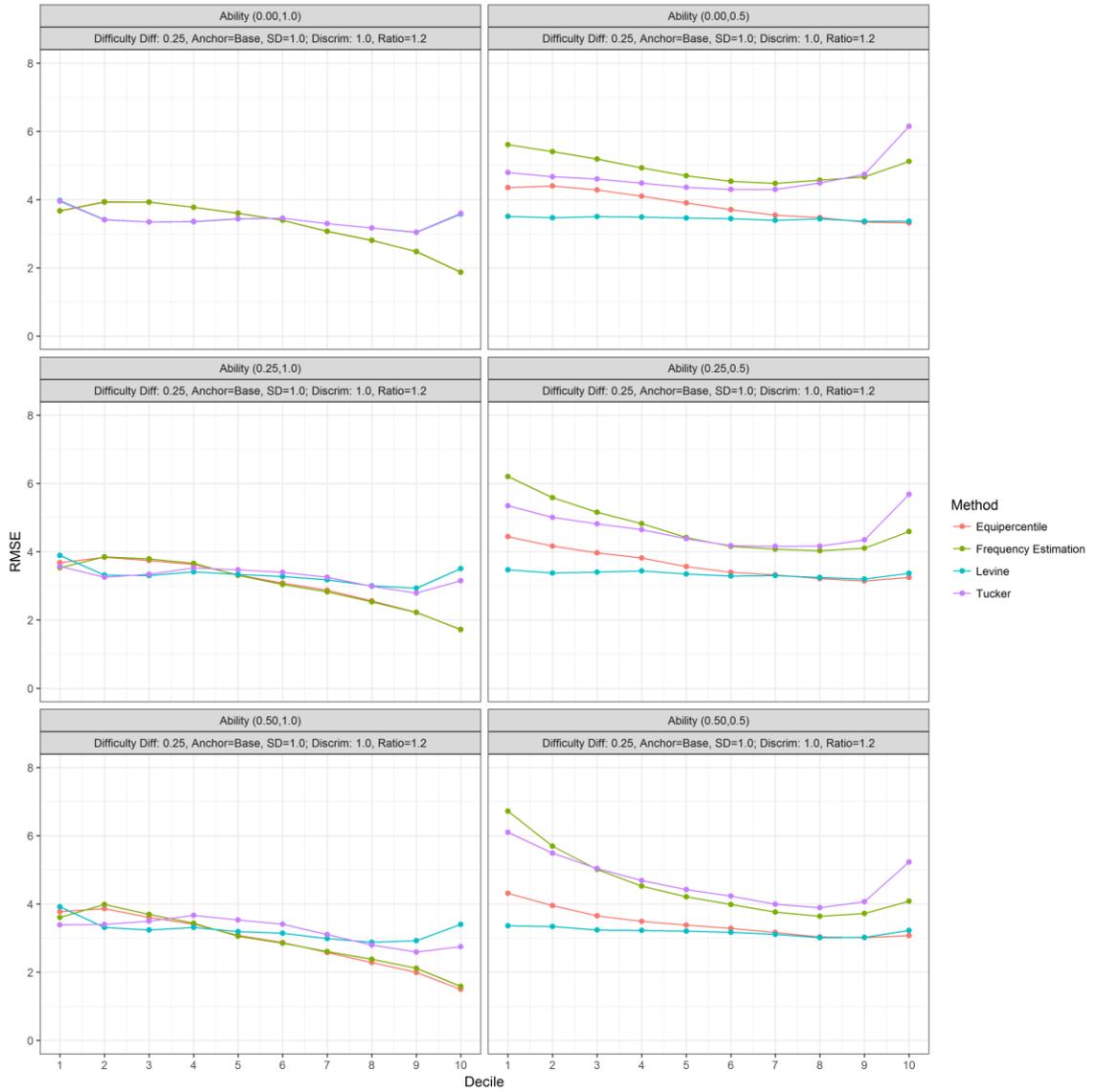


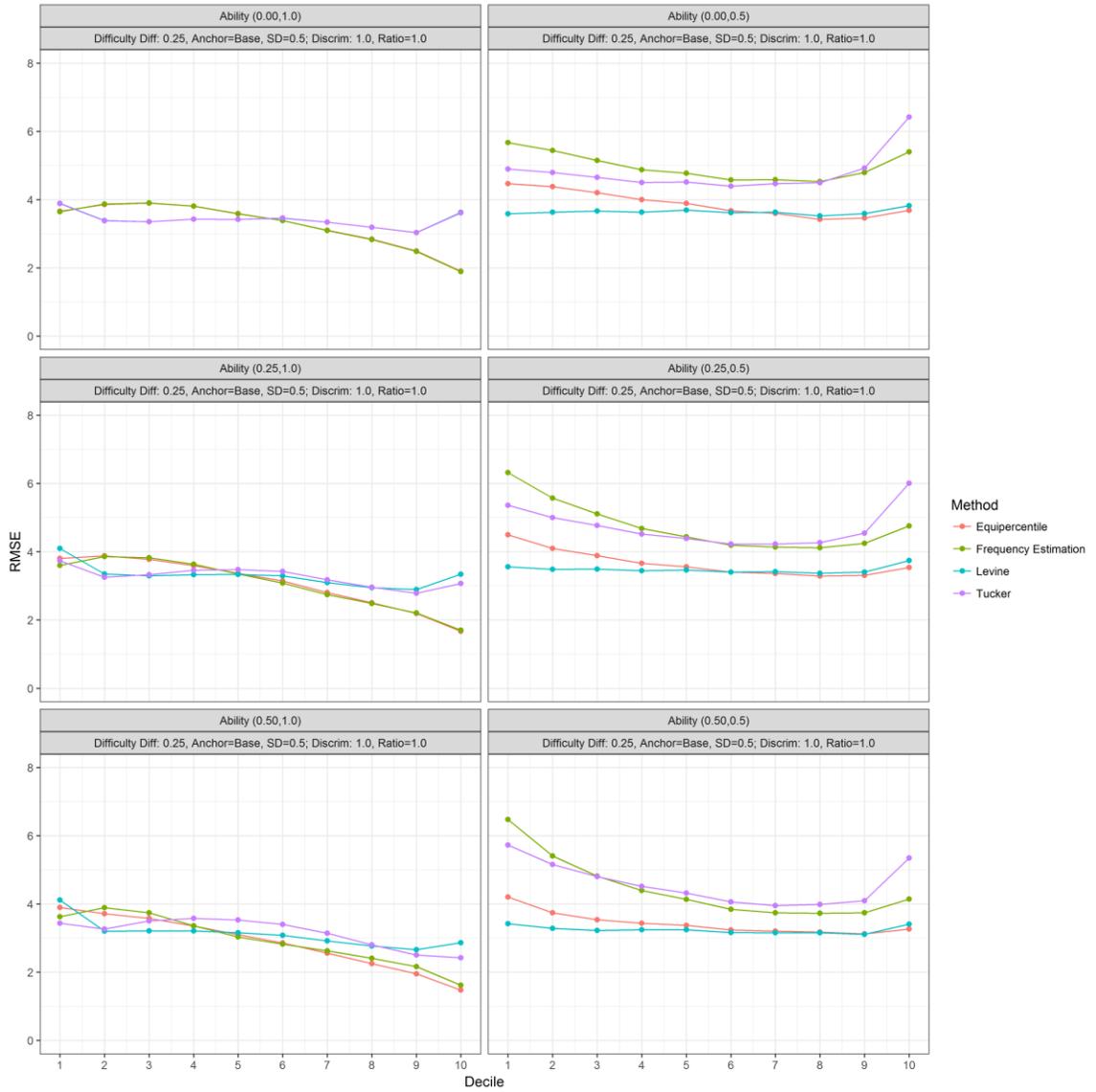


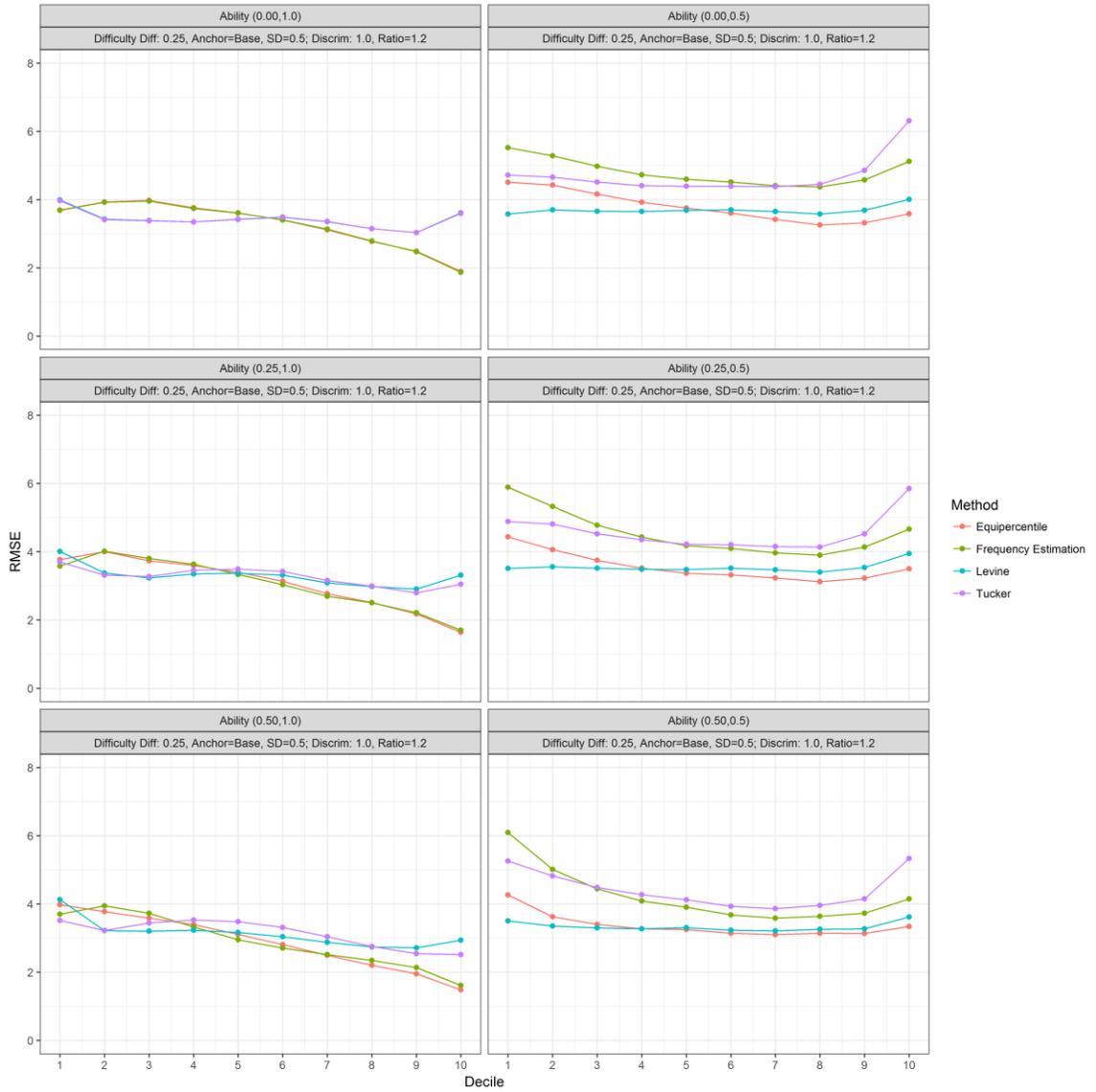


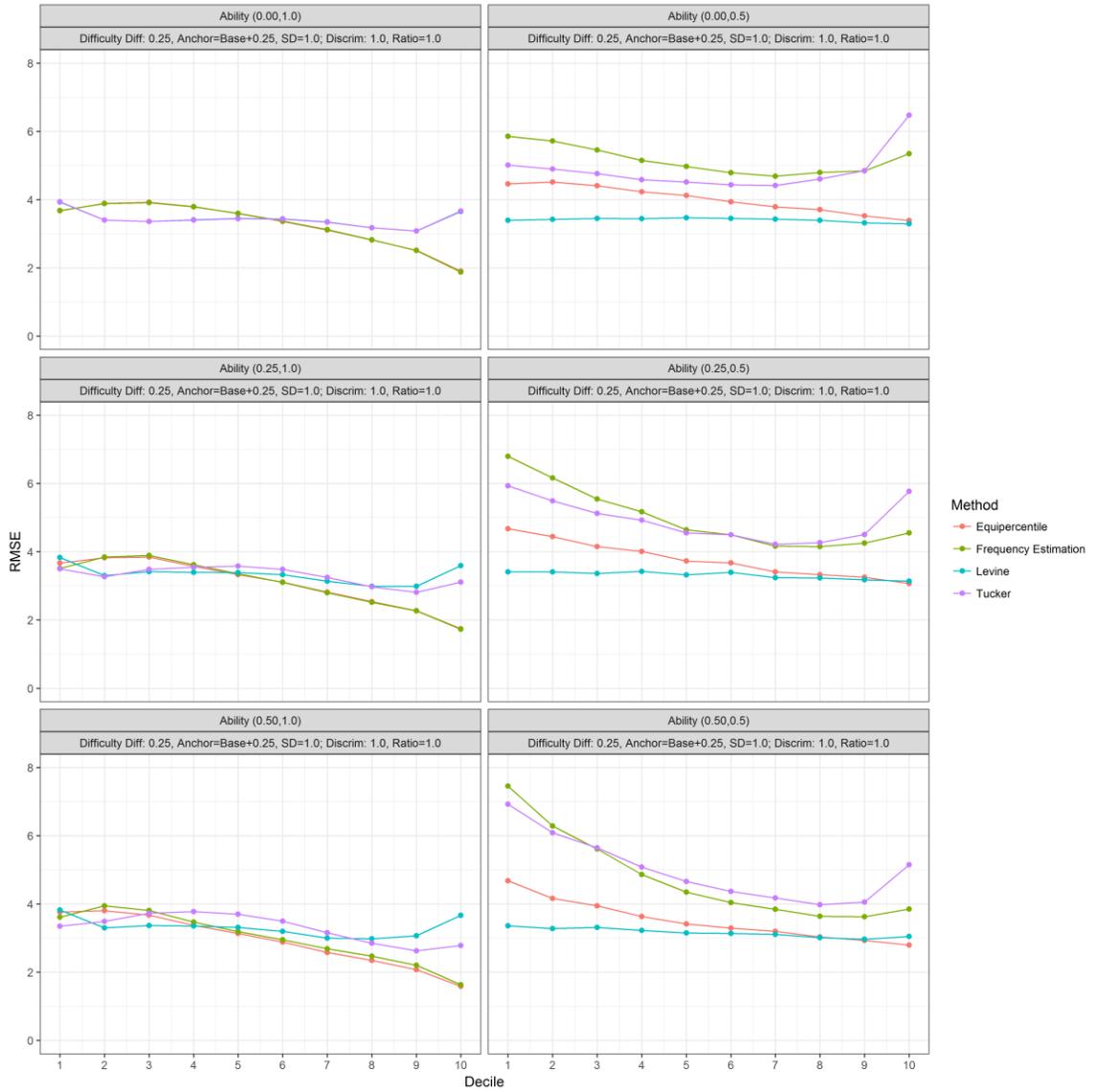


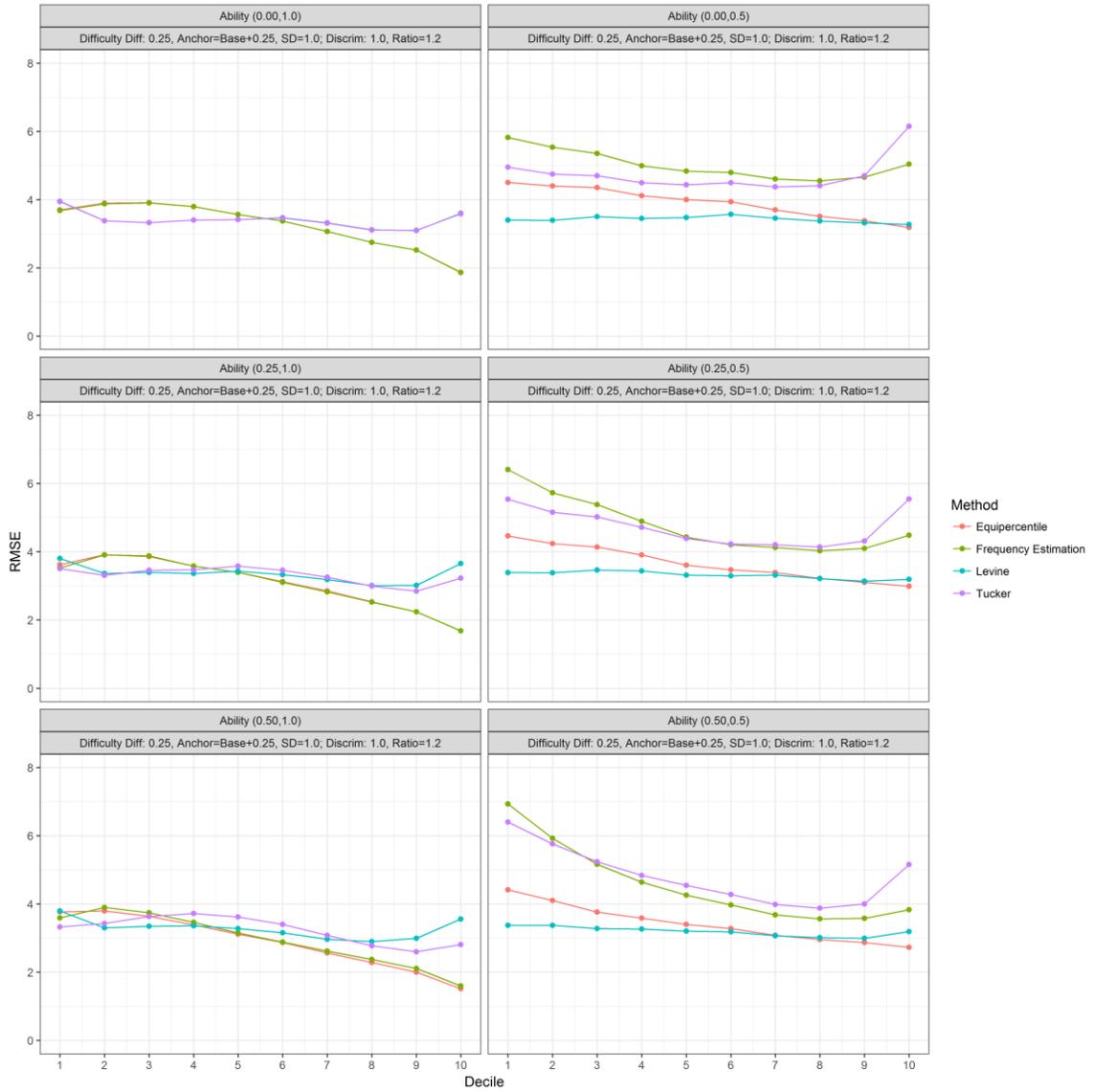


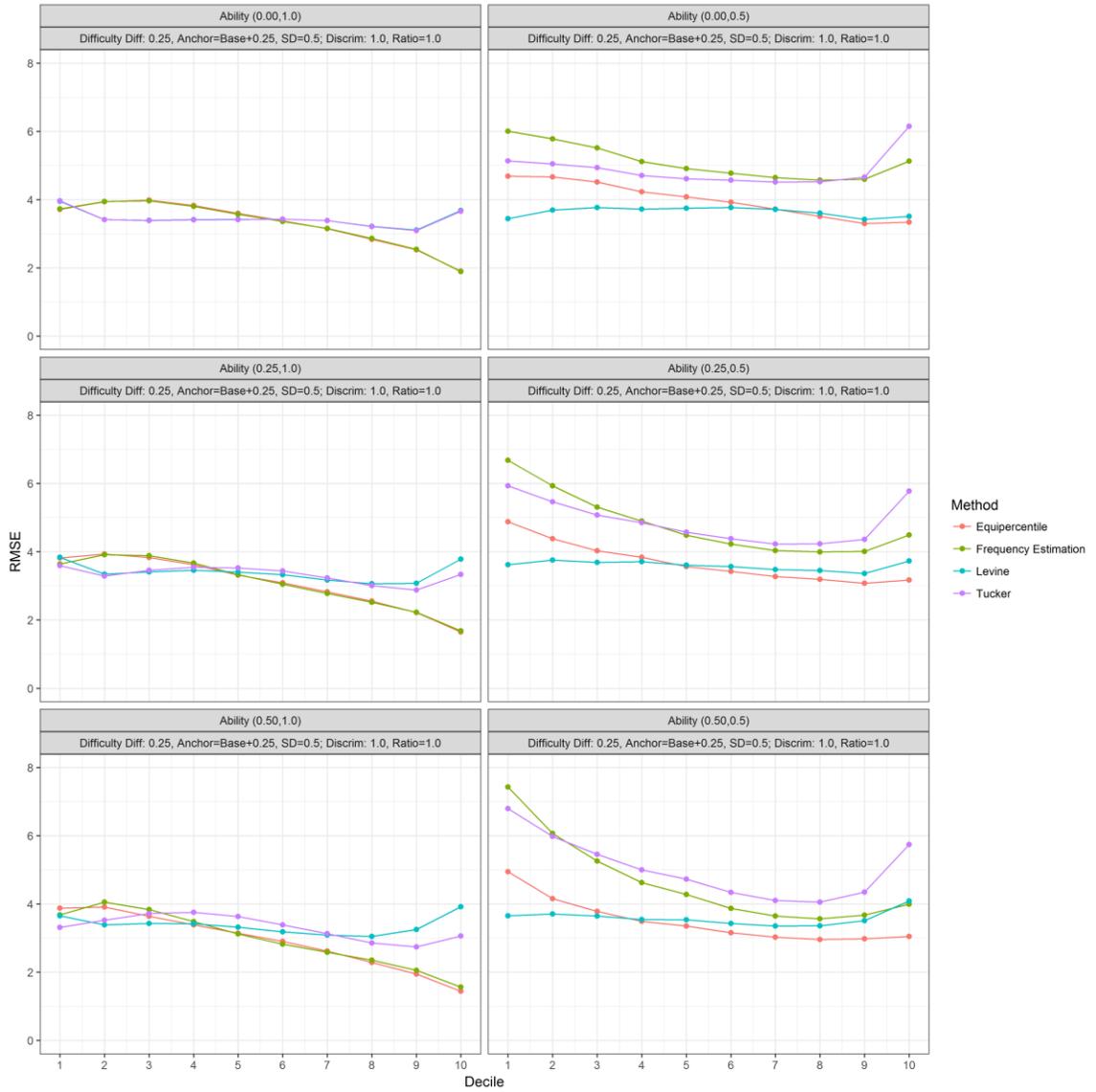


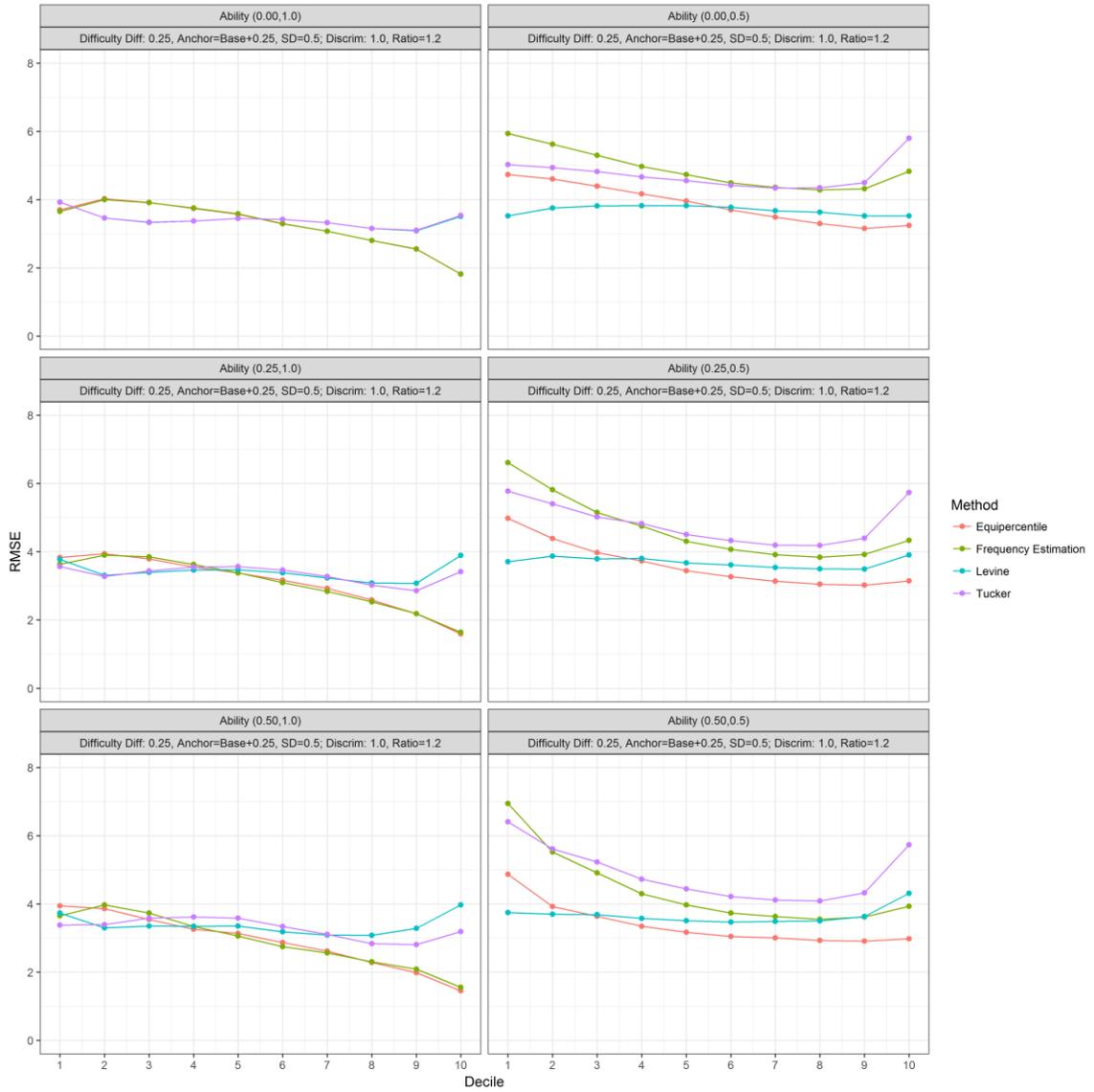


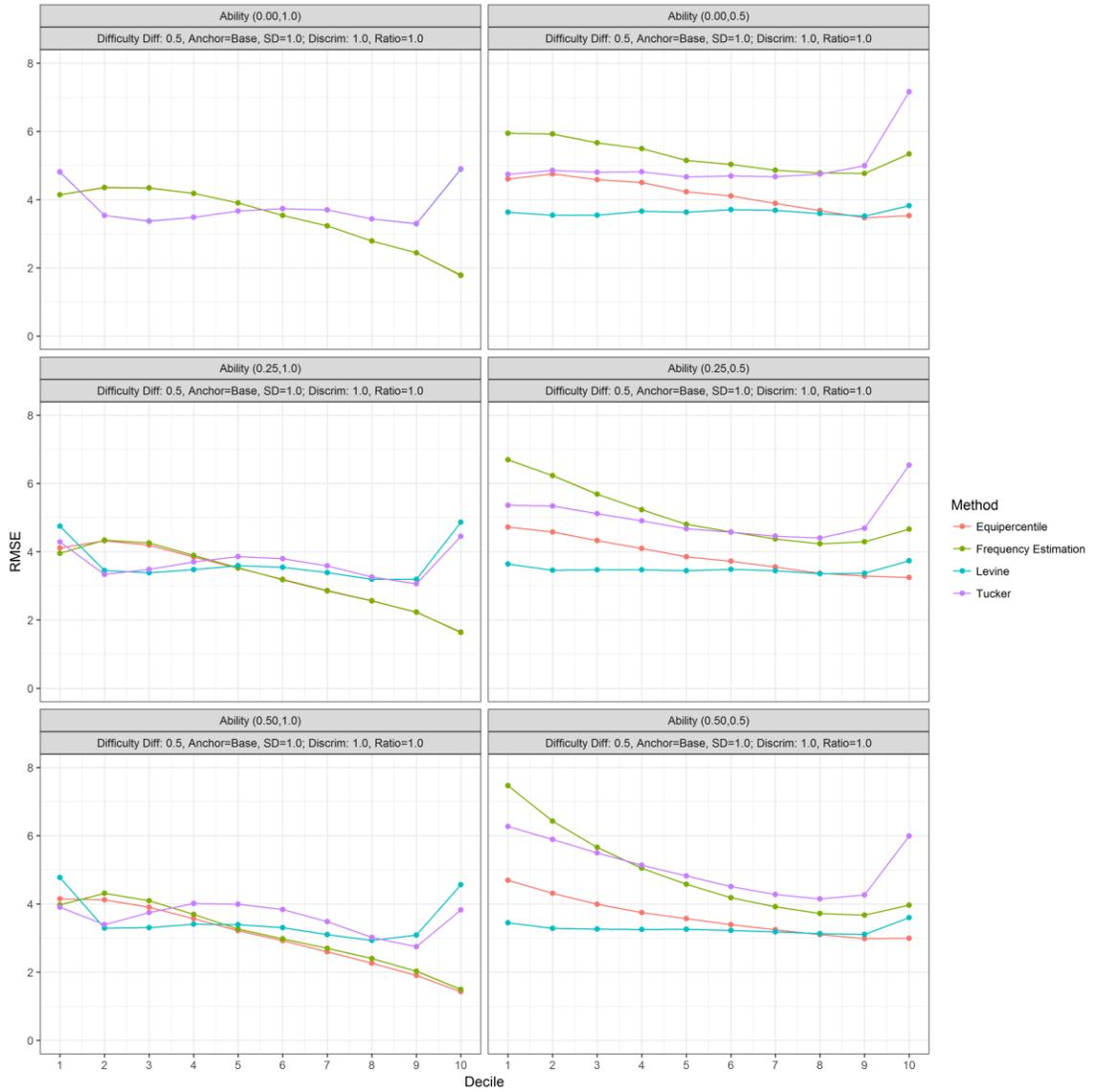


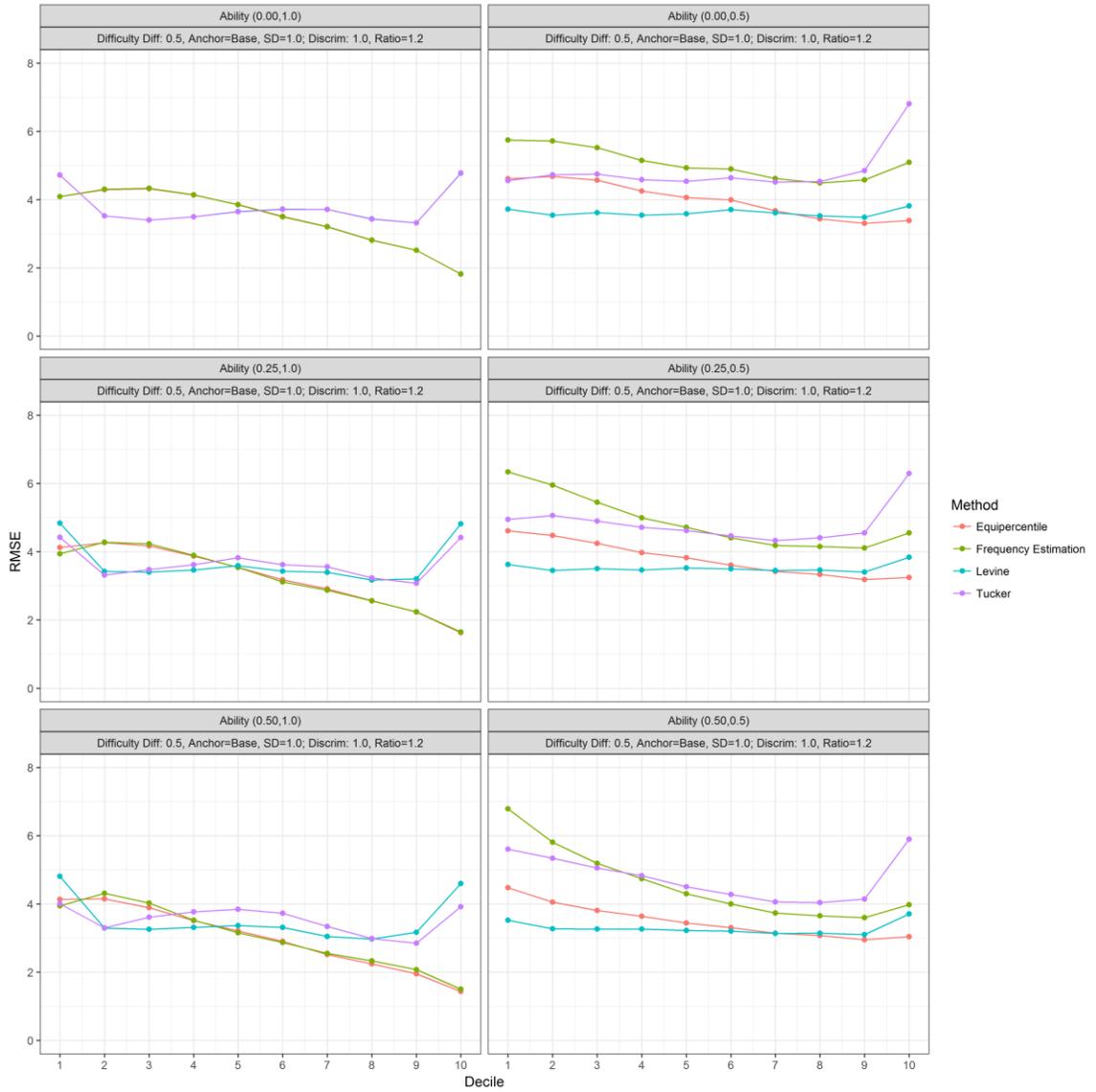


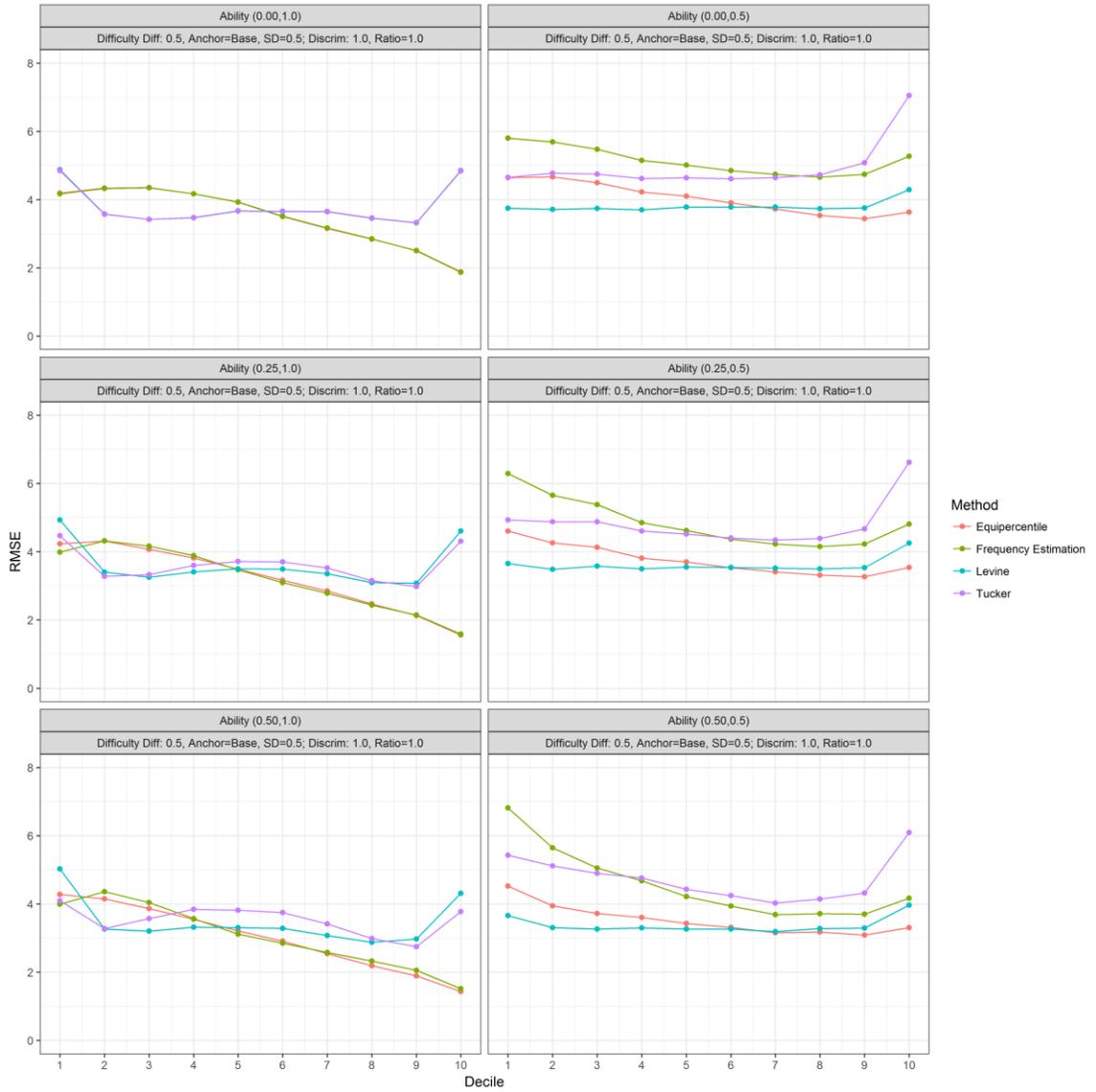


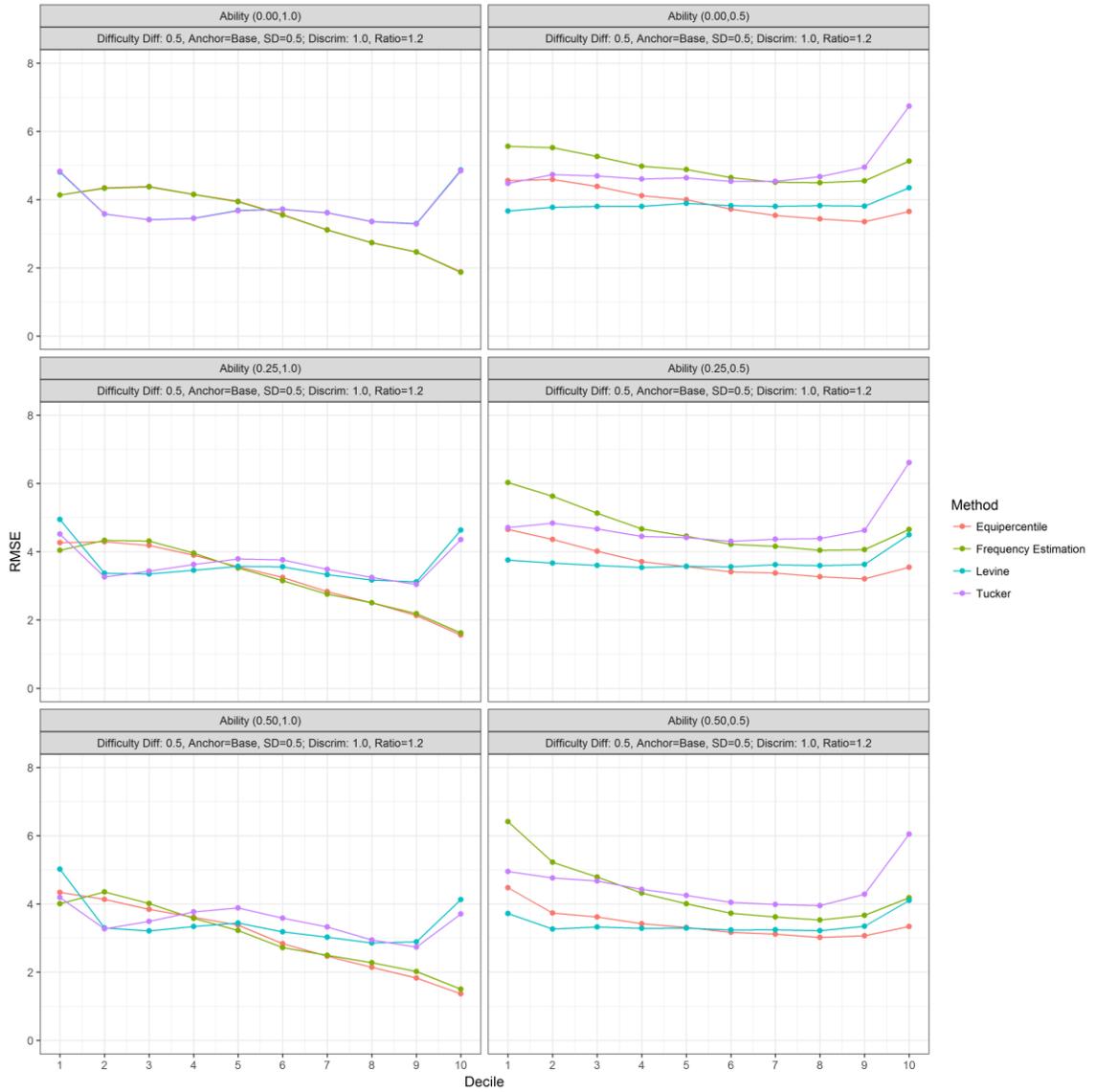


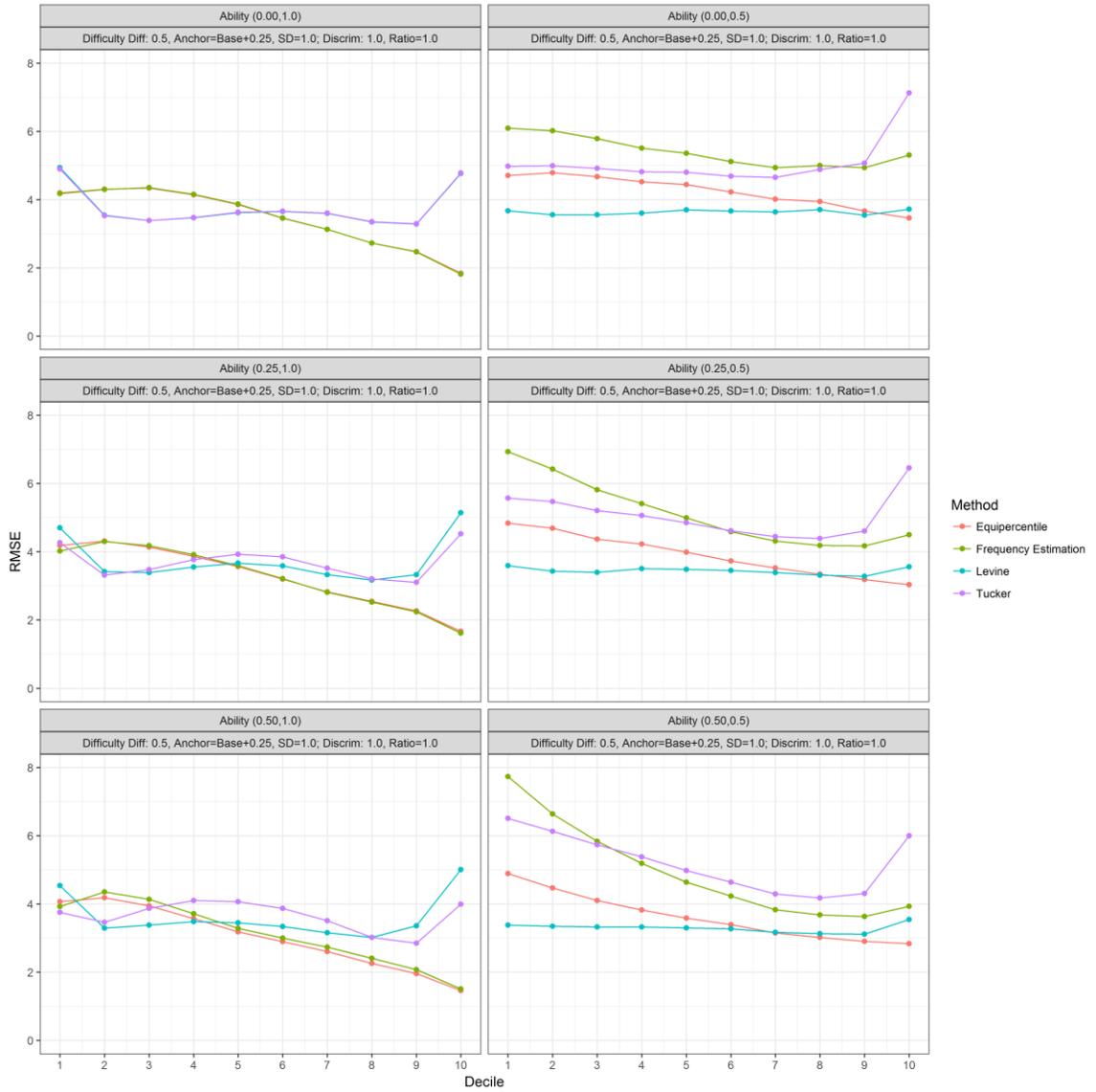


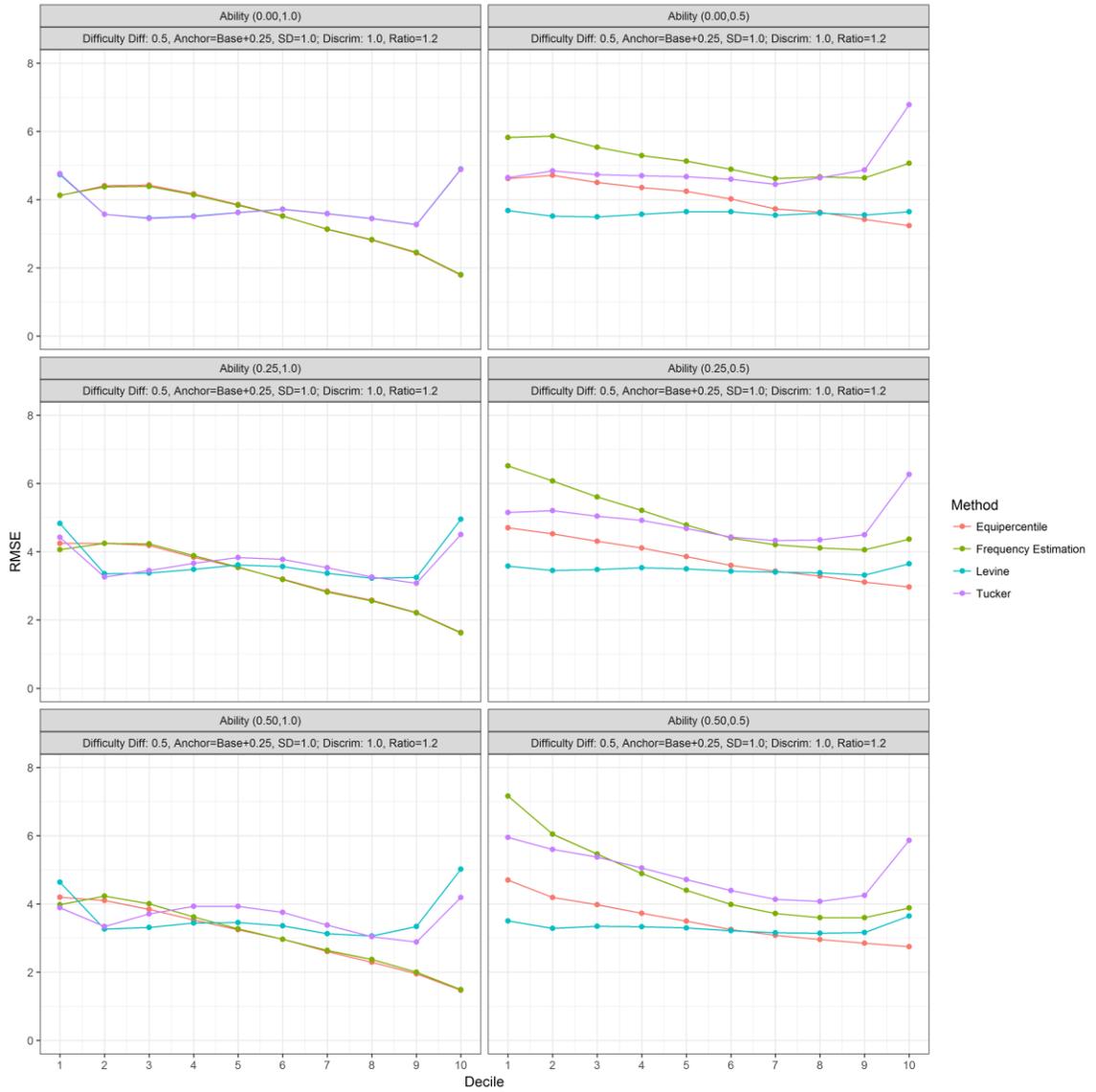


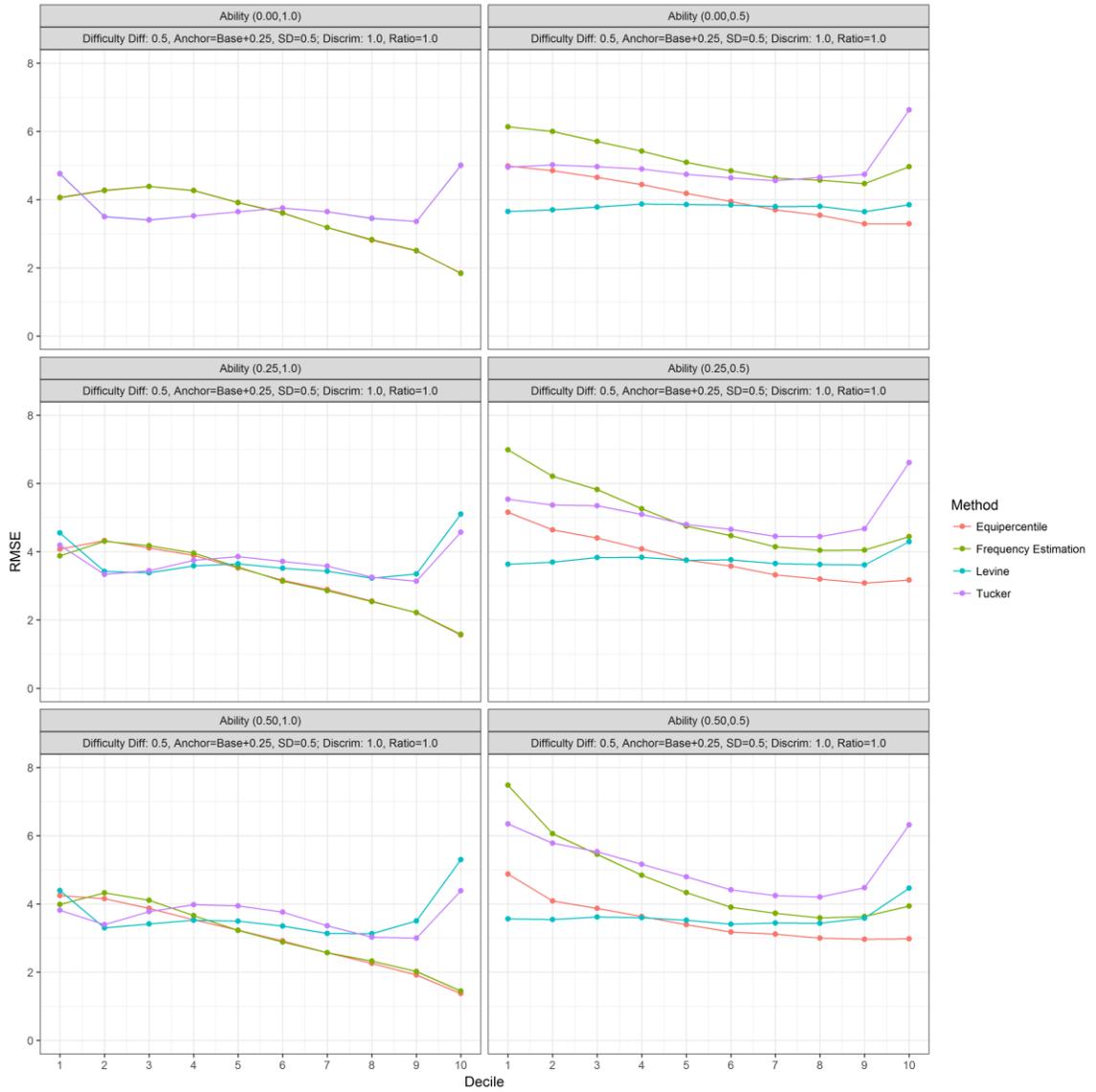


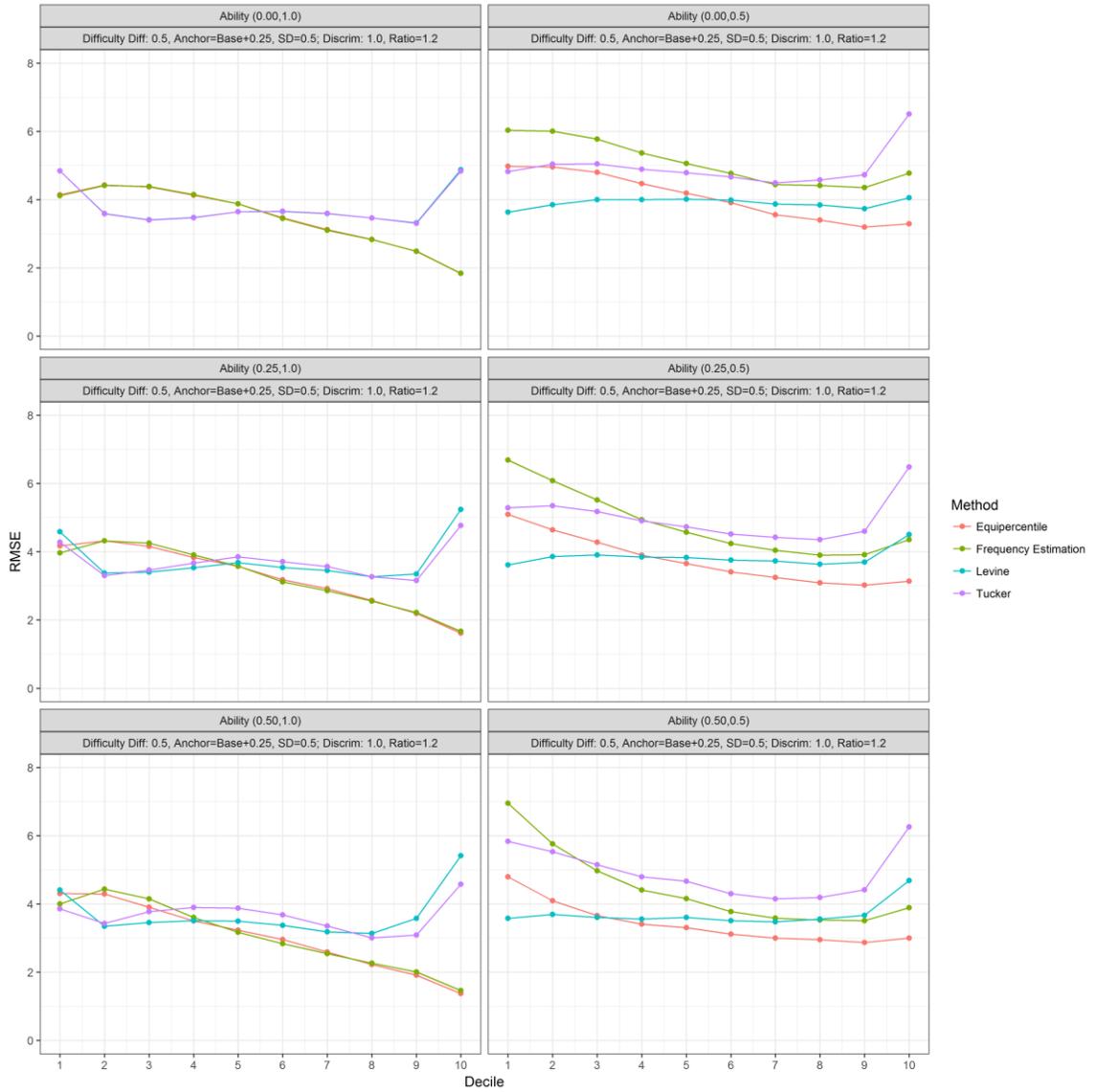




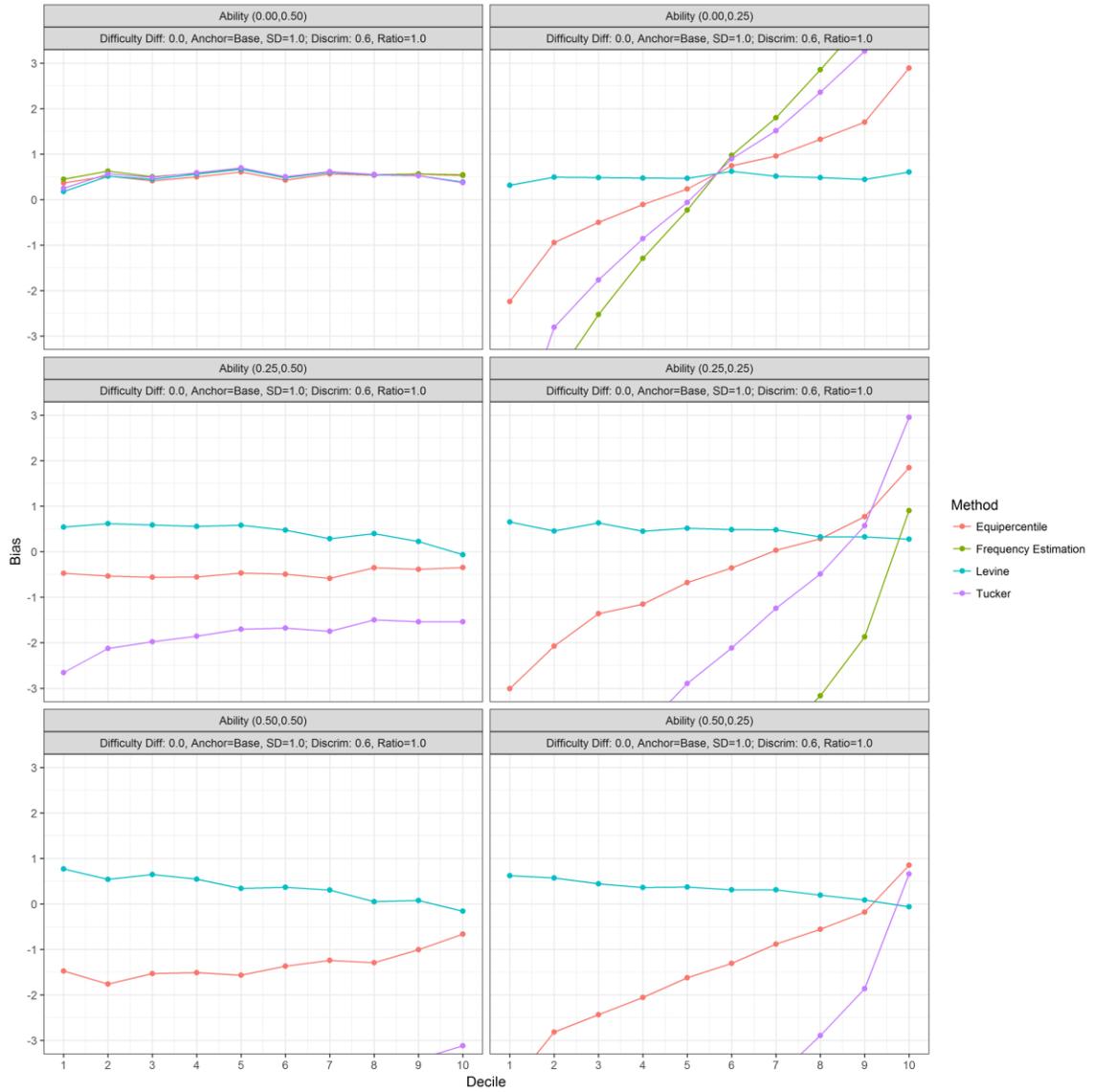


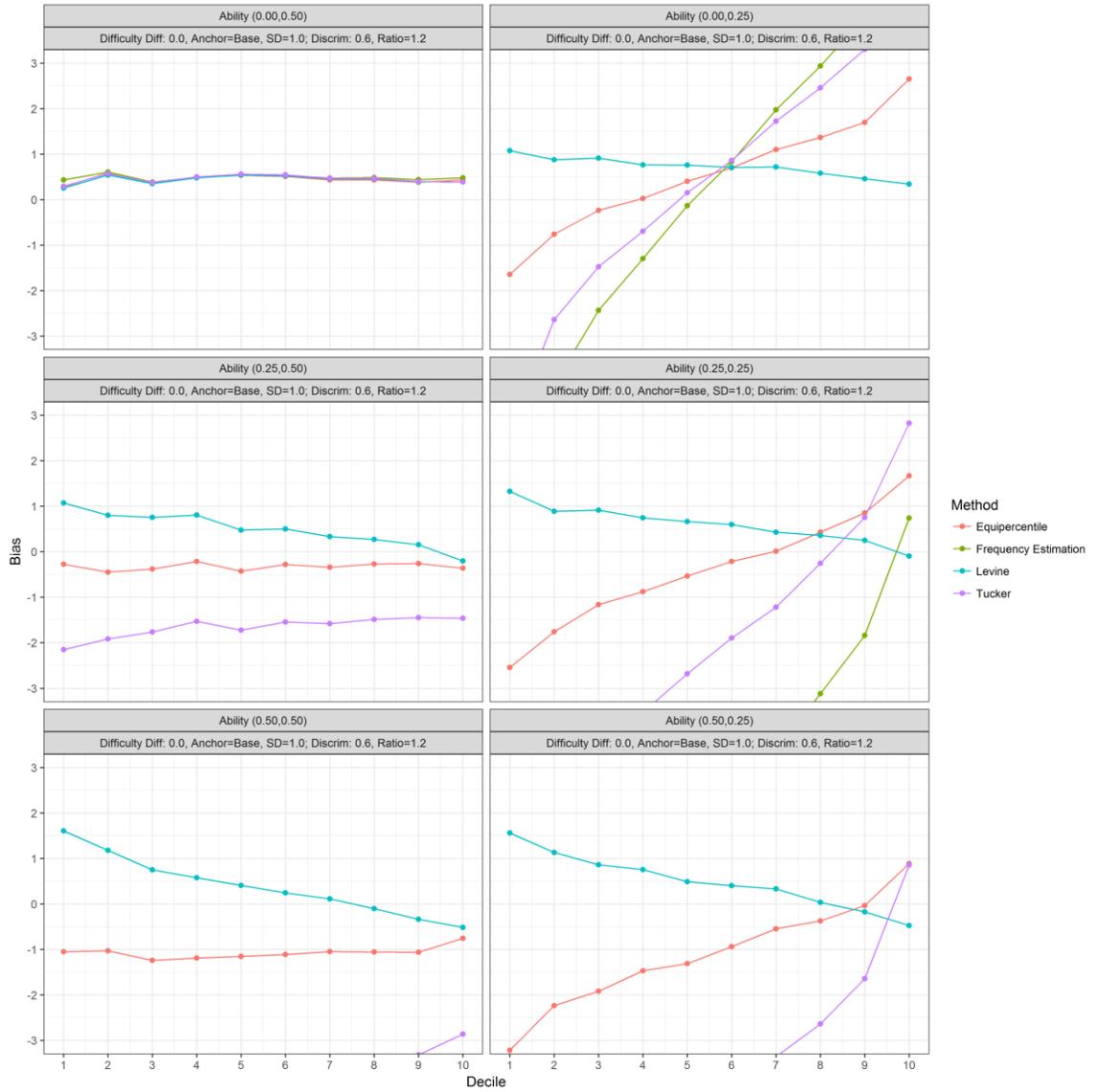


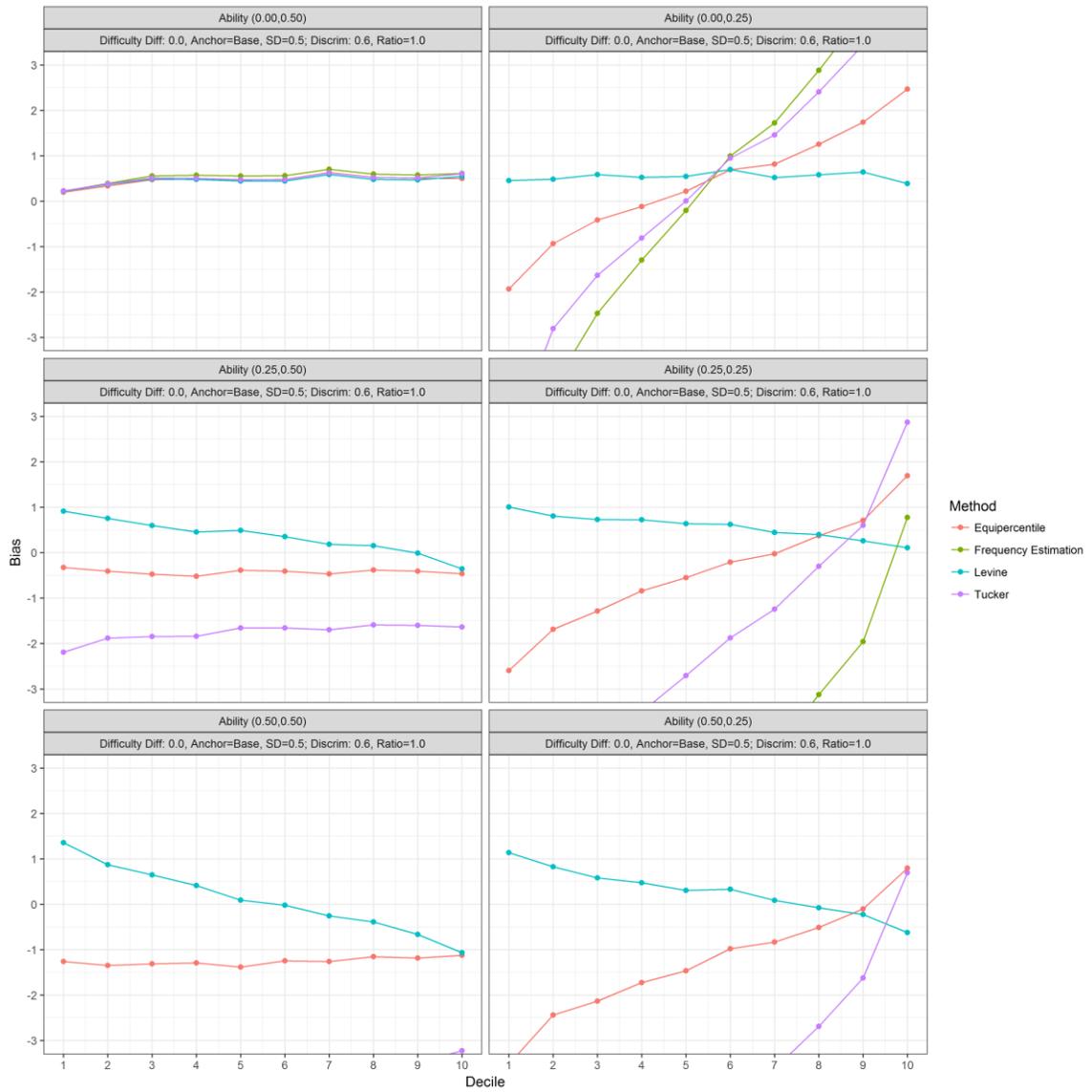


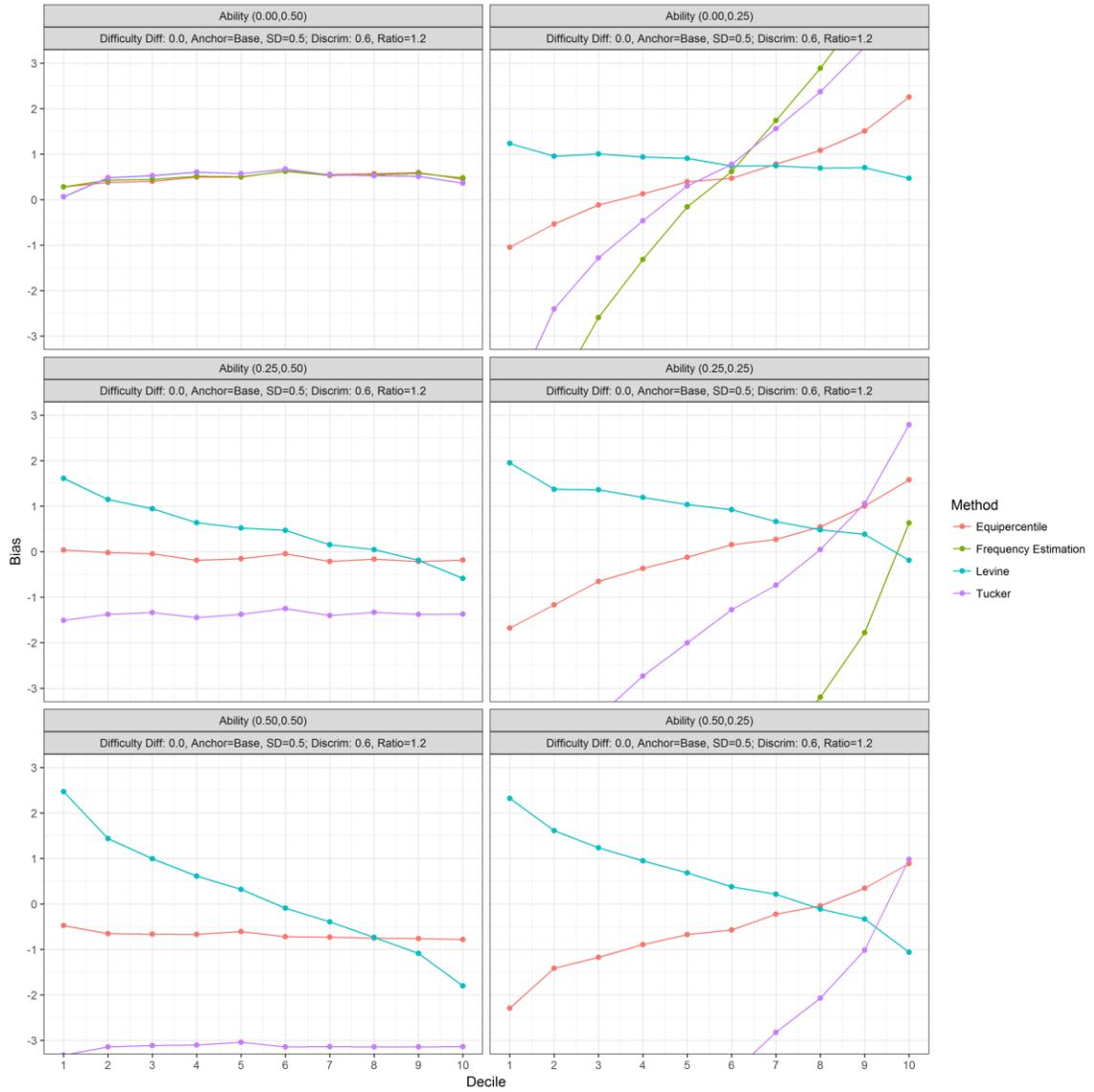


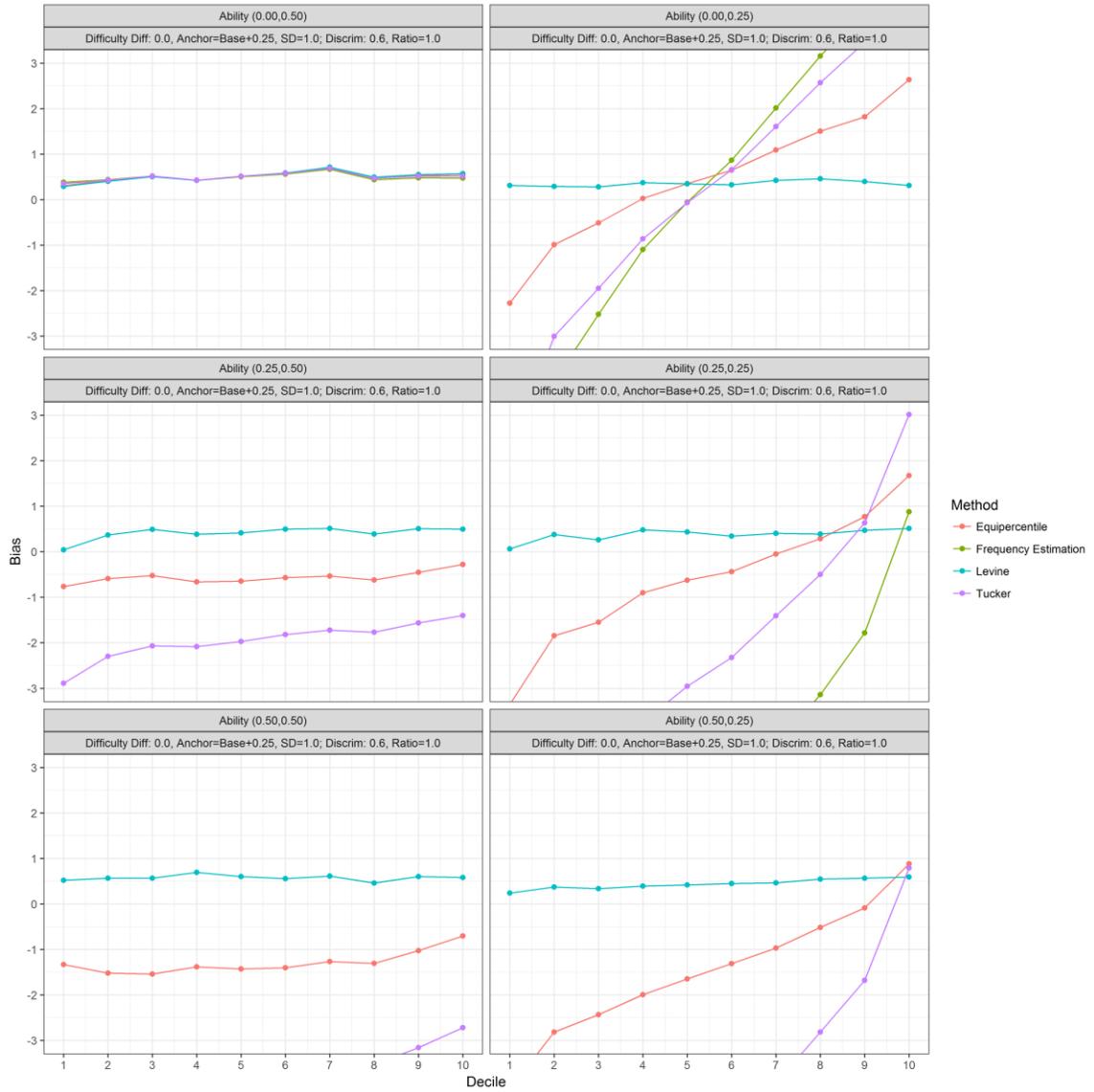
APPENDIX C  
CERTIFICATION BIAS RESULTS

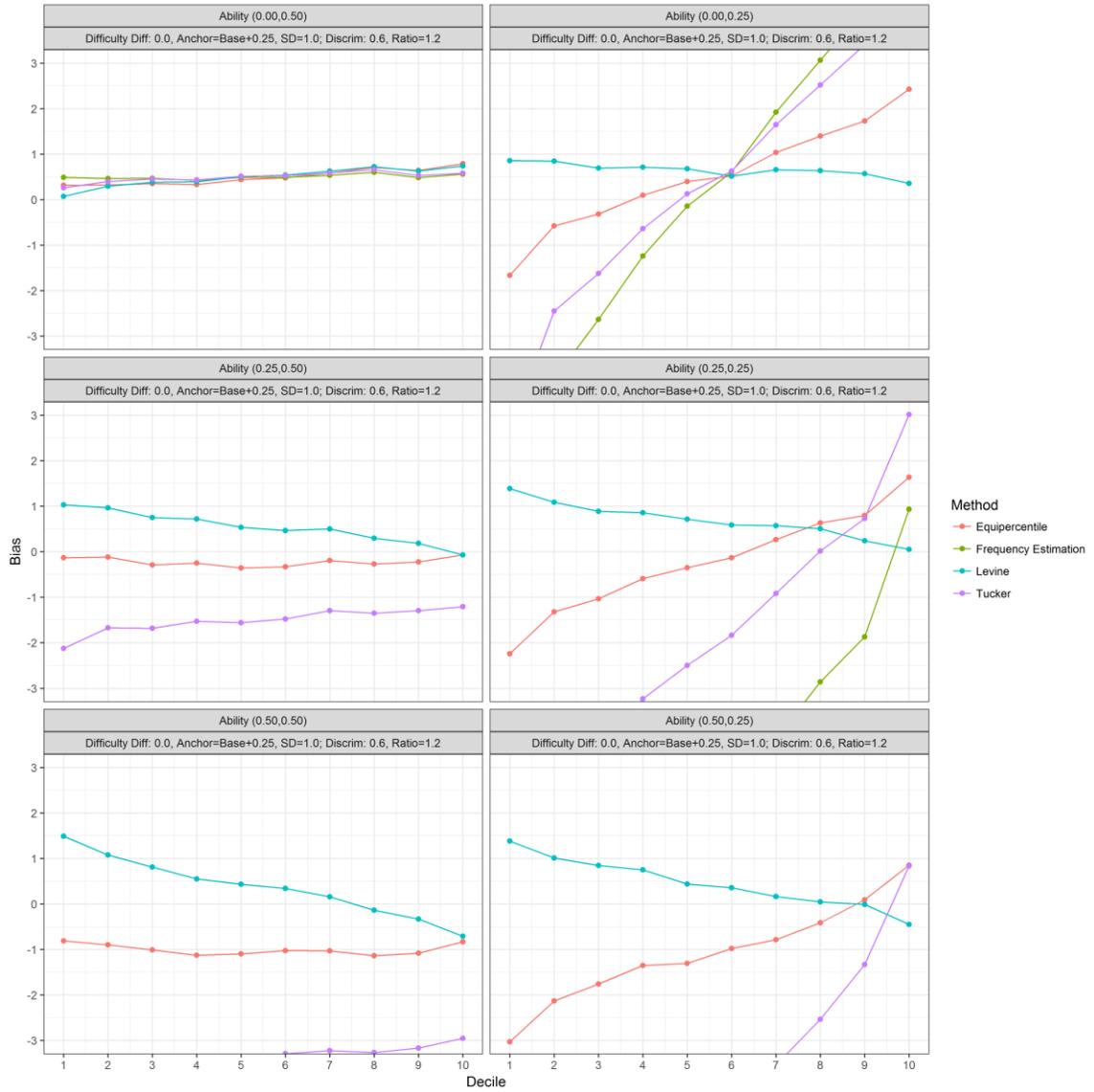


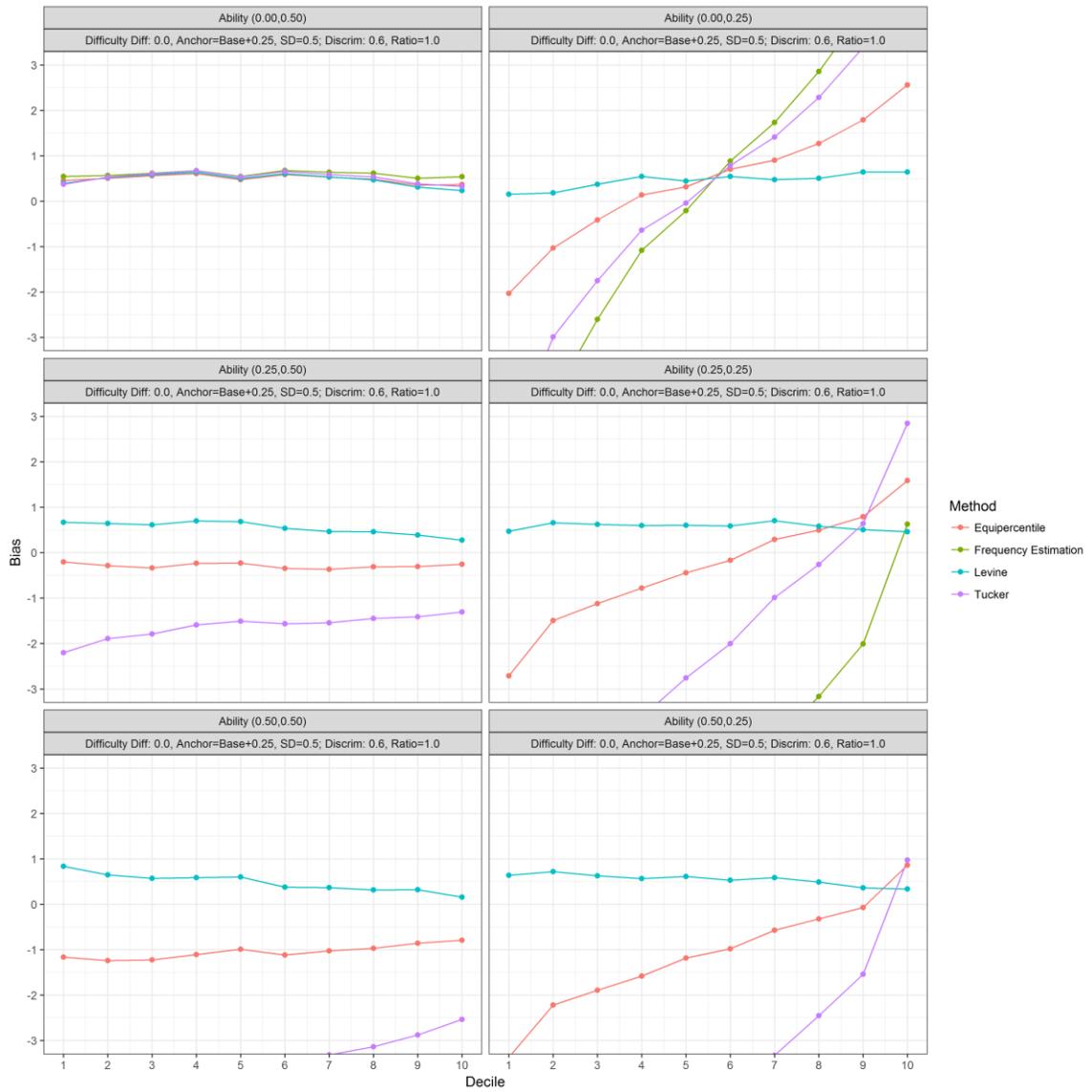


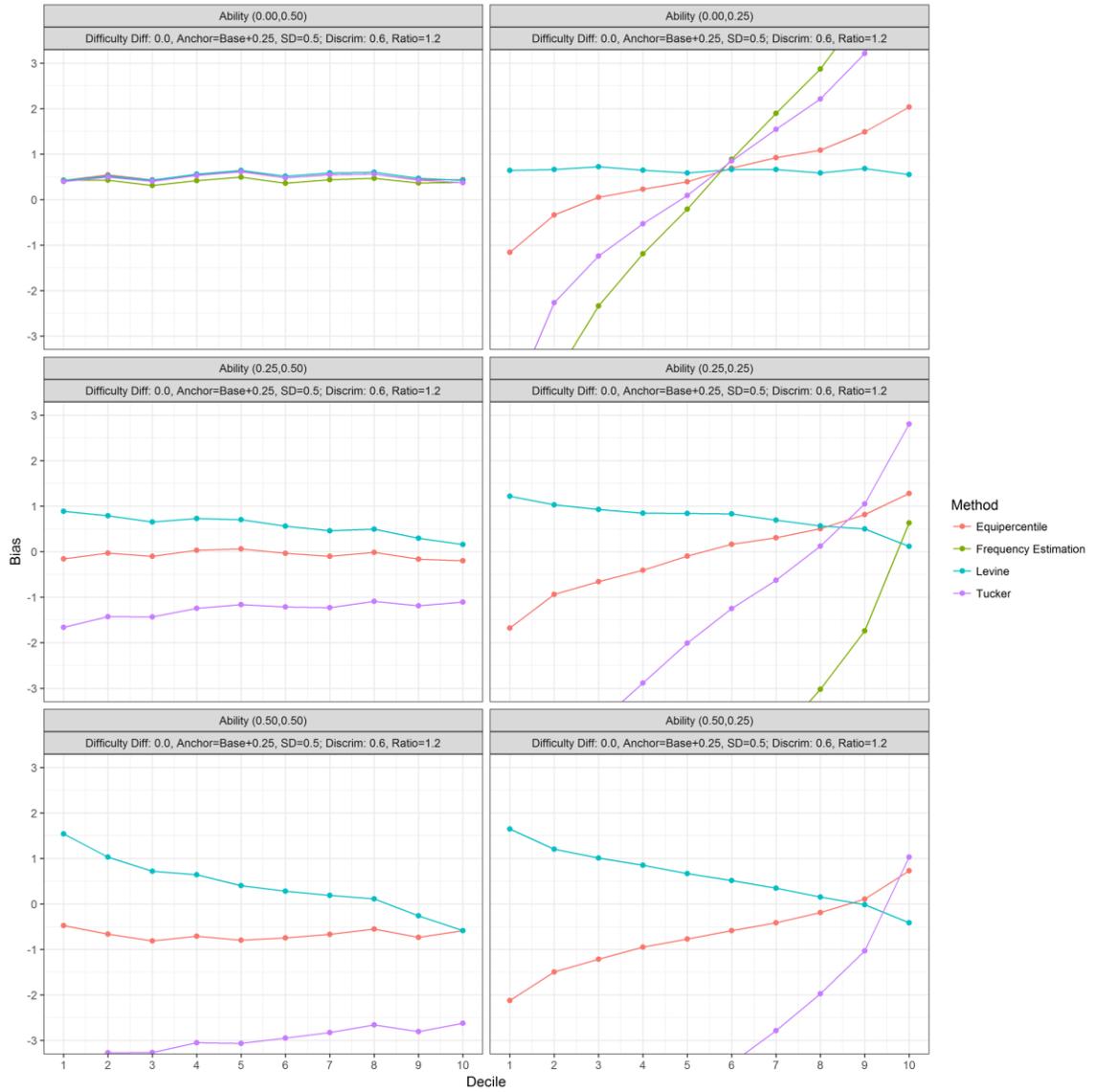


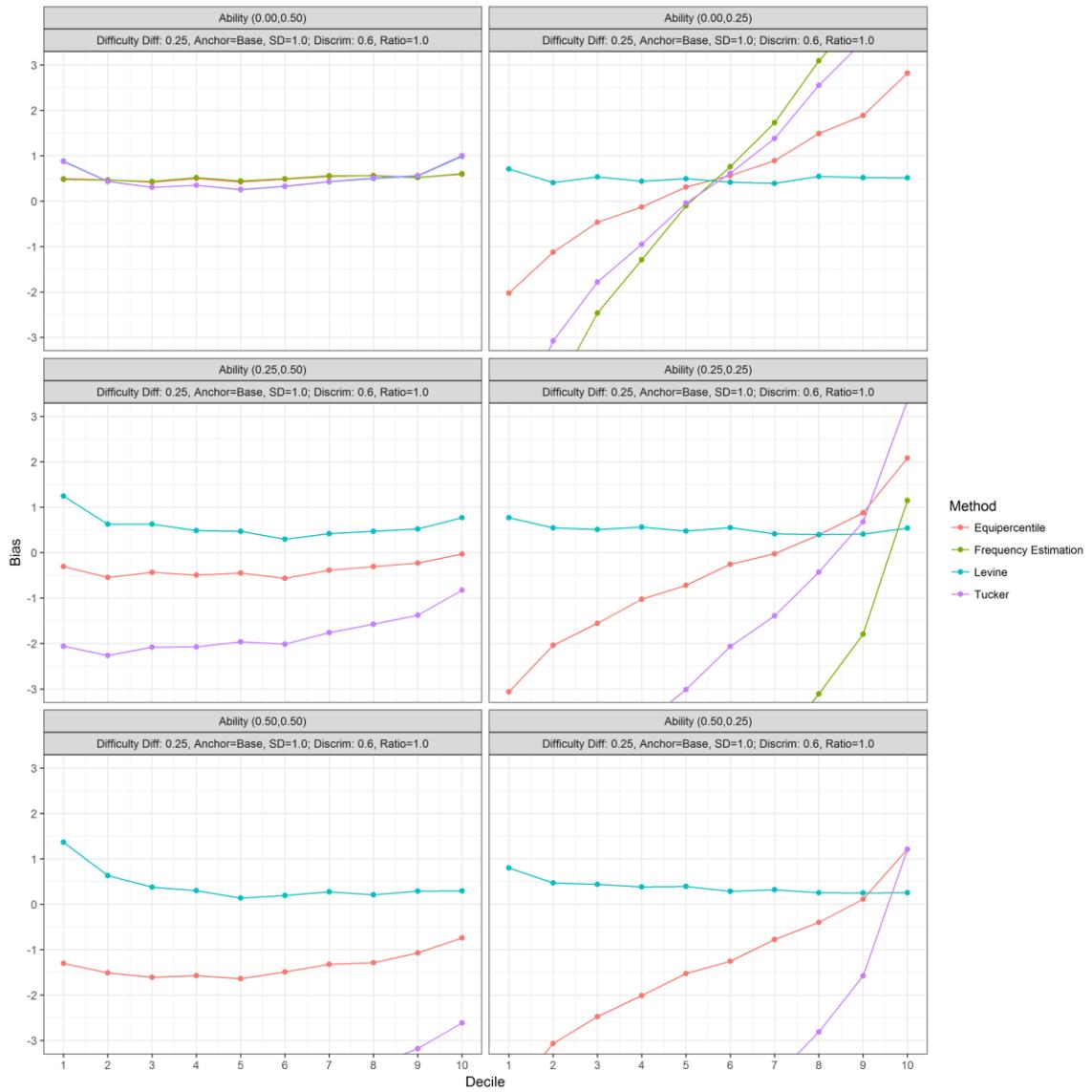


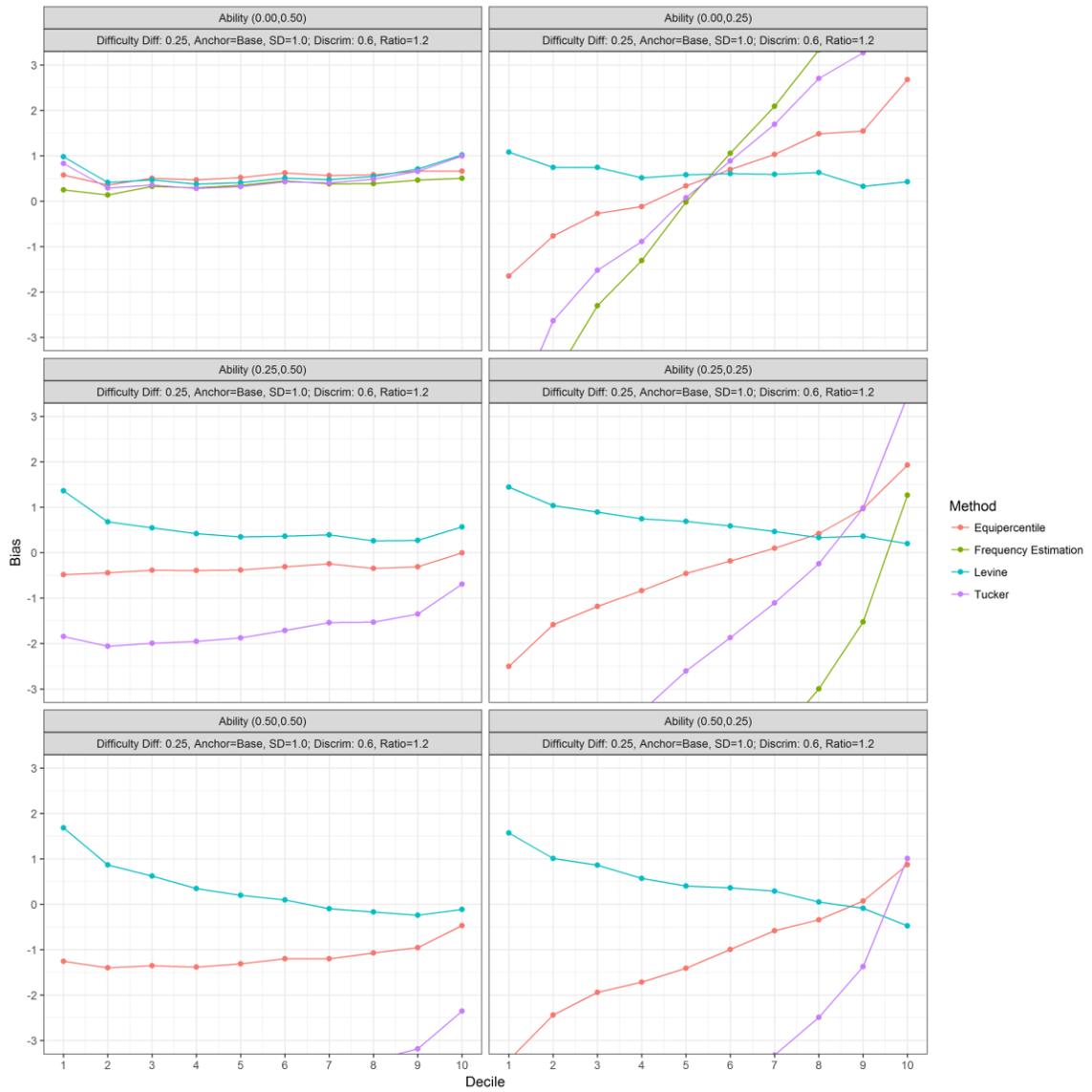


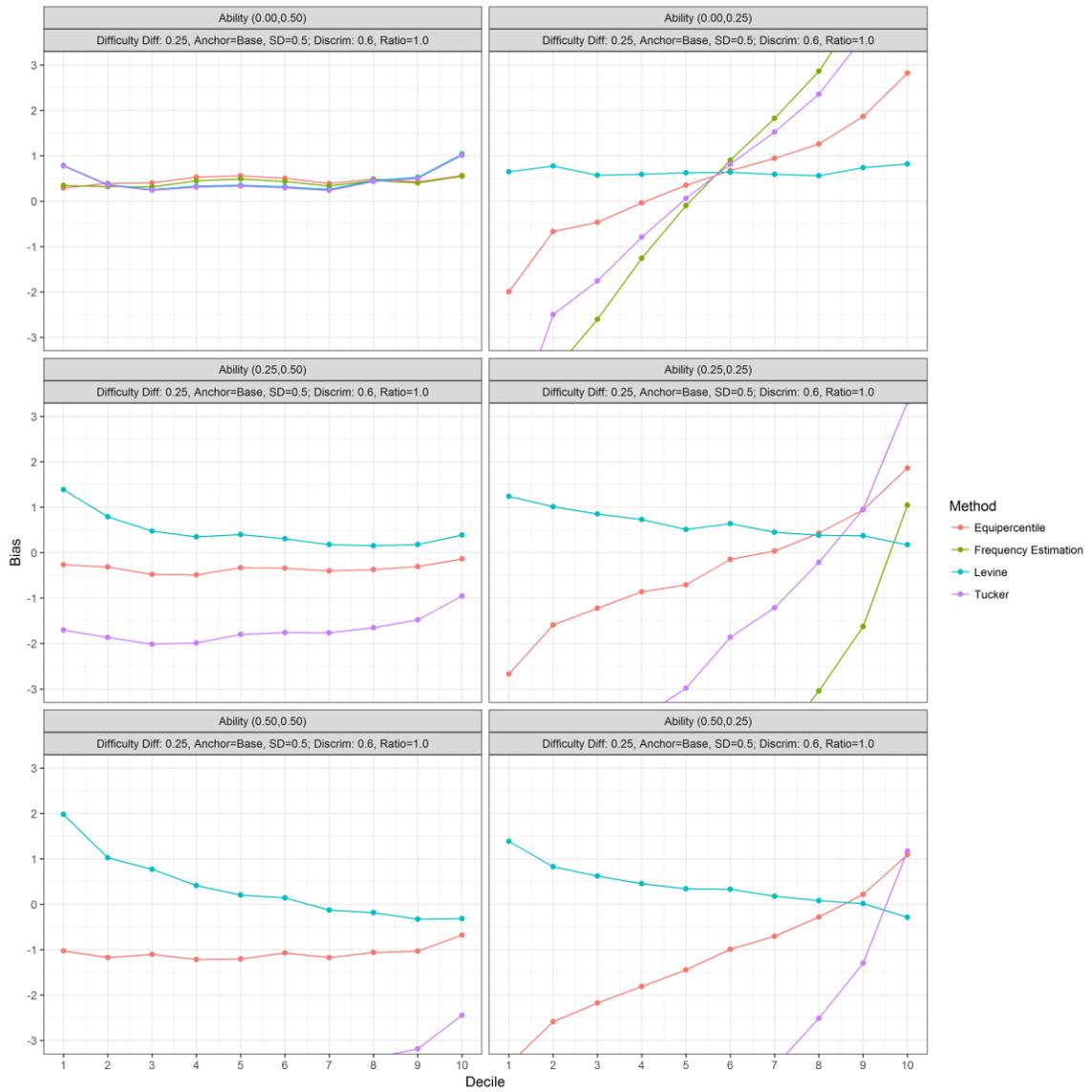


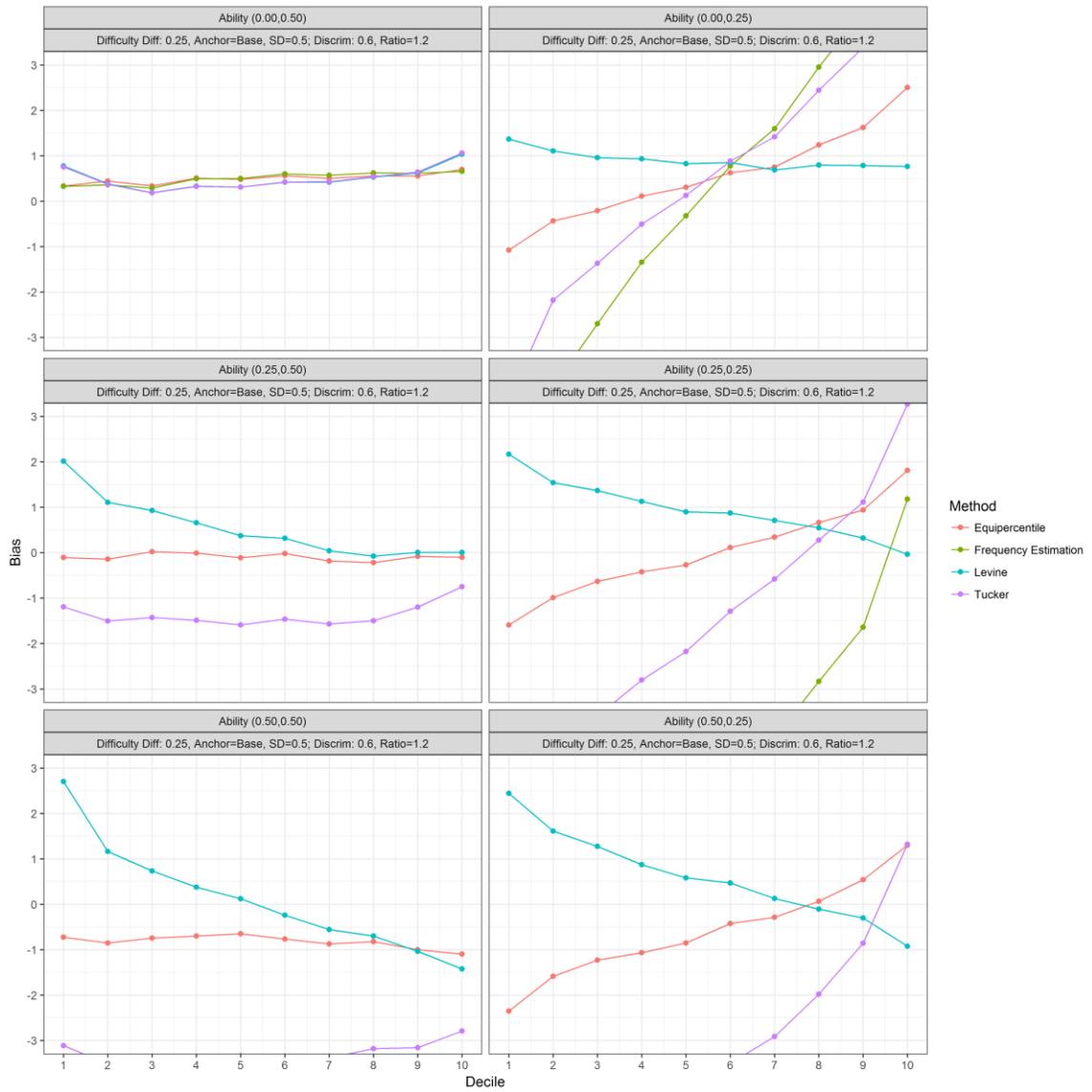


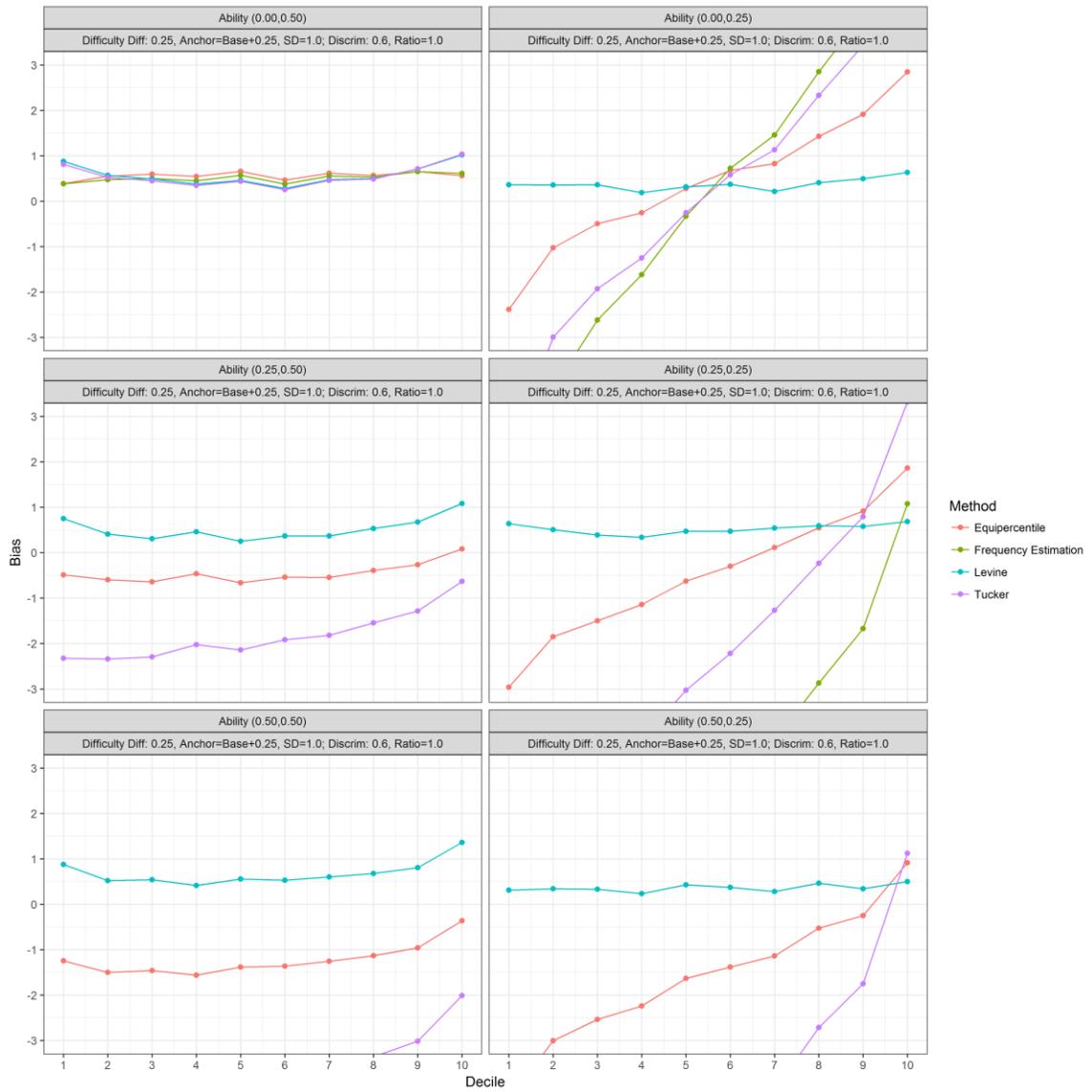


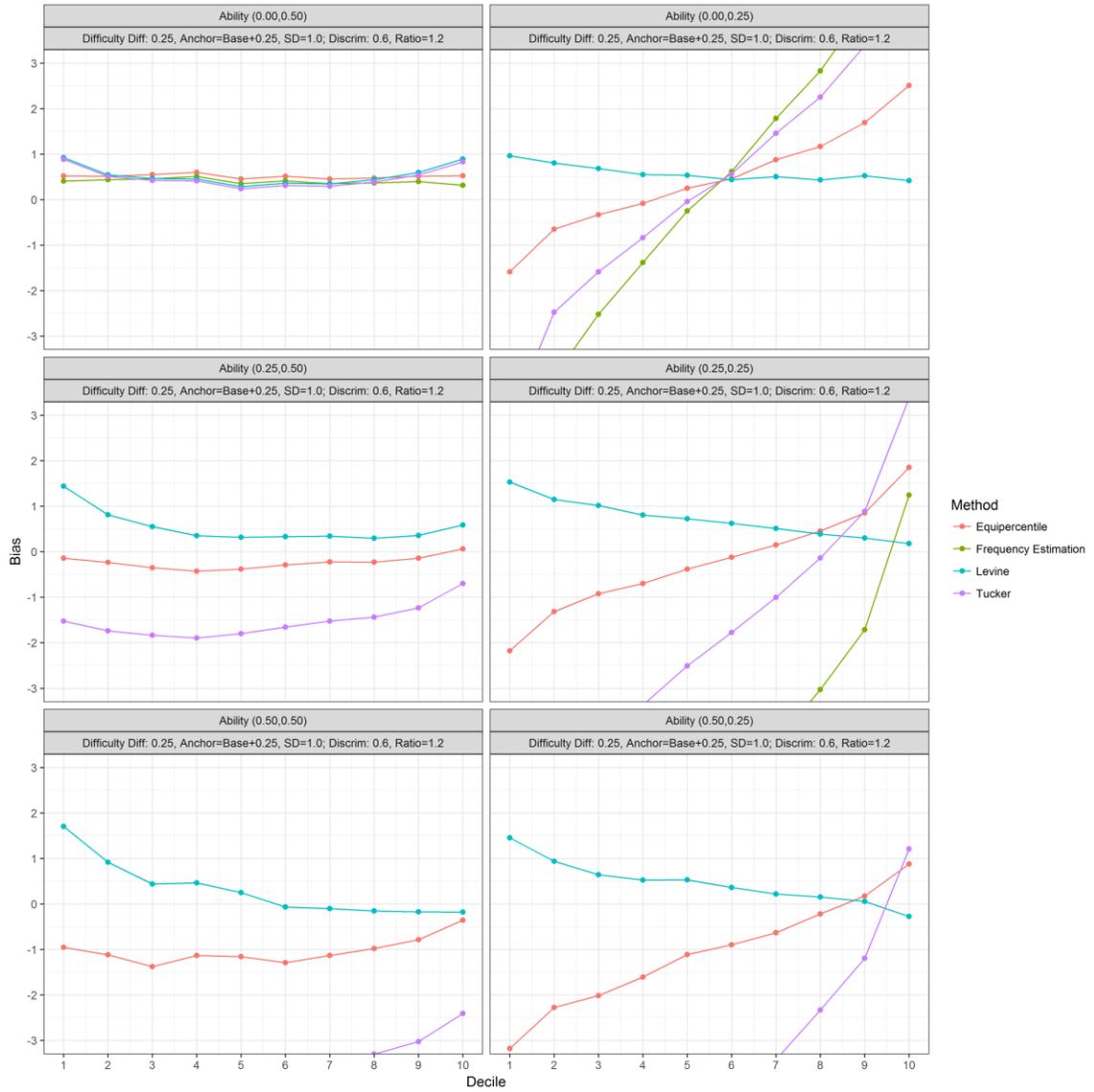


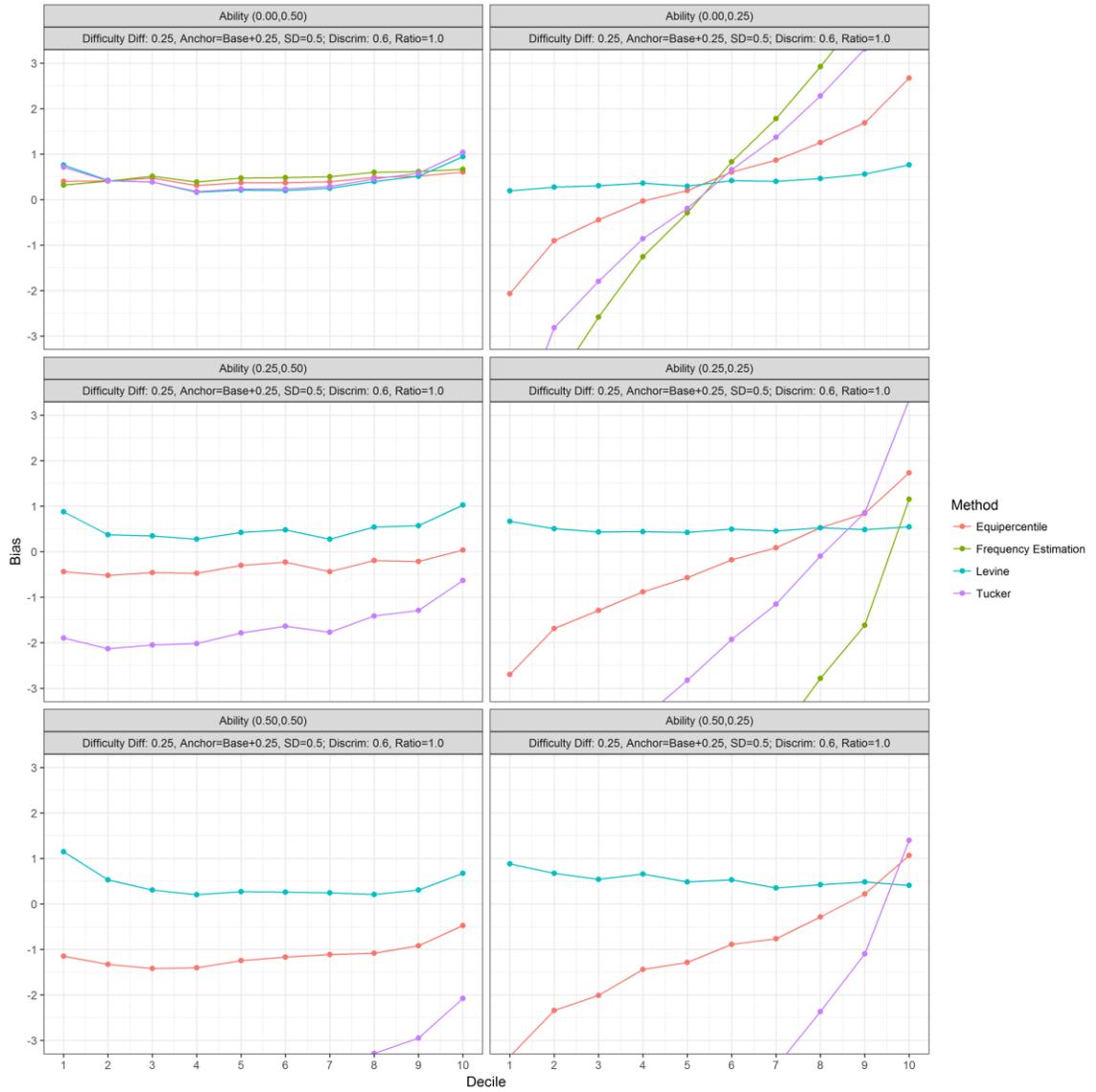


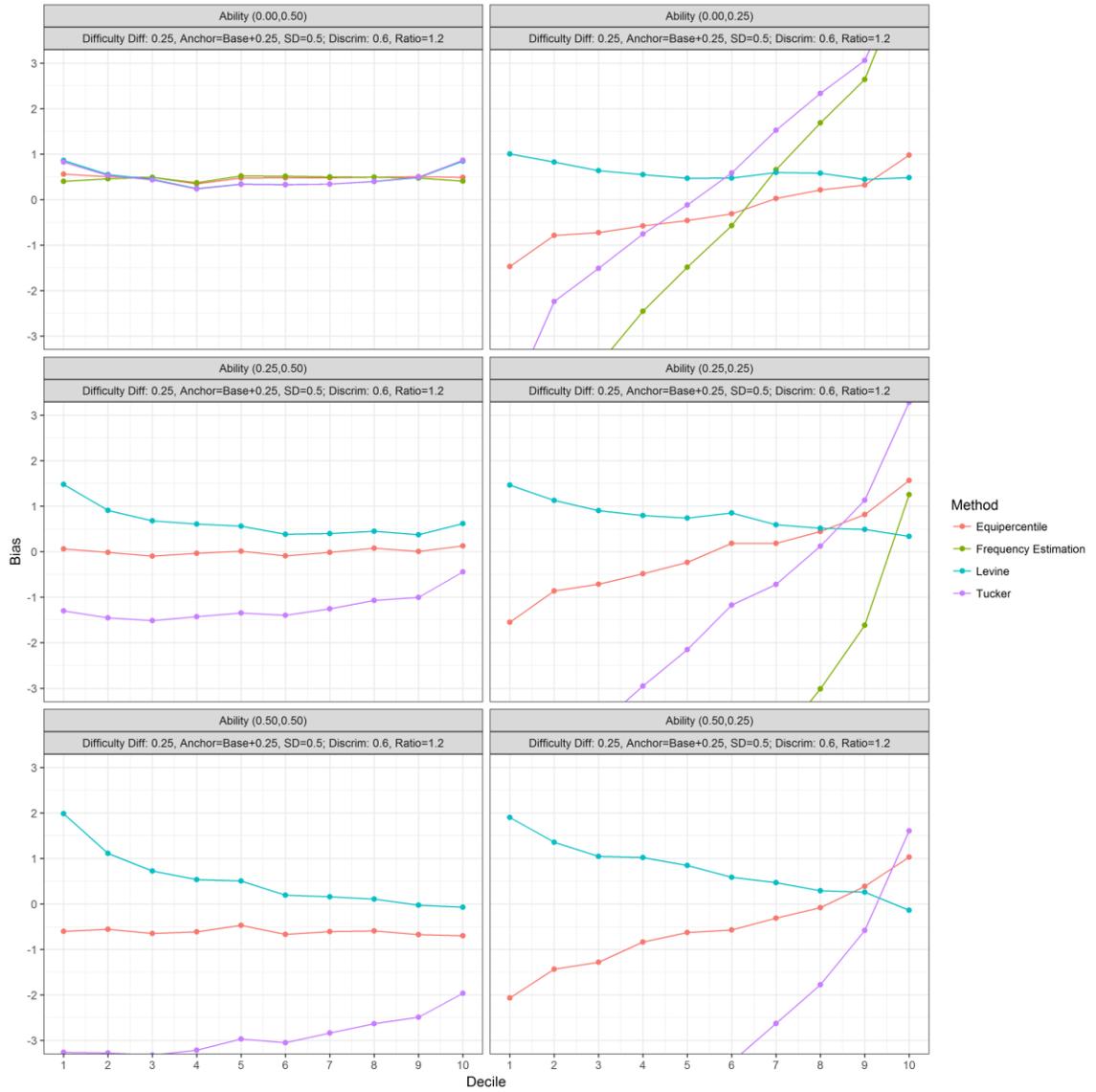


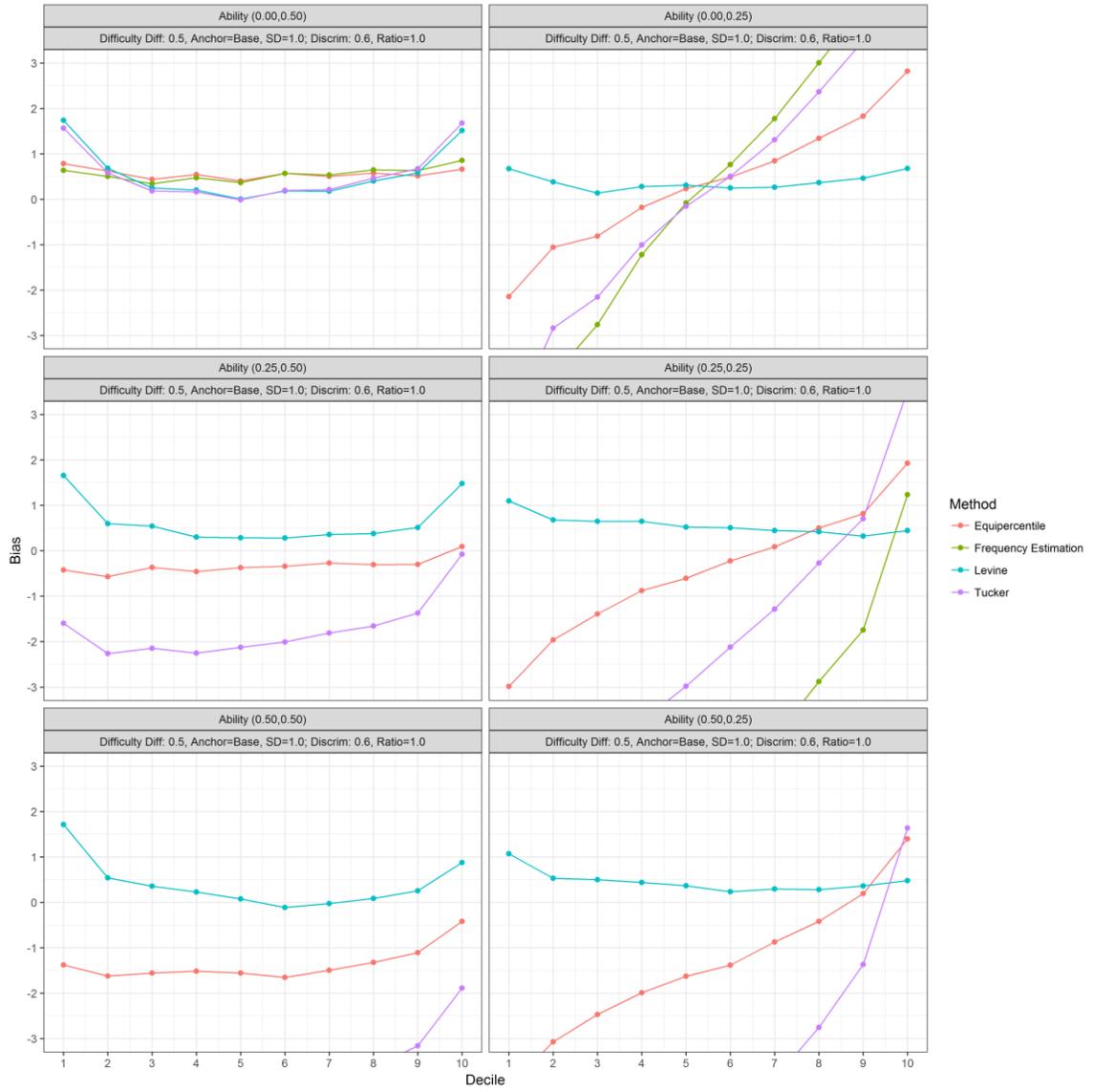


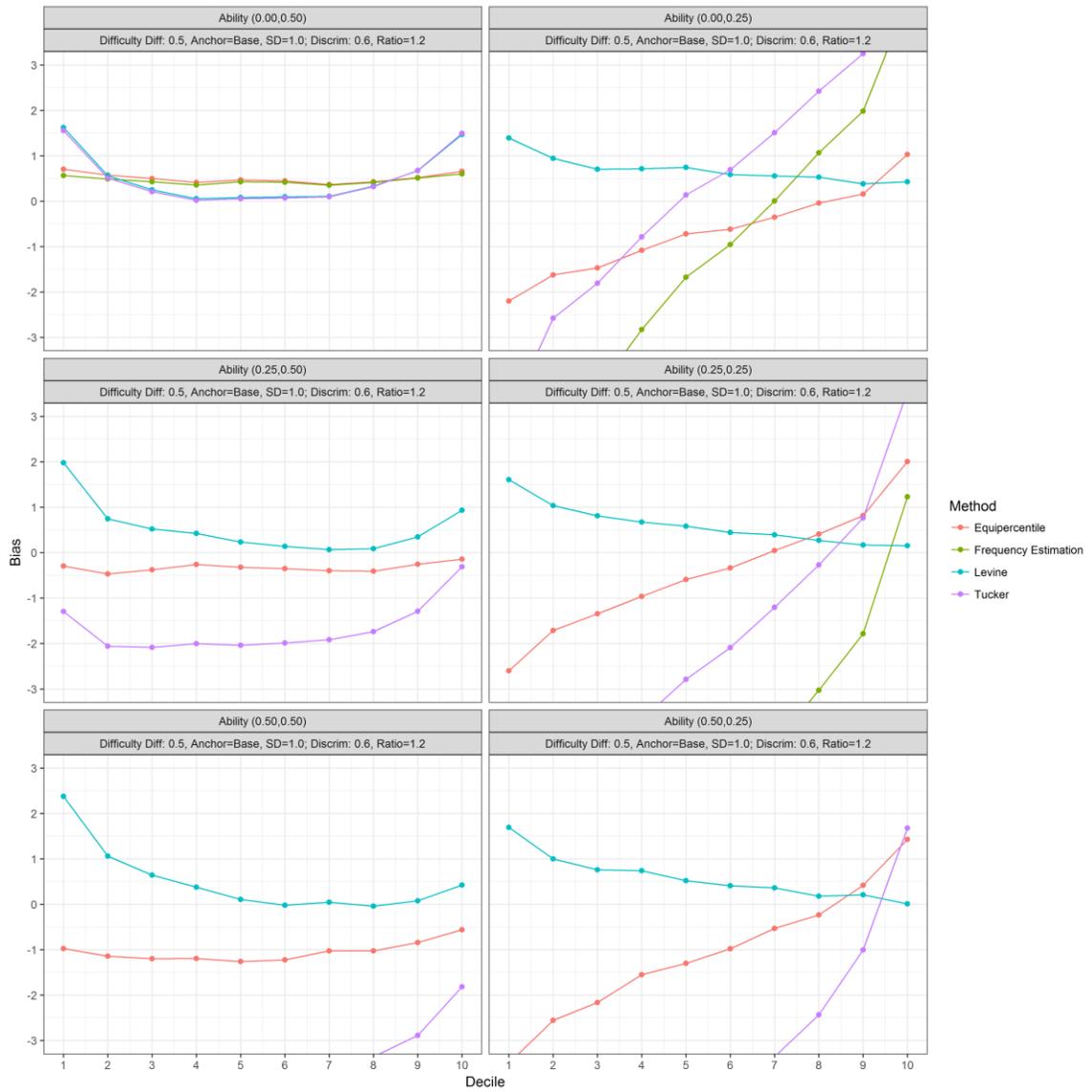


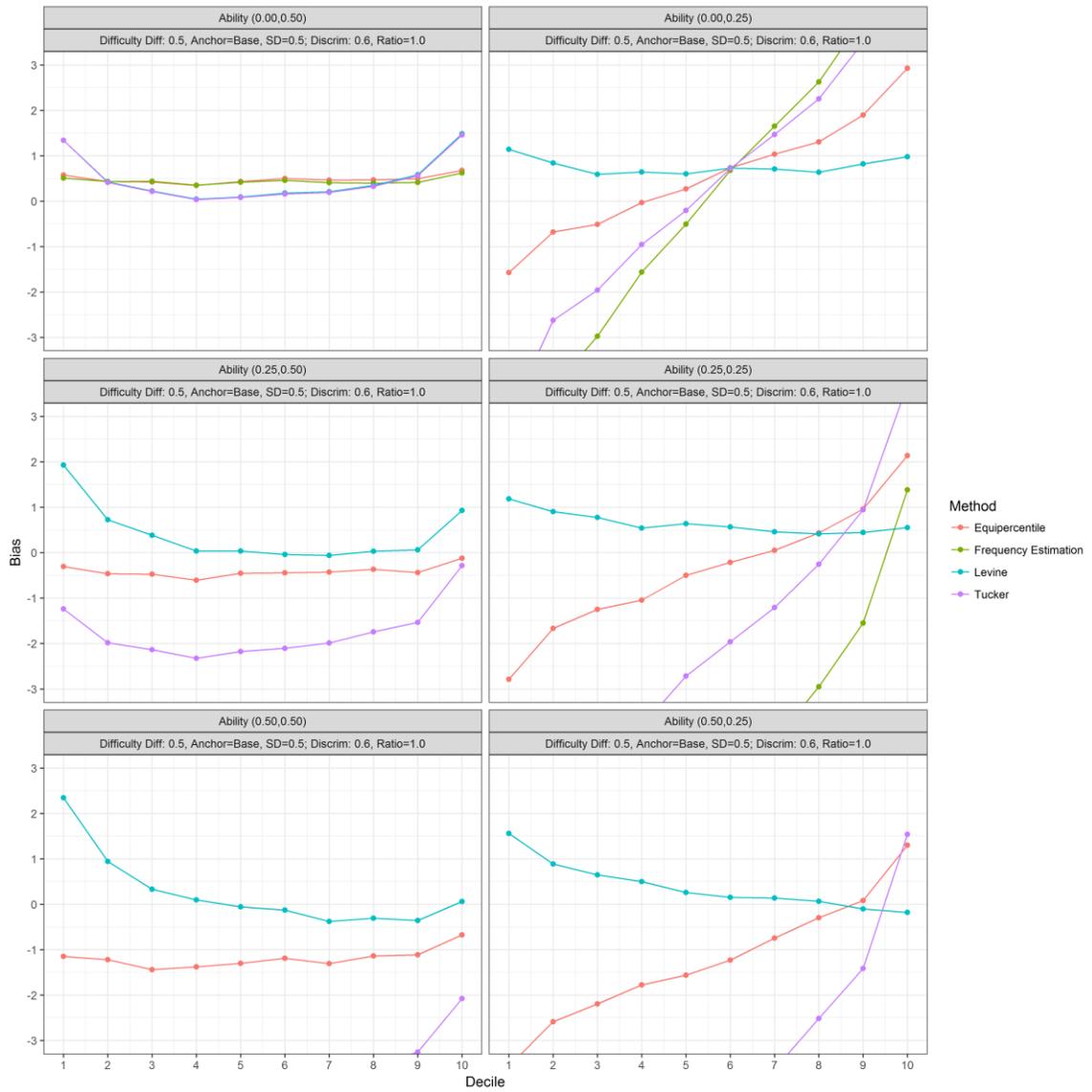


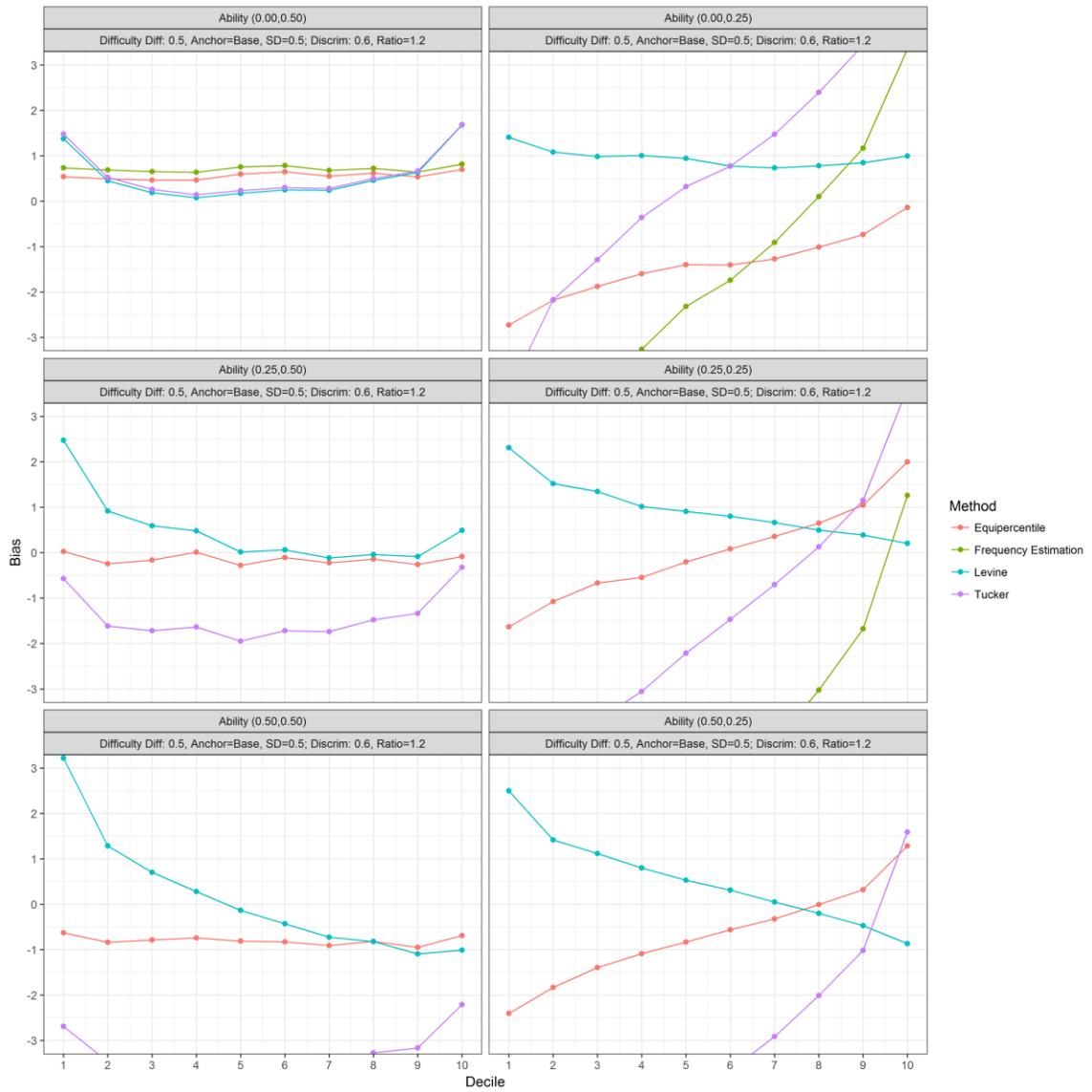


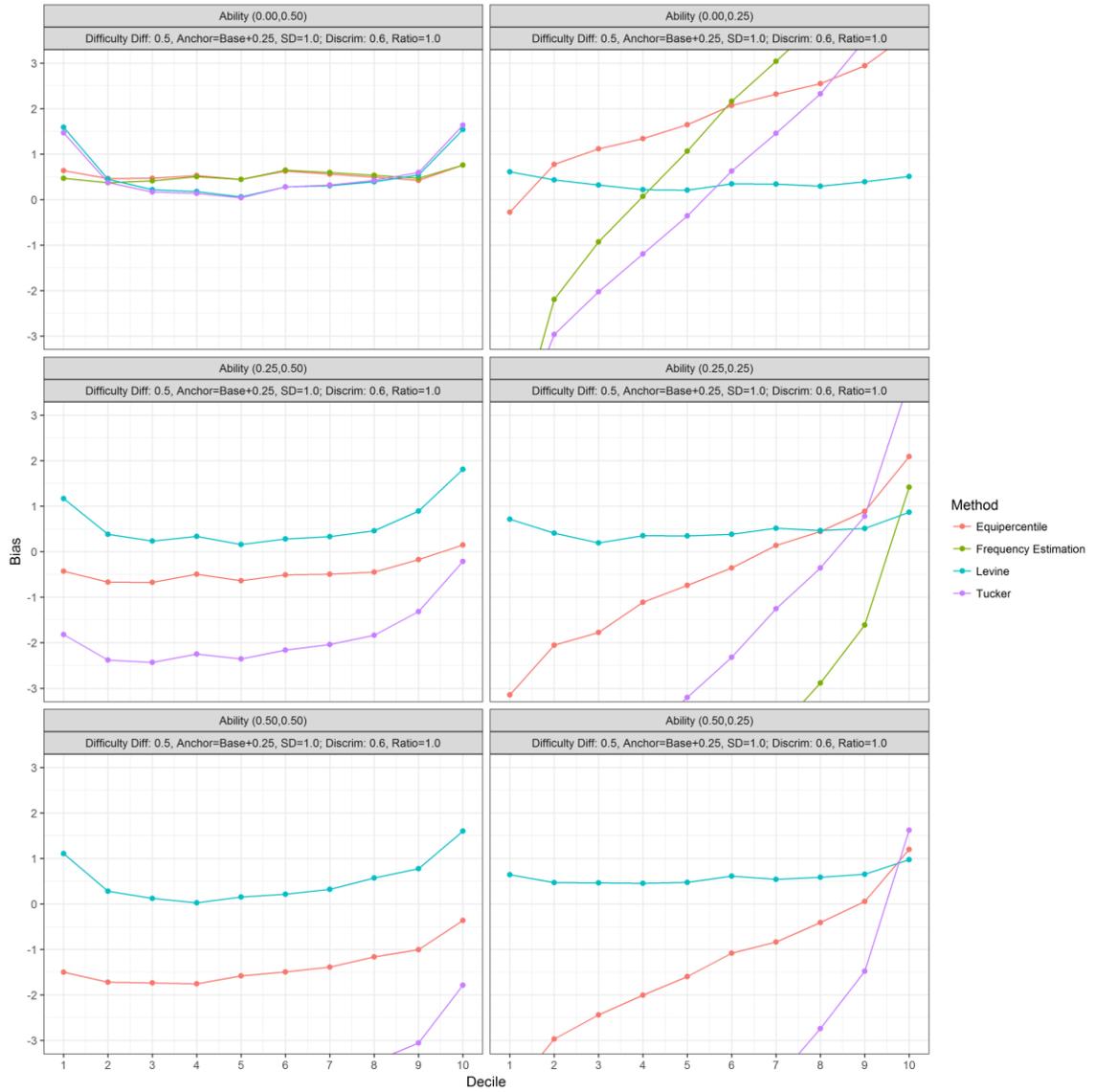


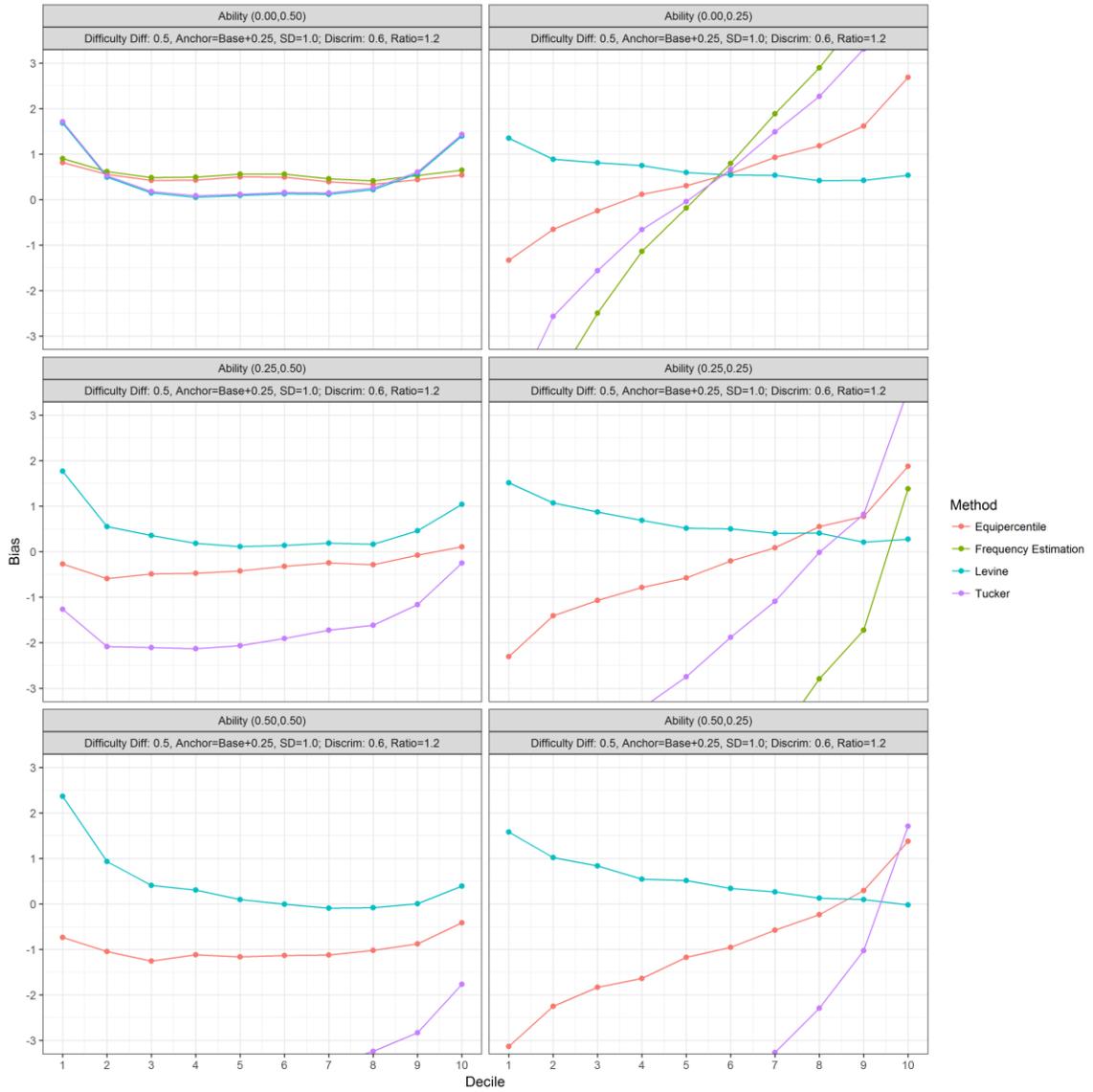


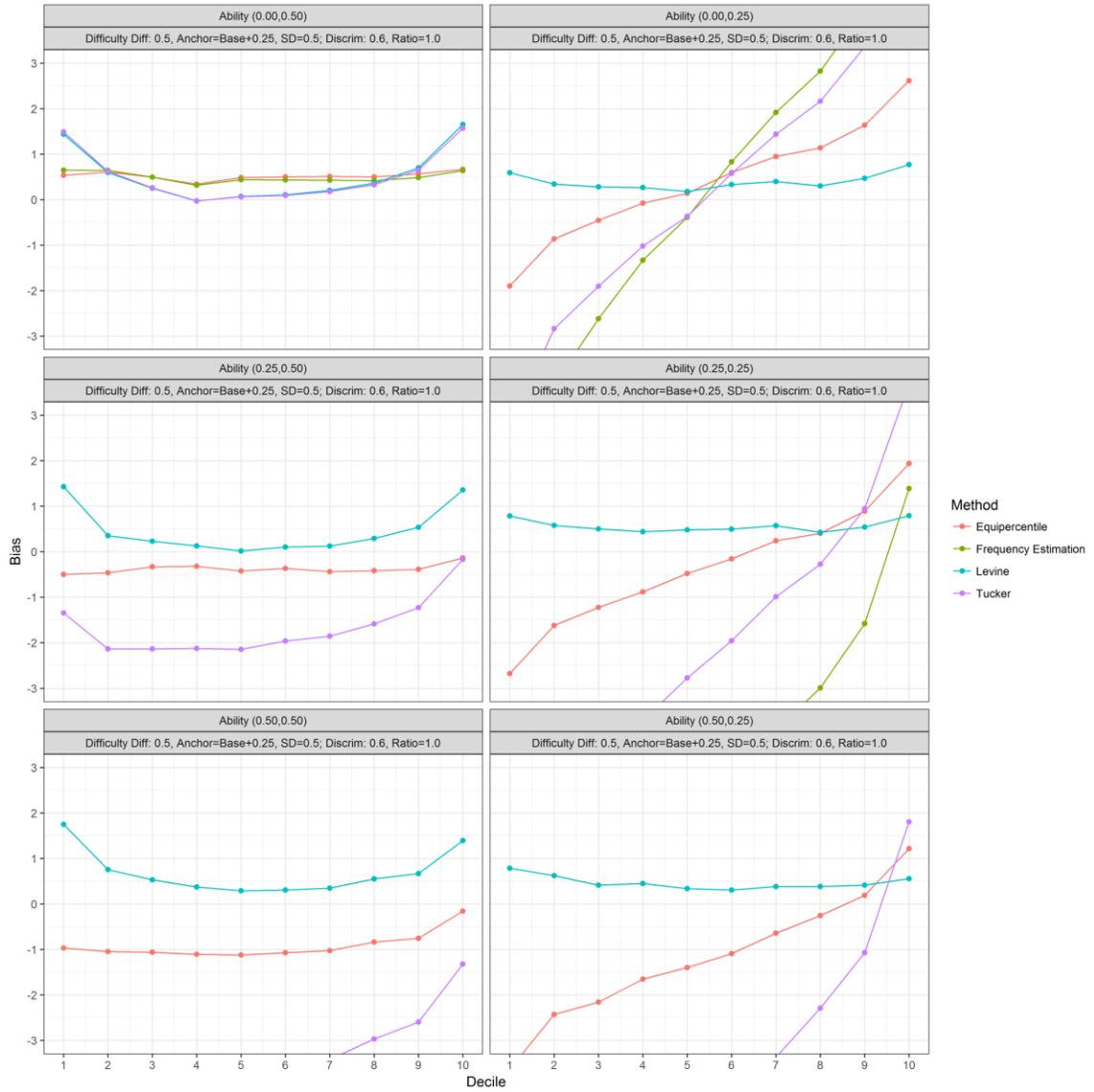


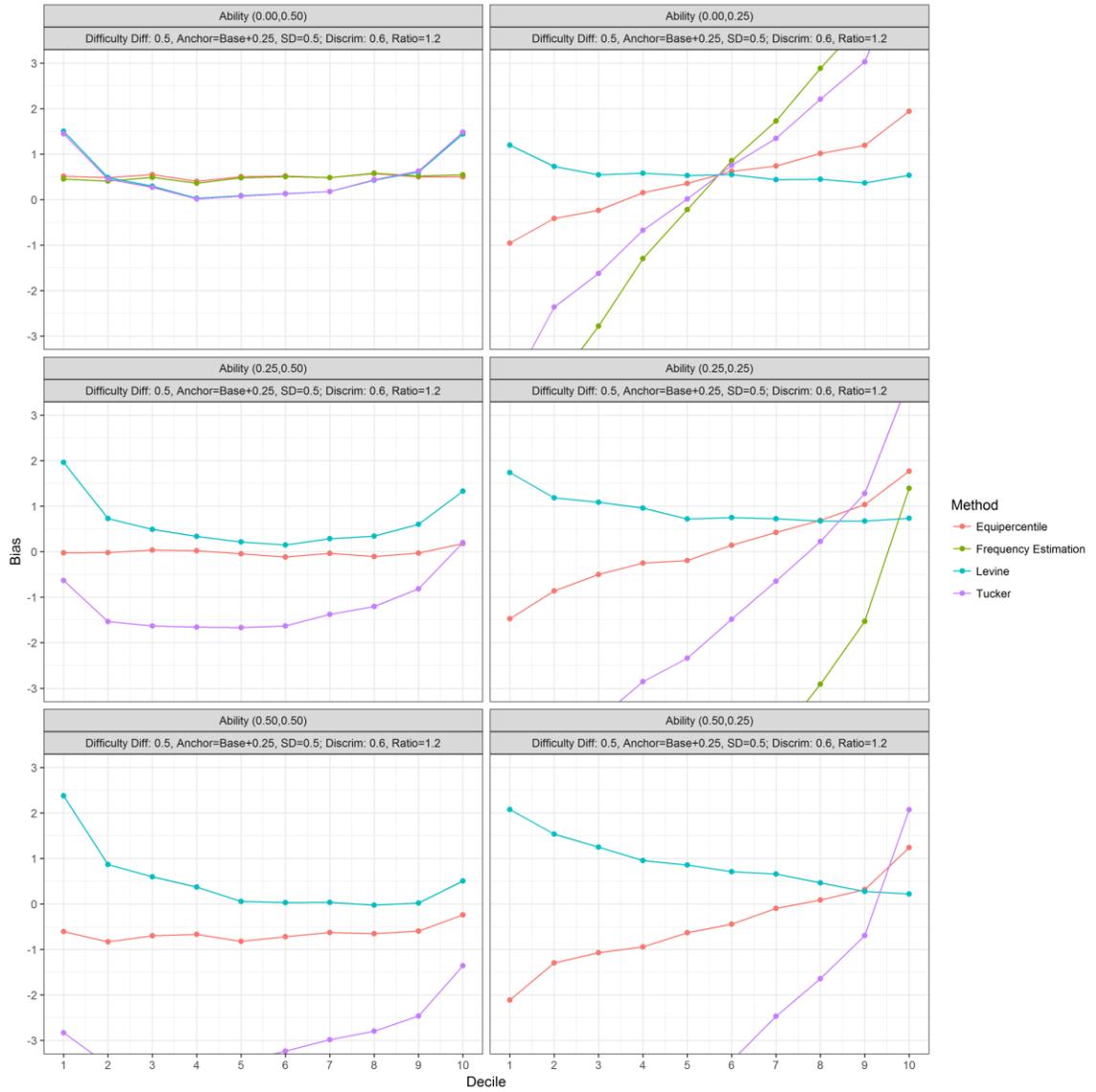


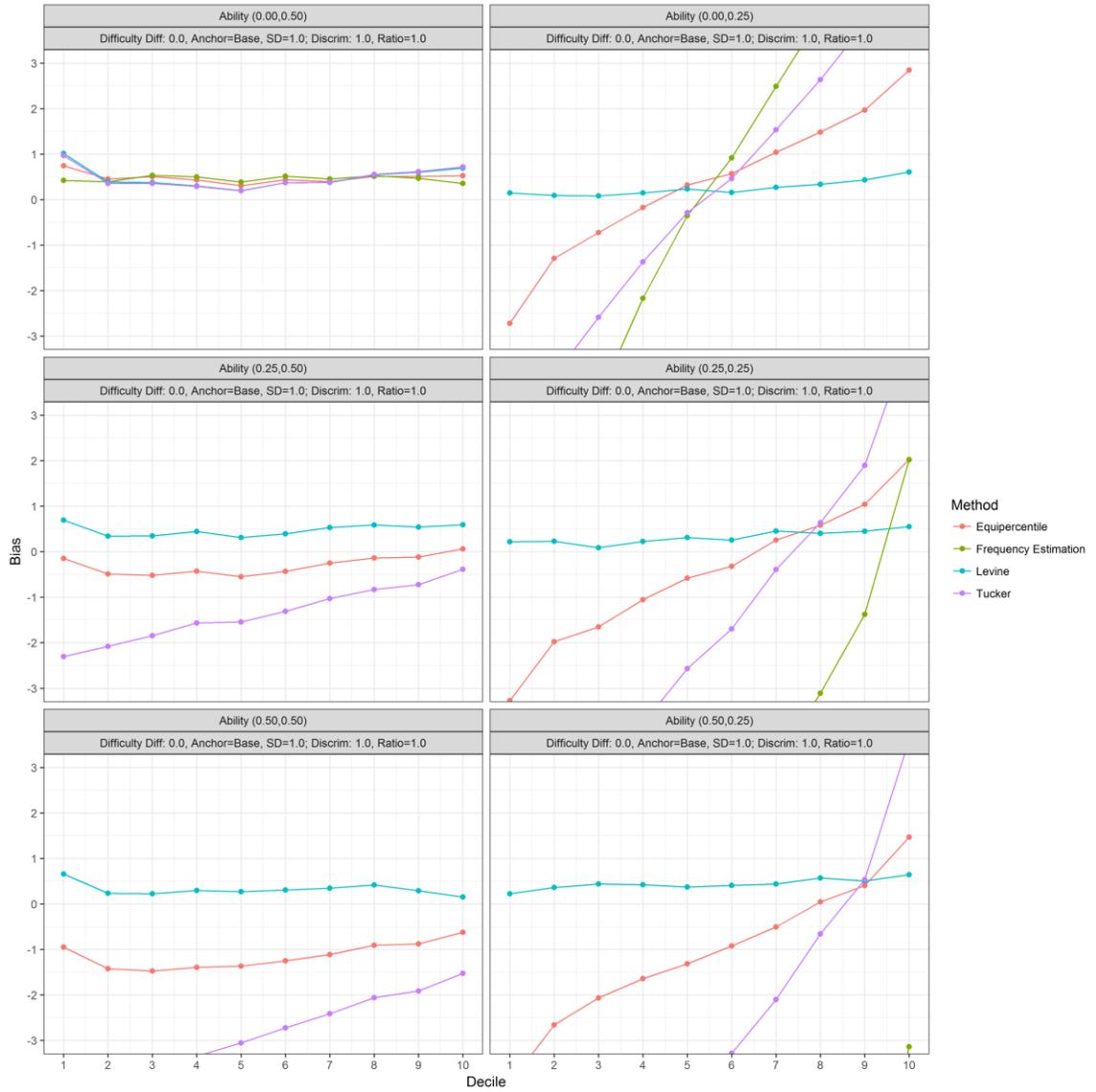


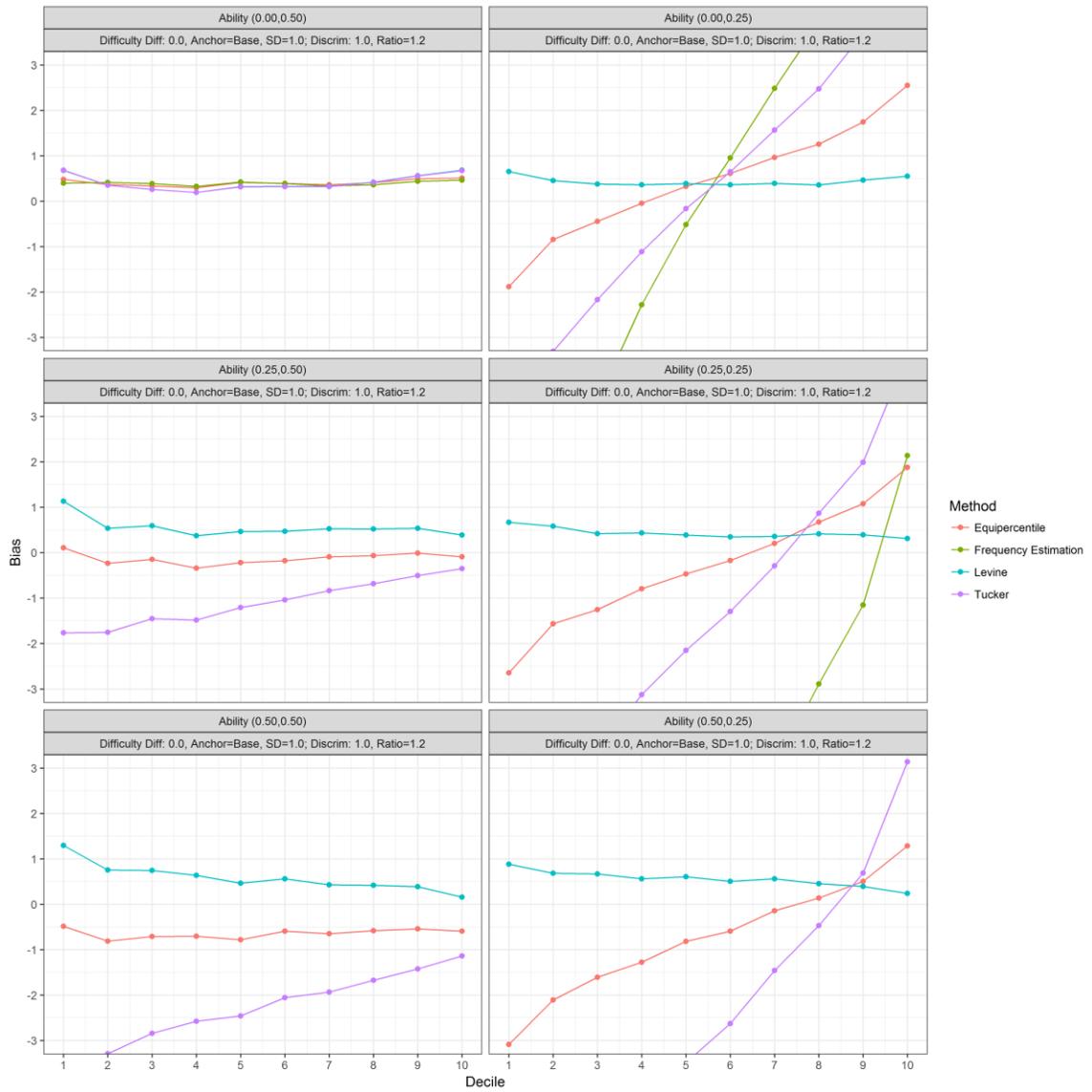


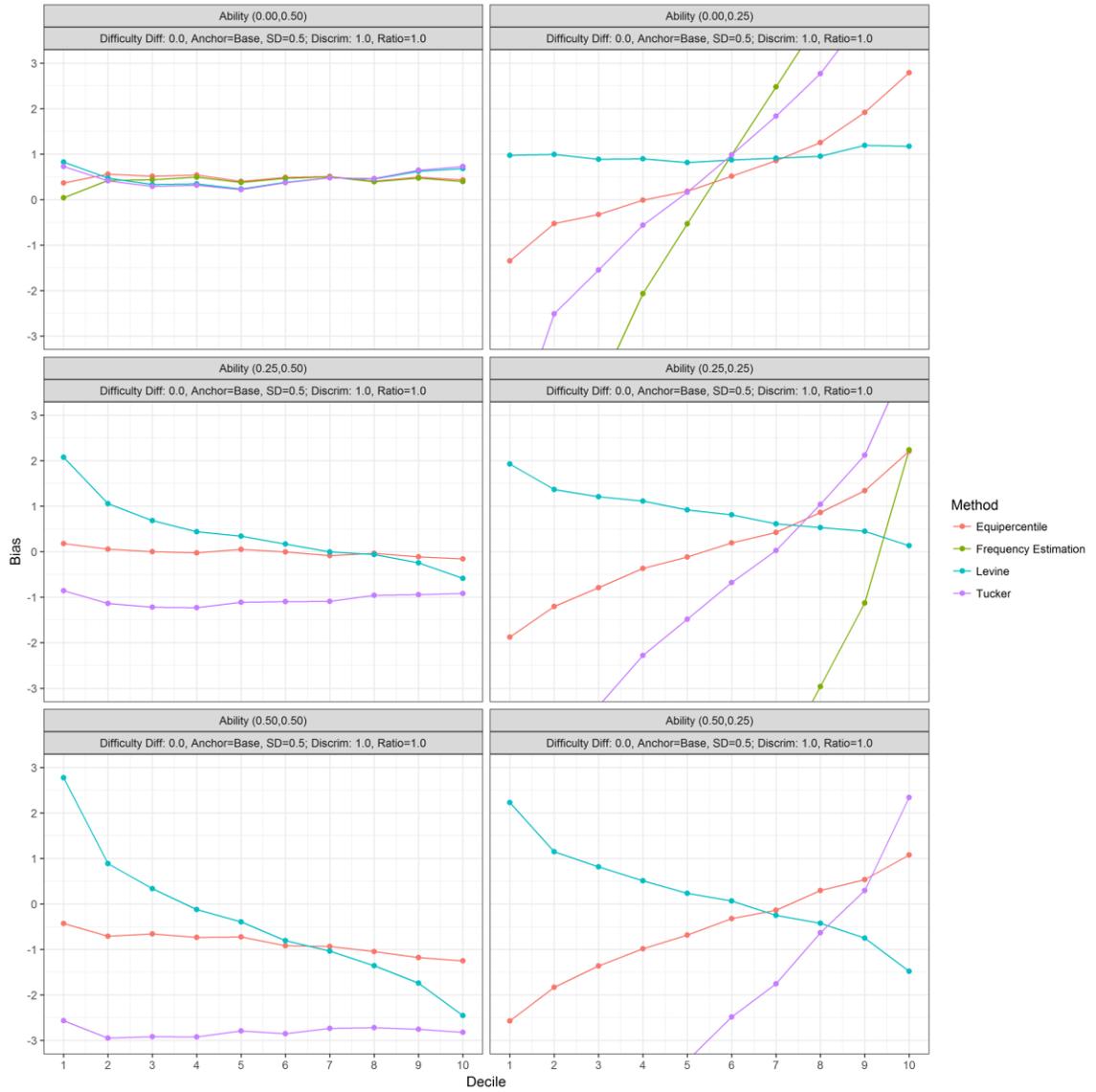


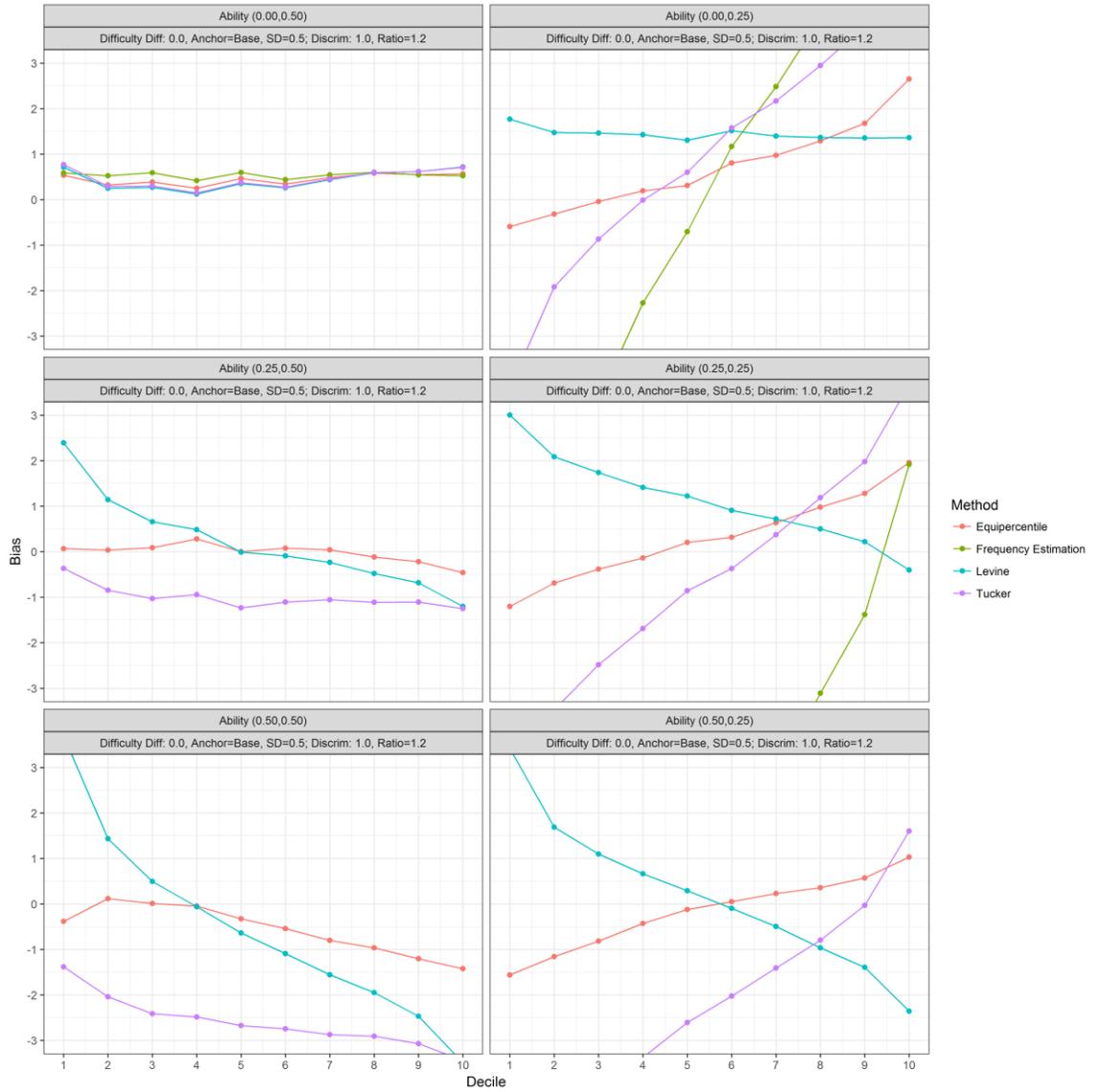


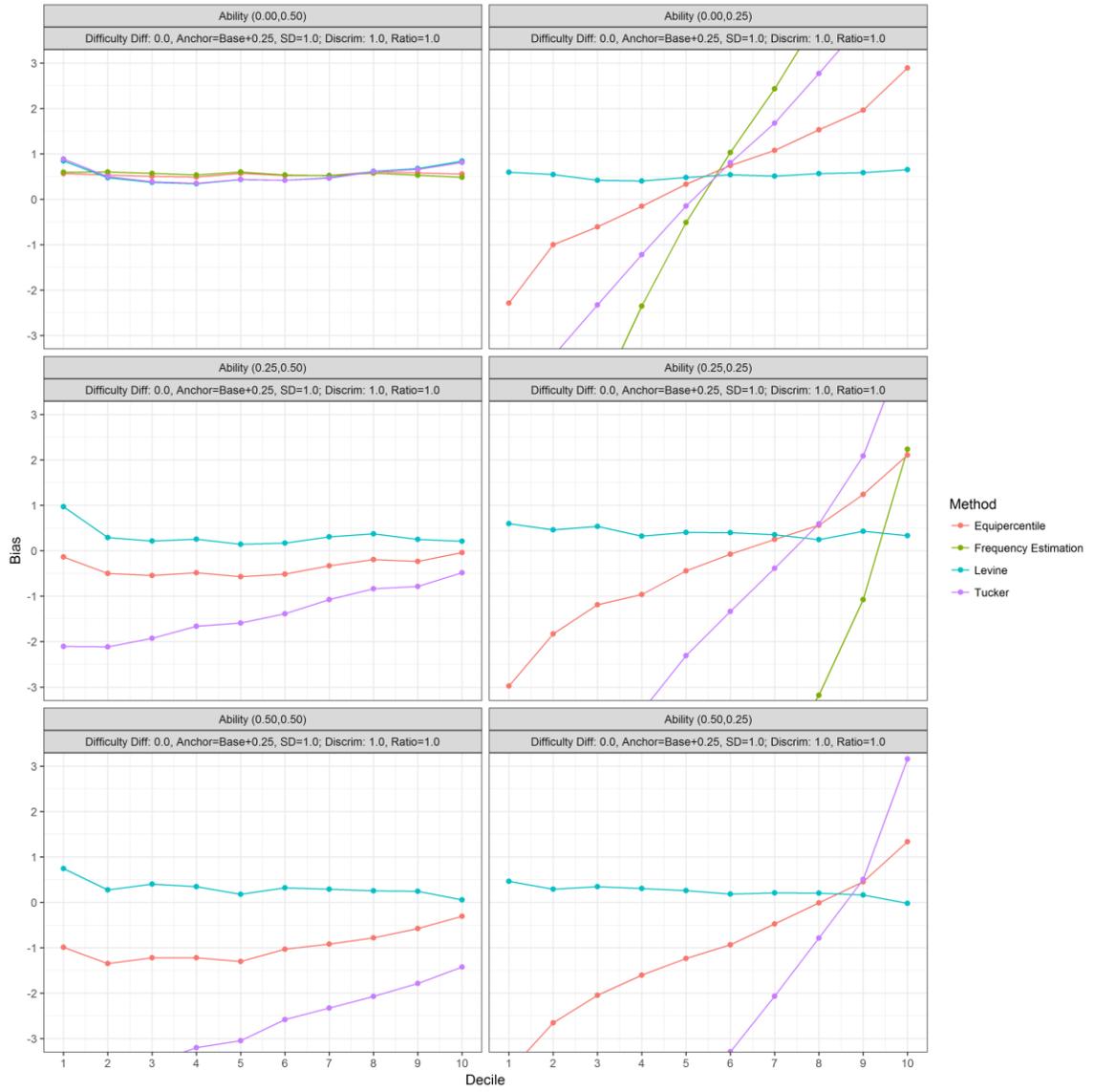


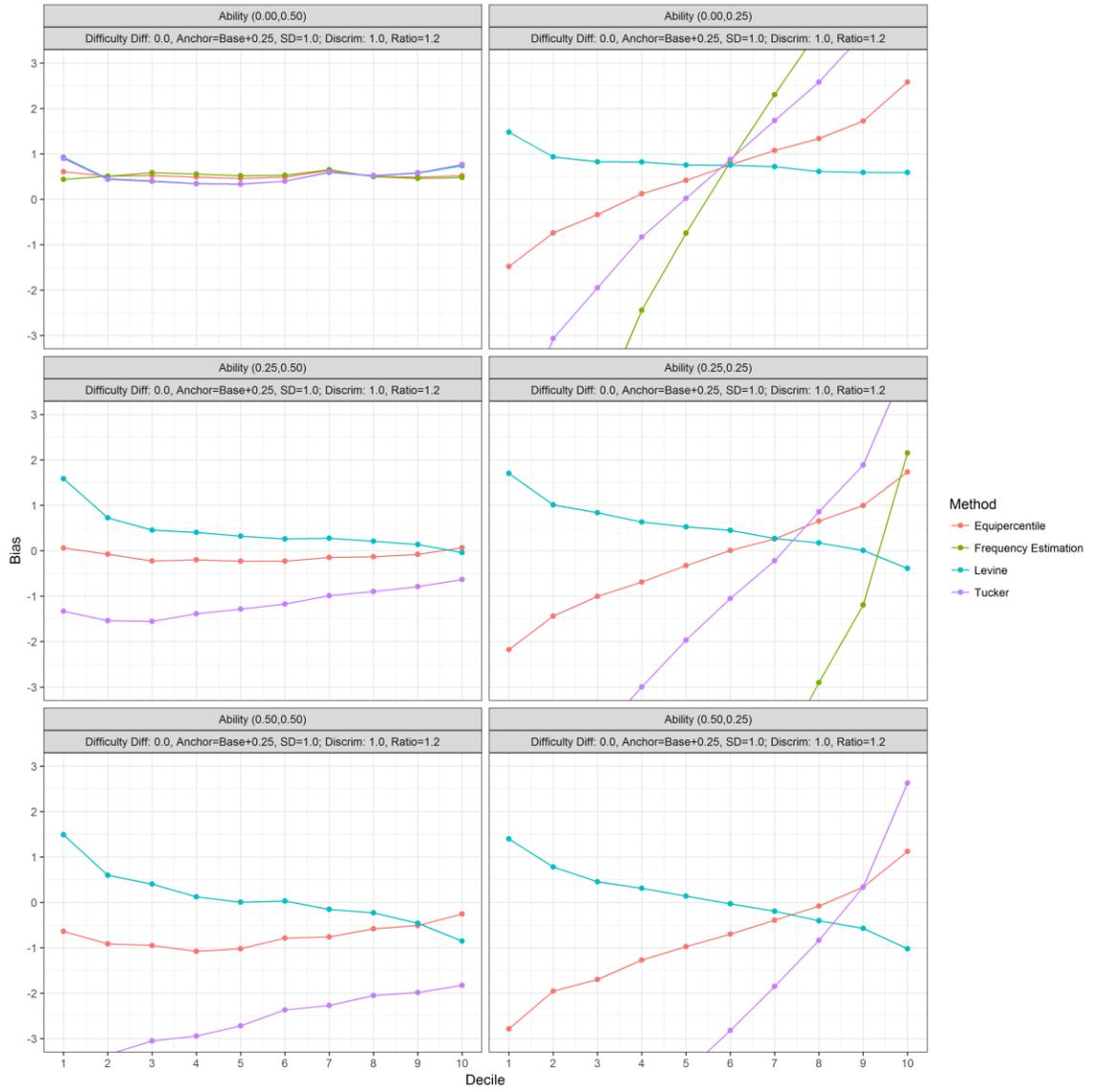


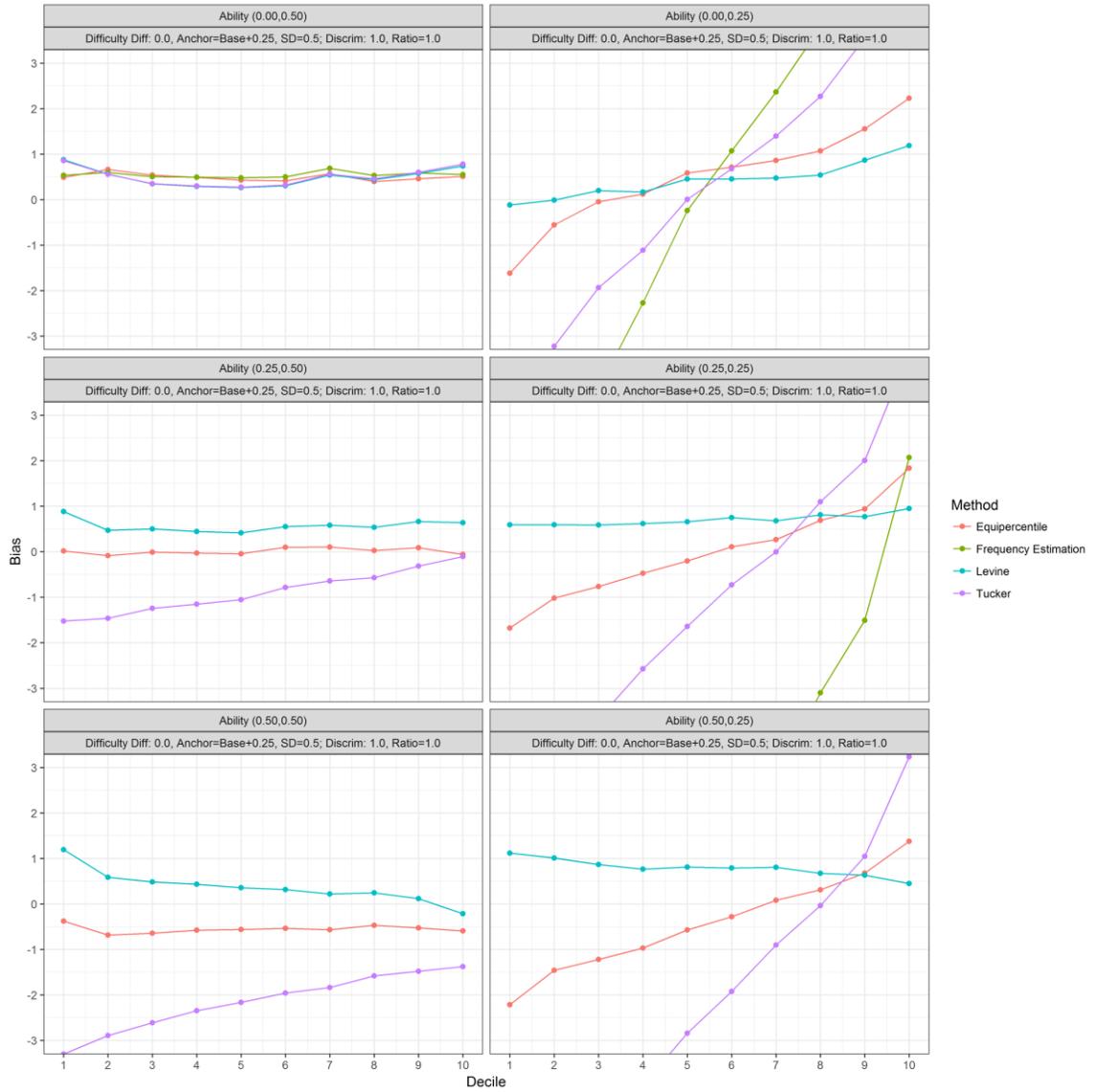


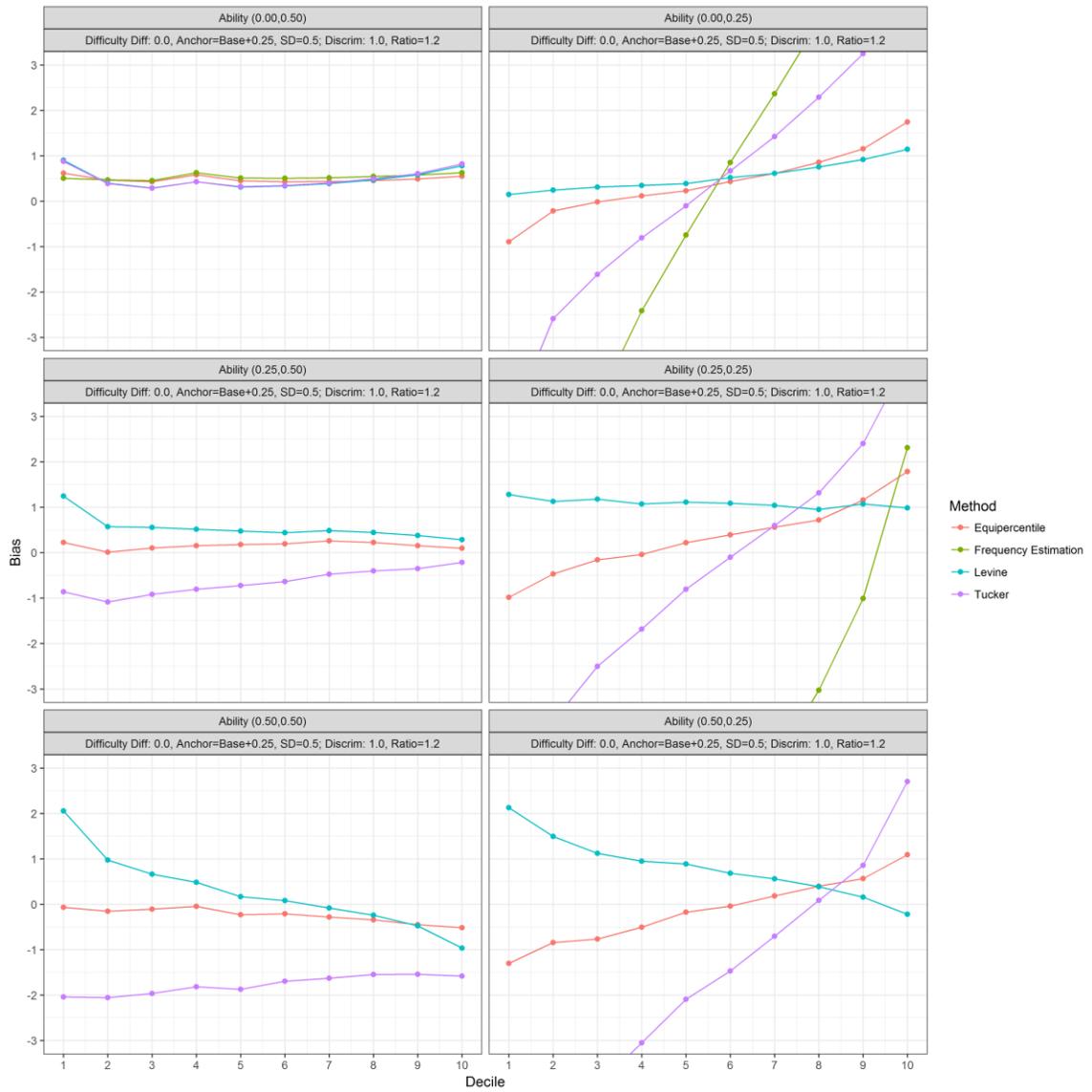


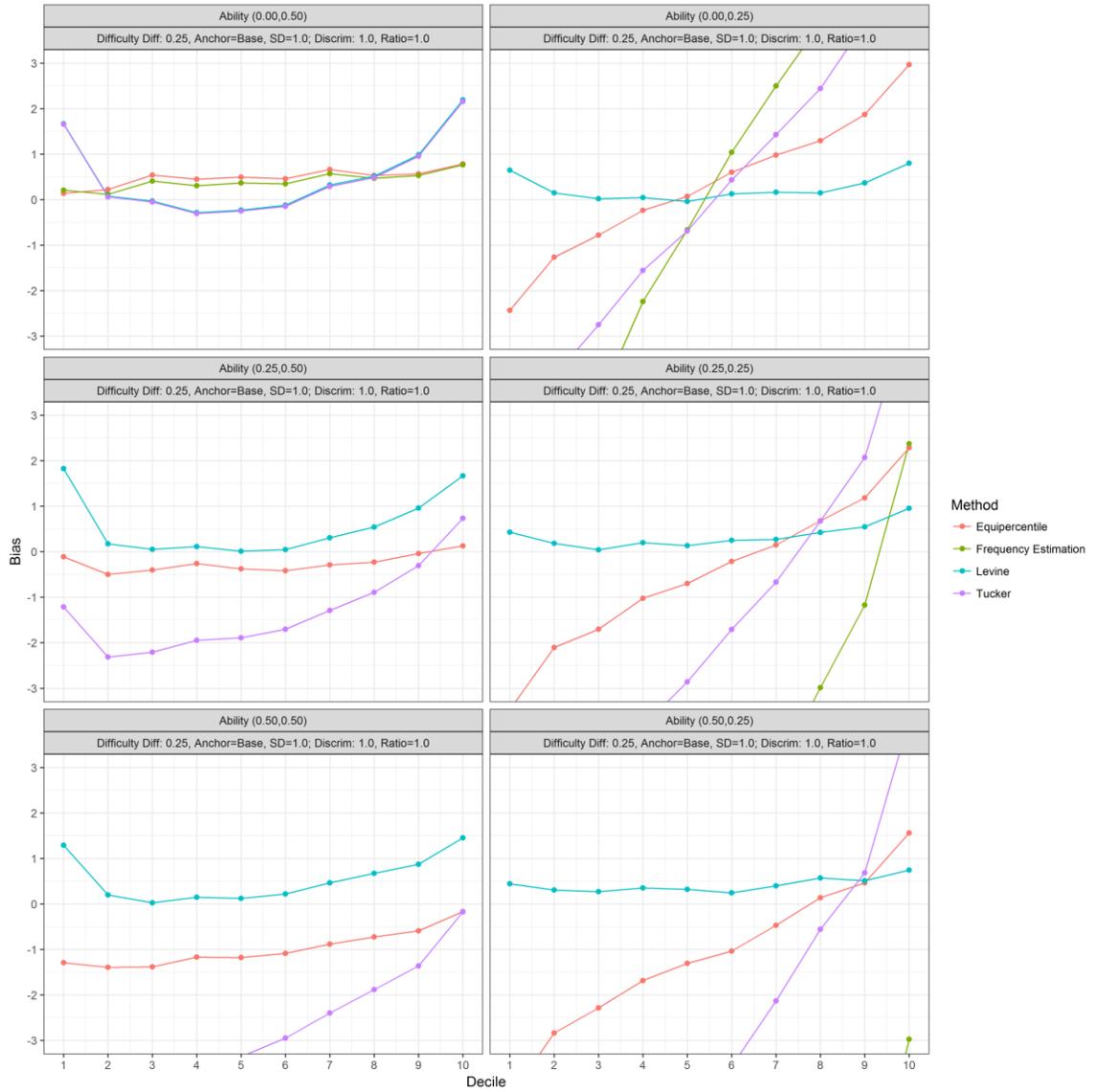


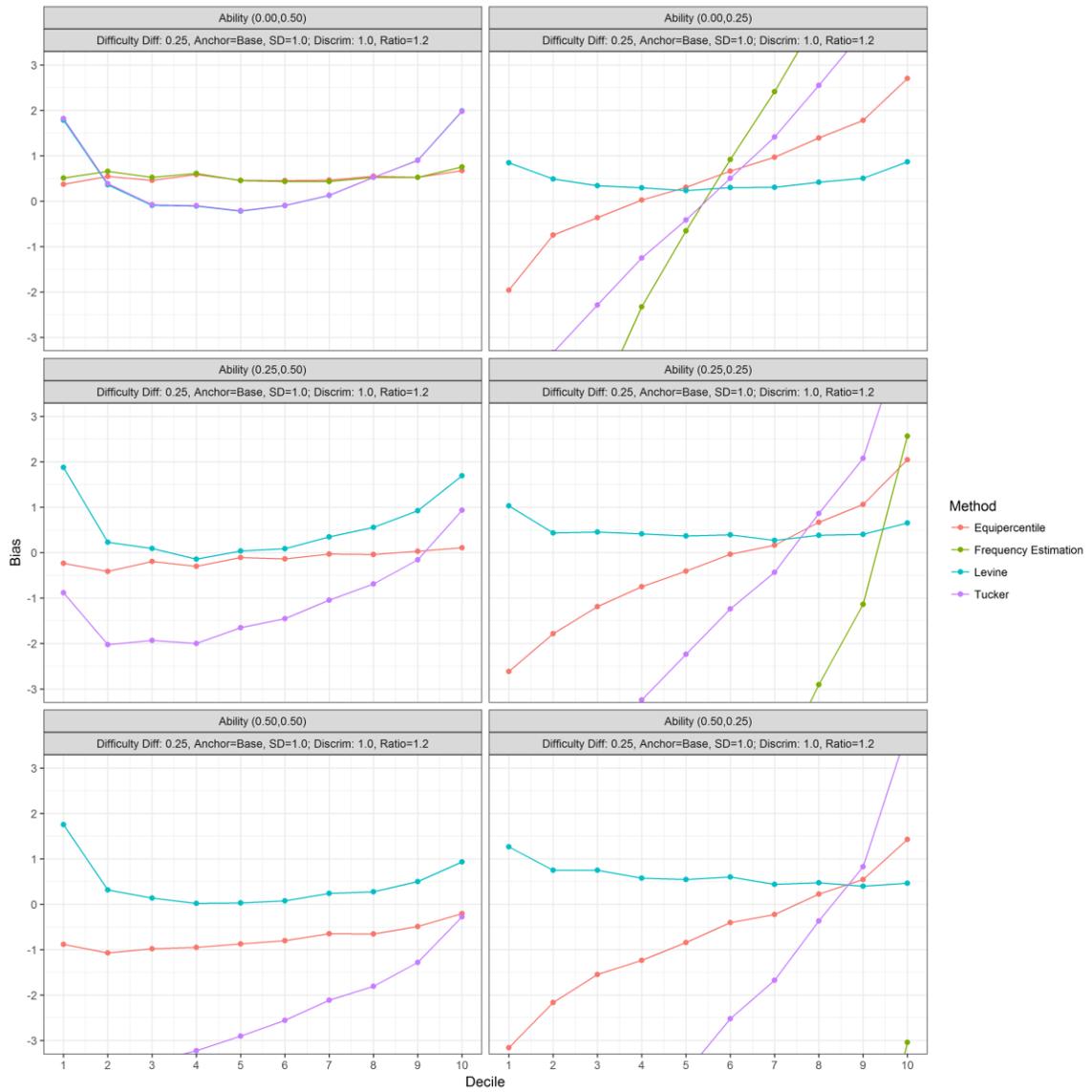


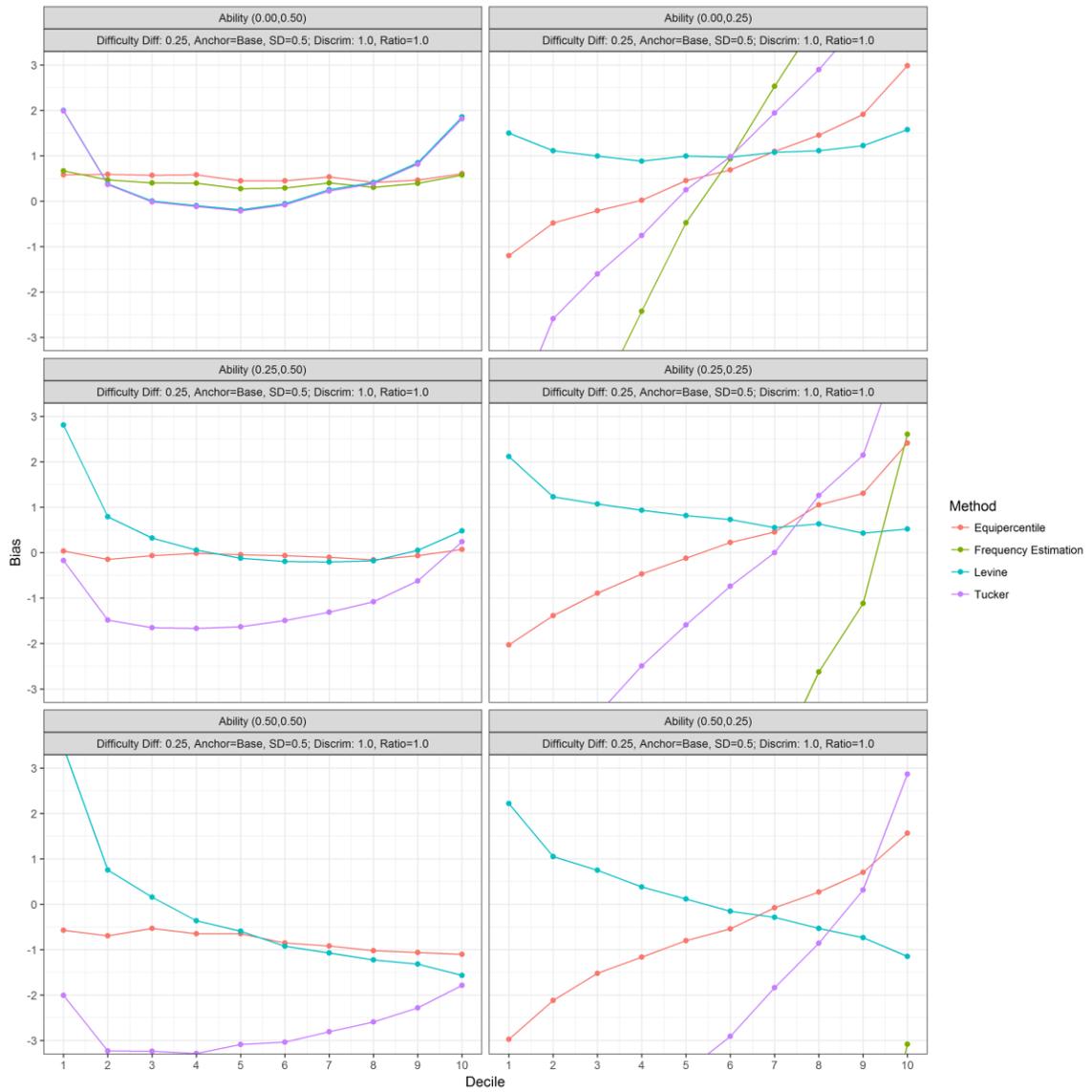


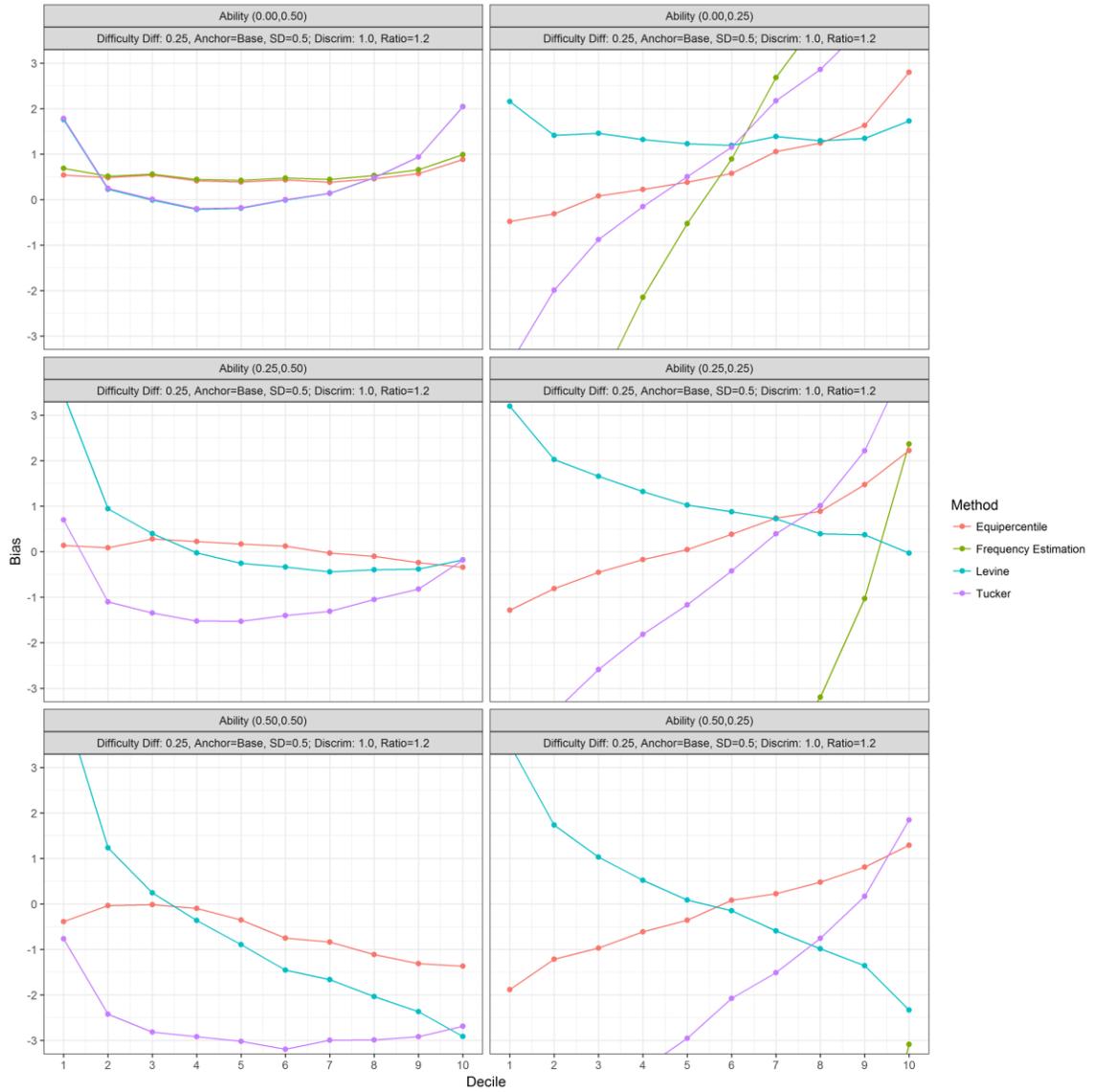


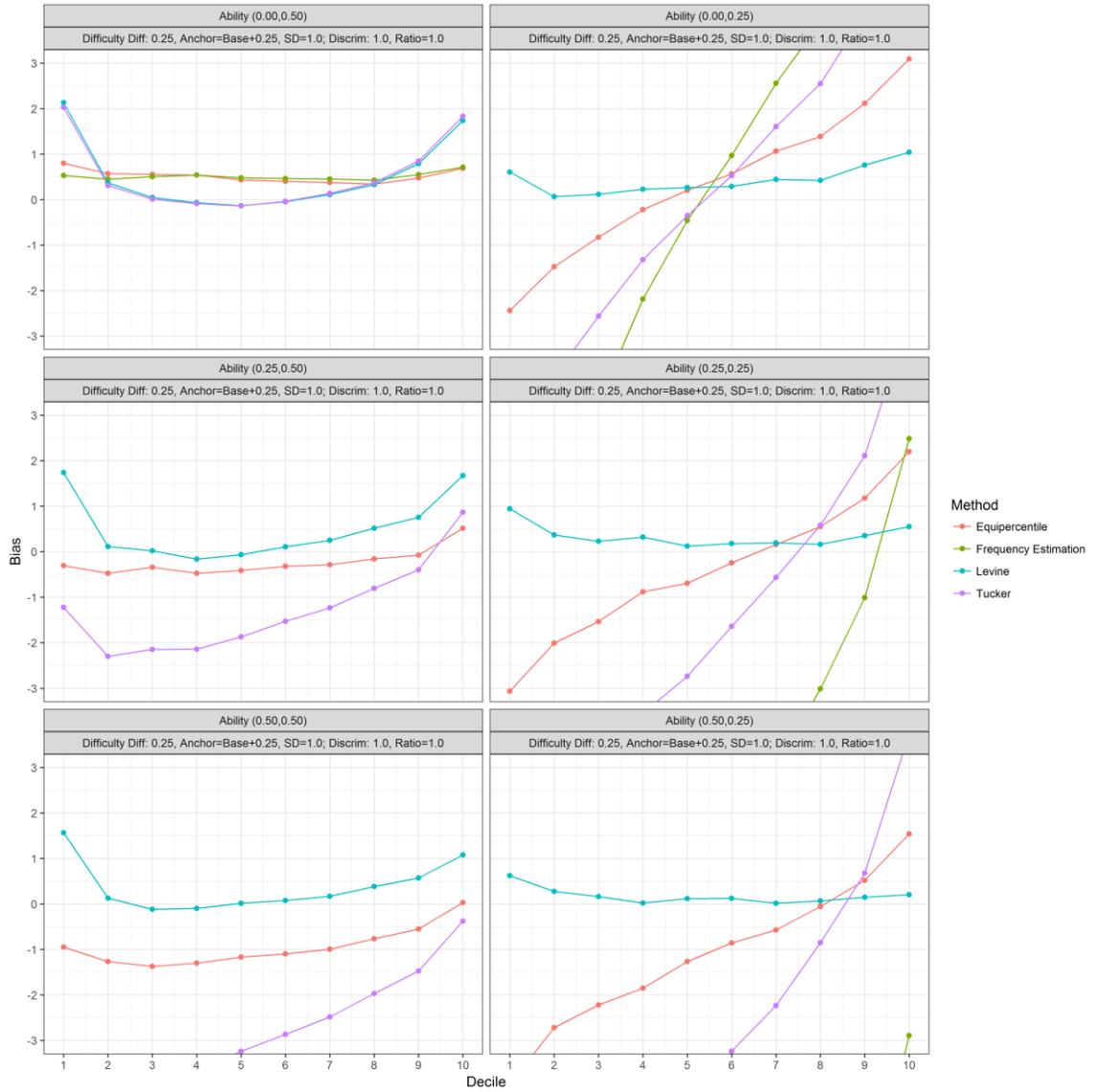


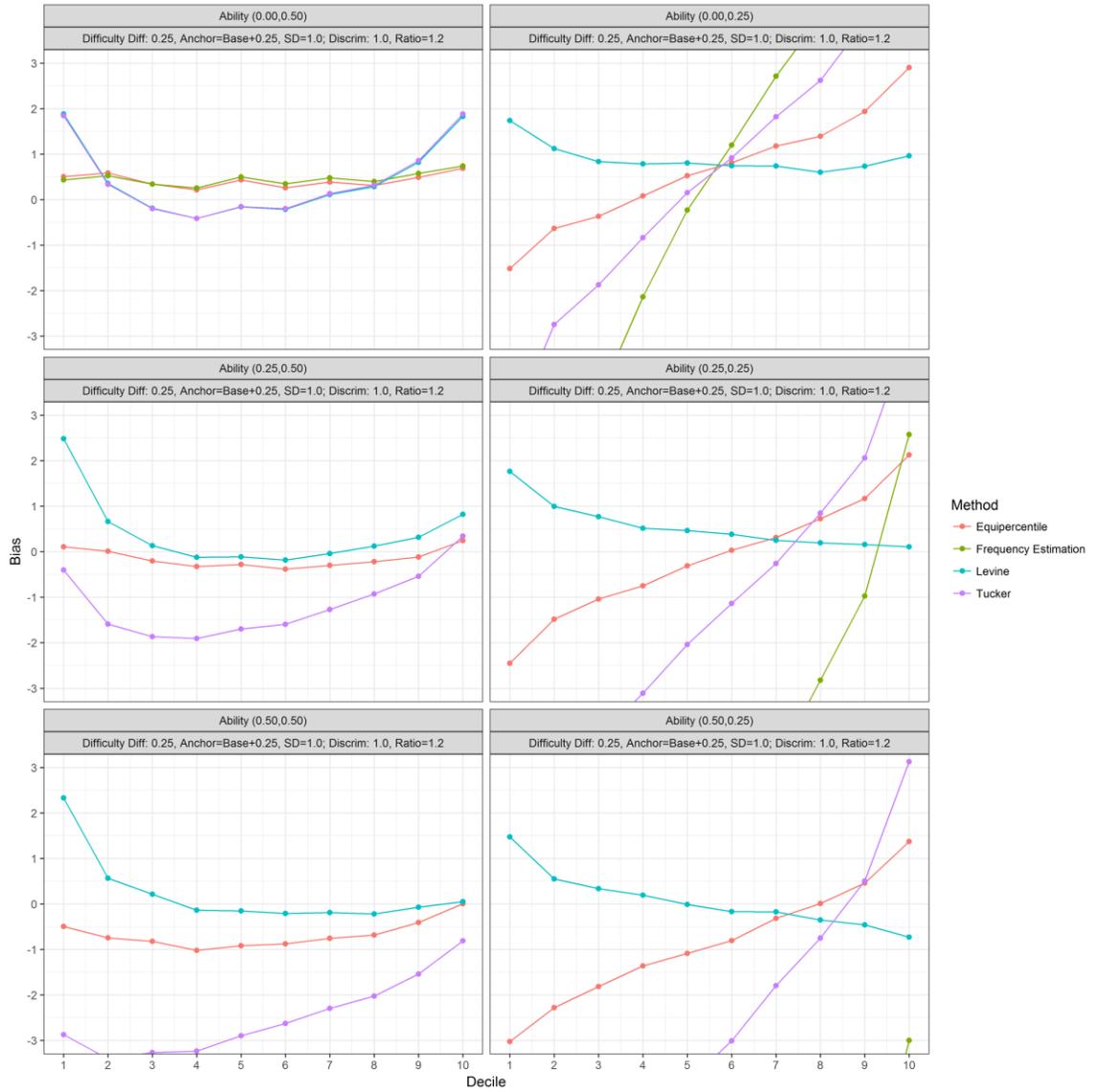


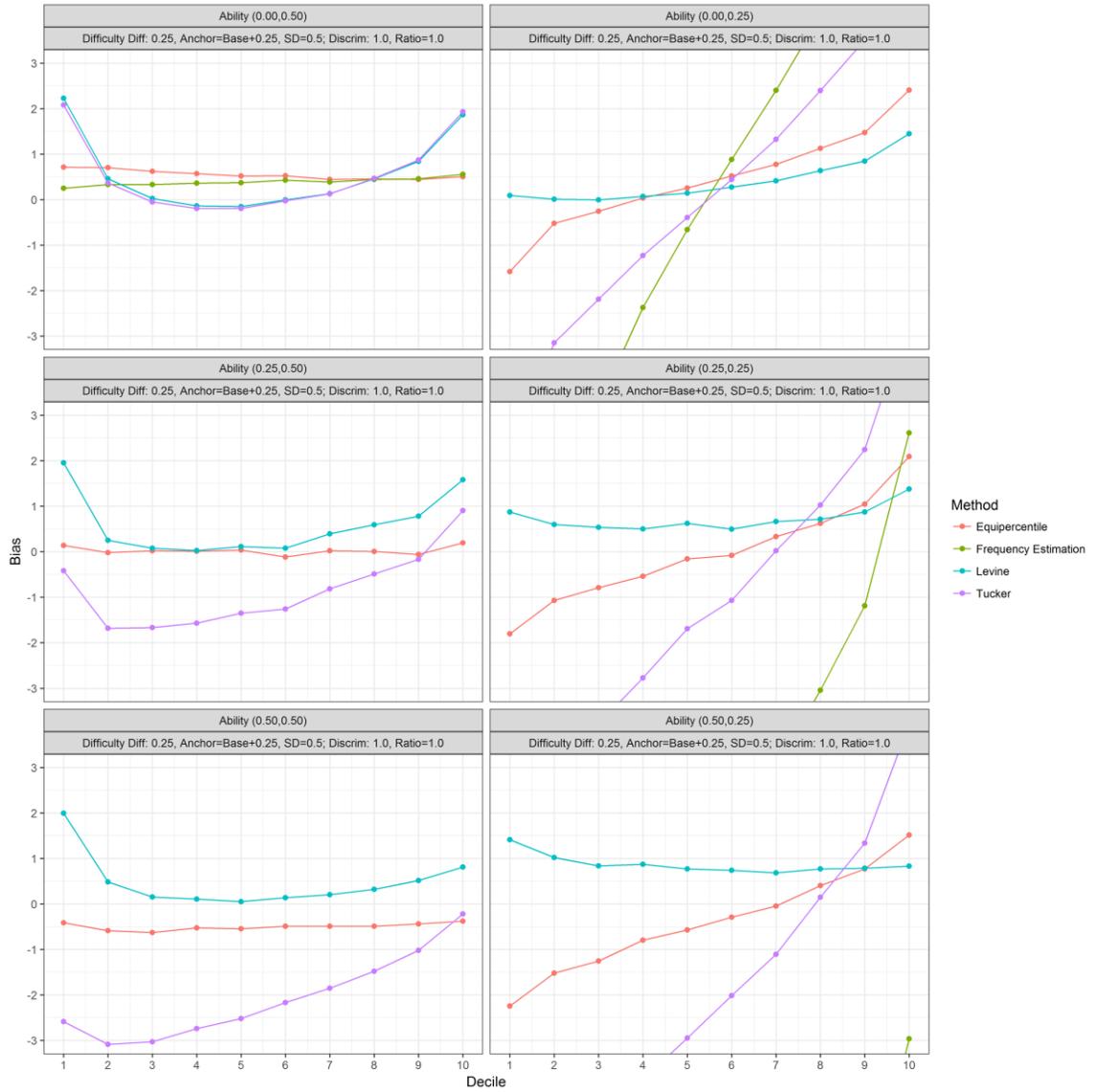


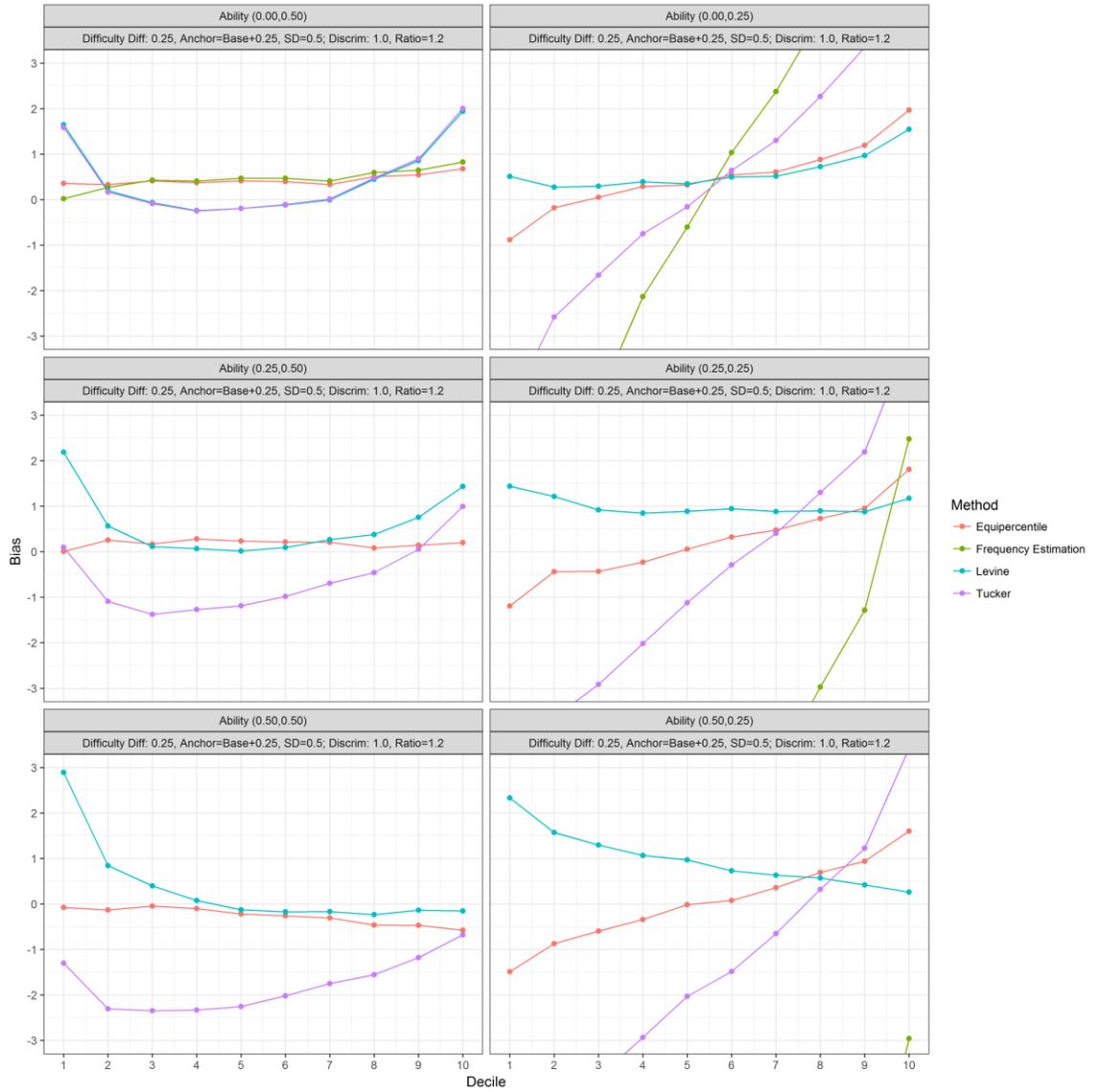


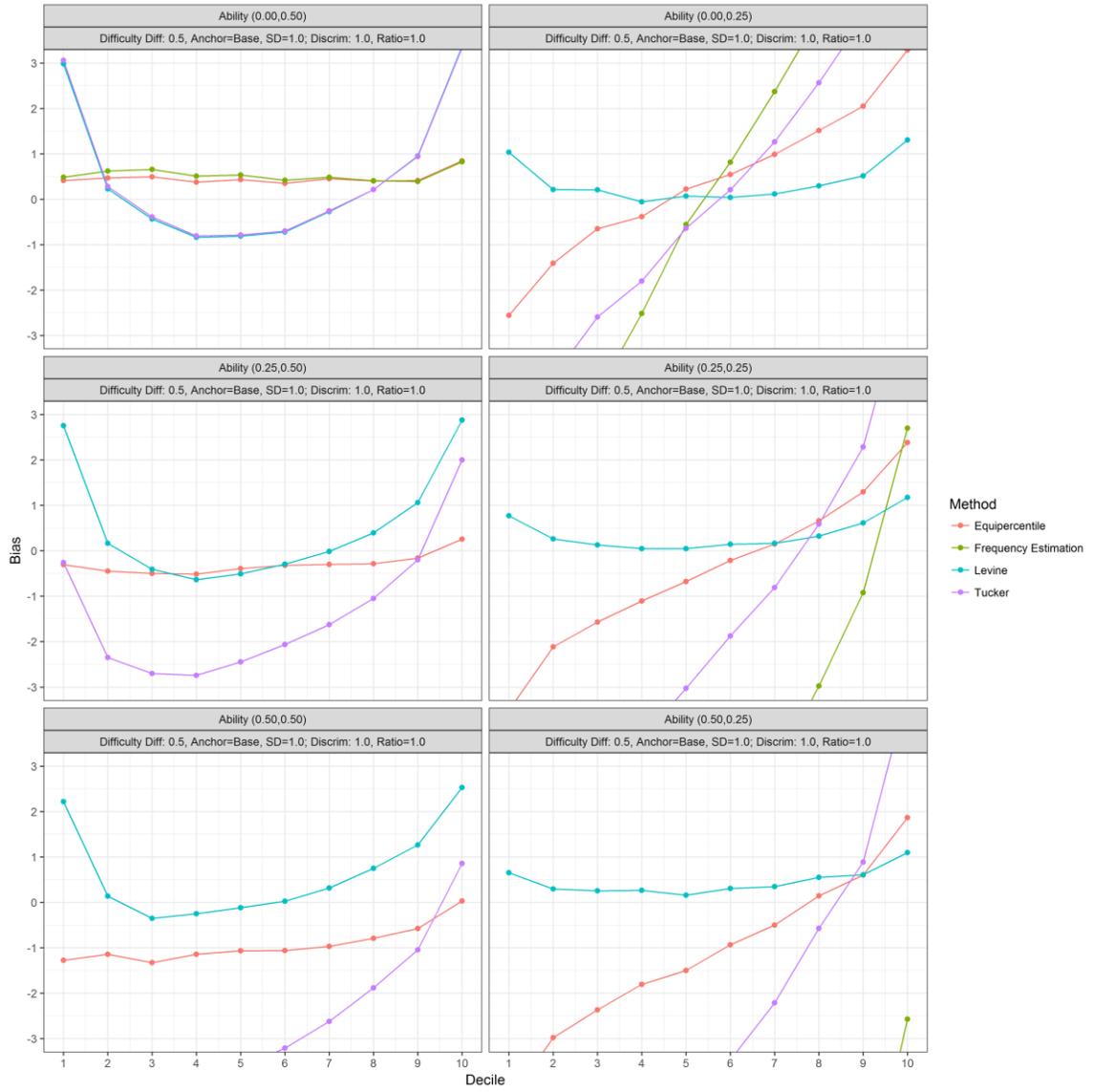


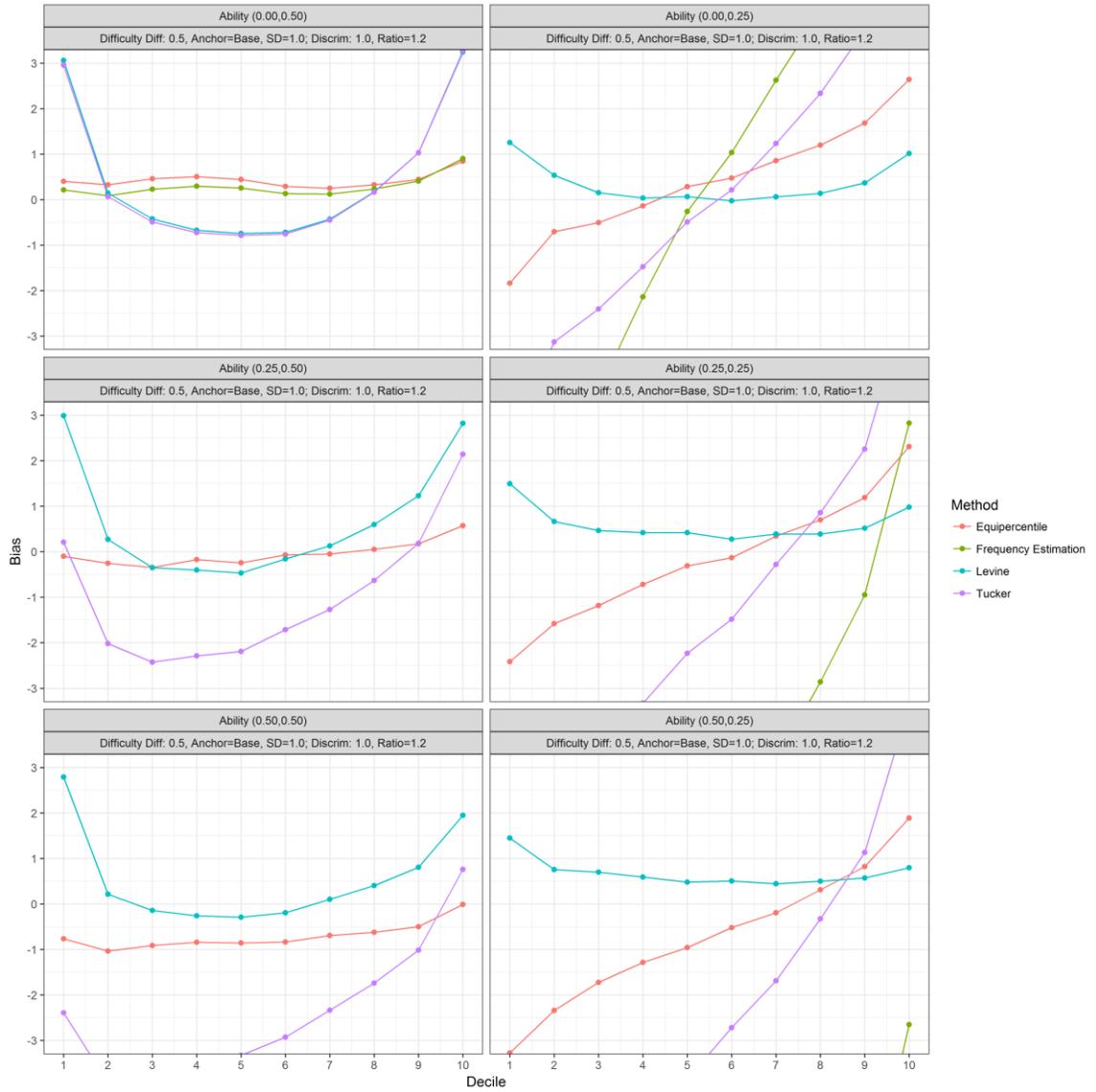


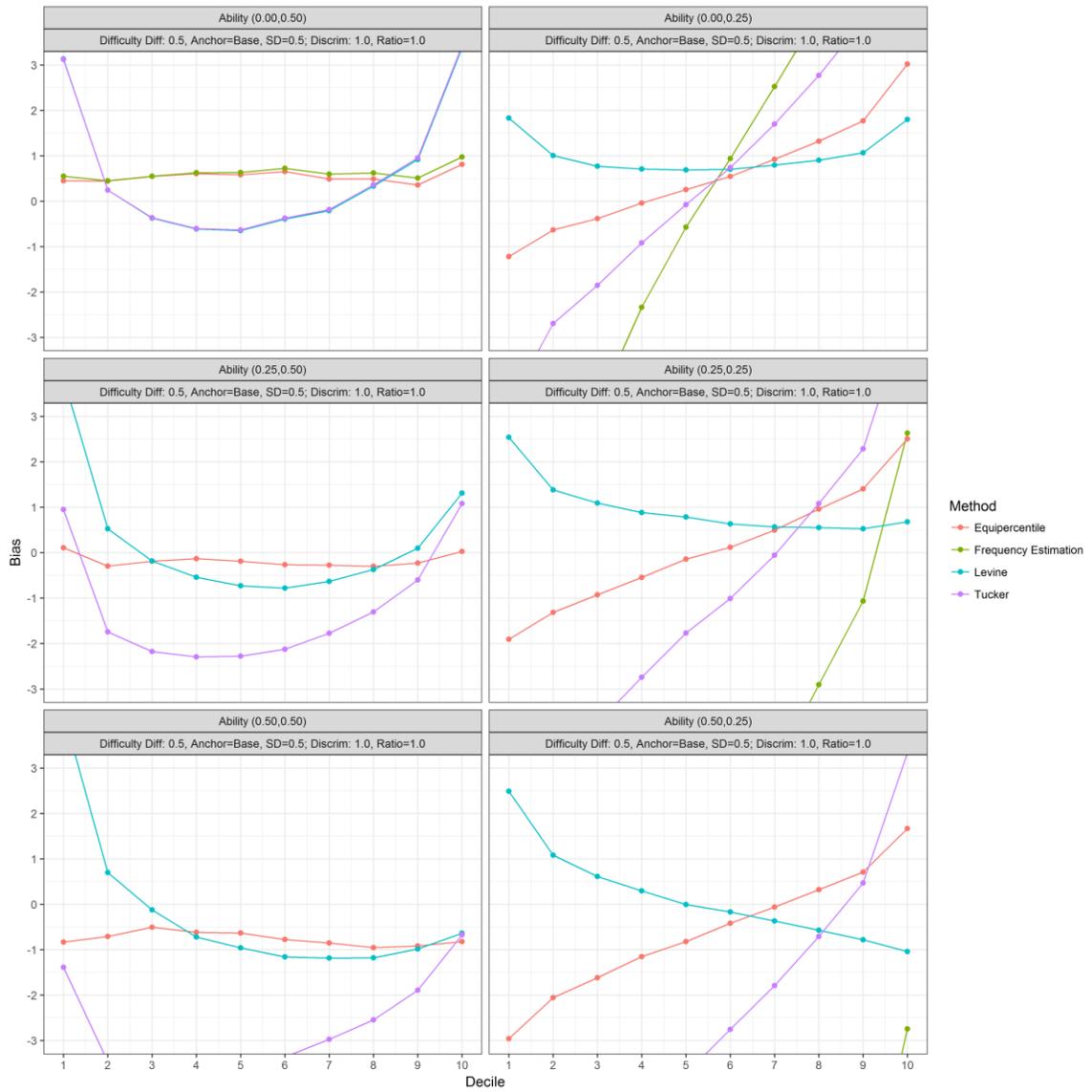


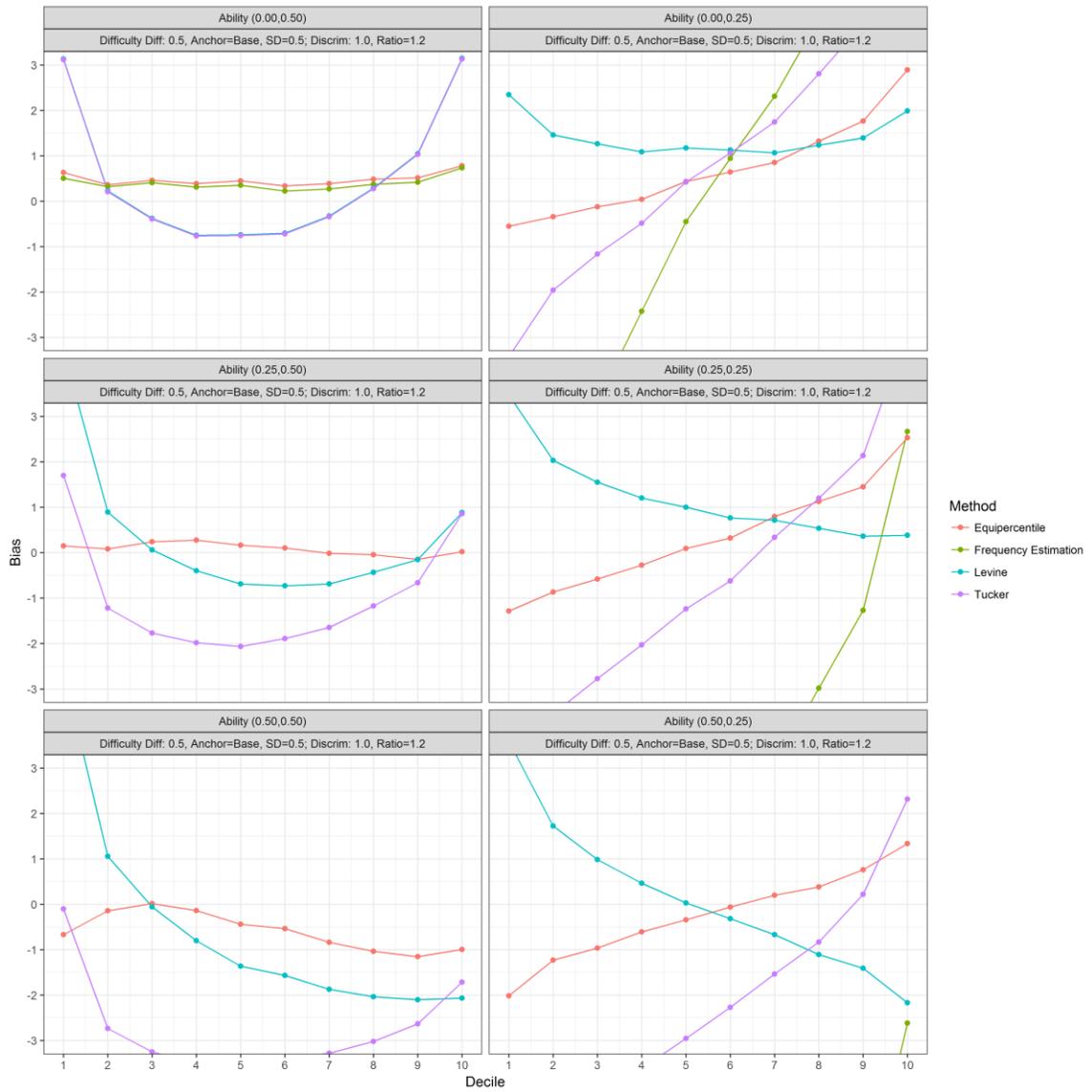


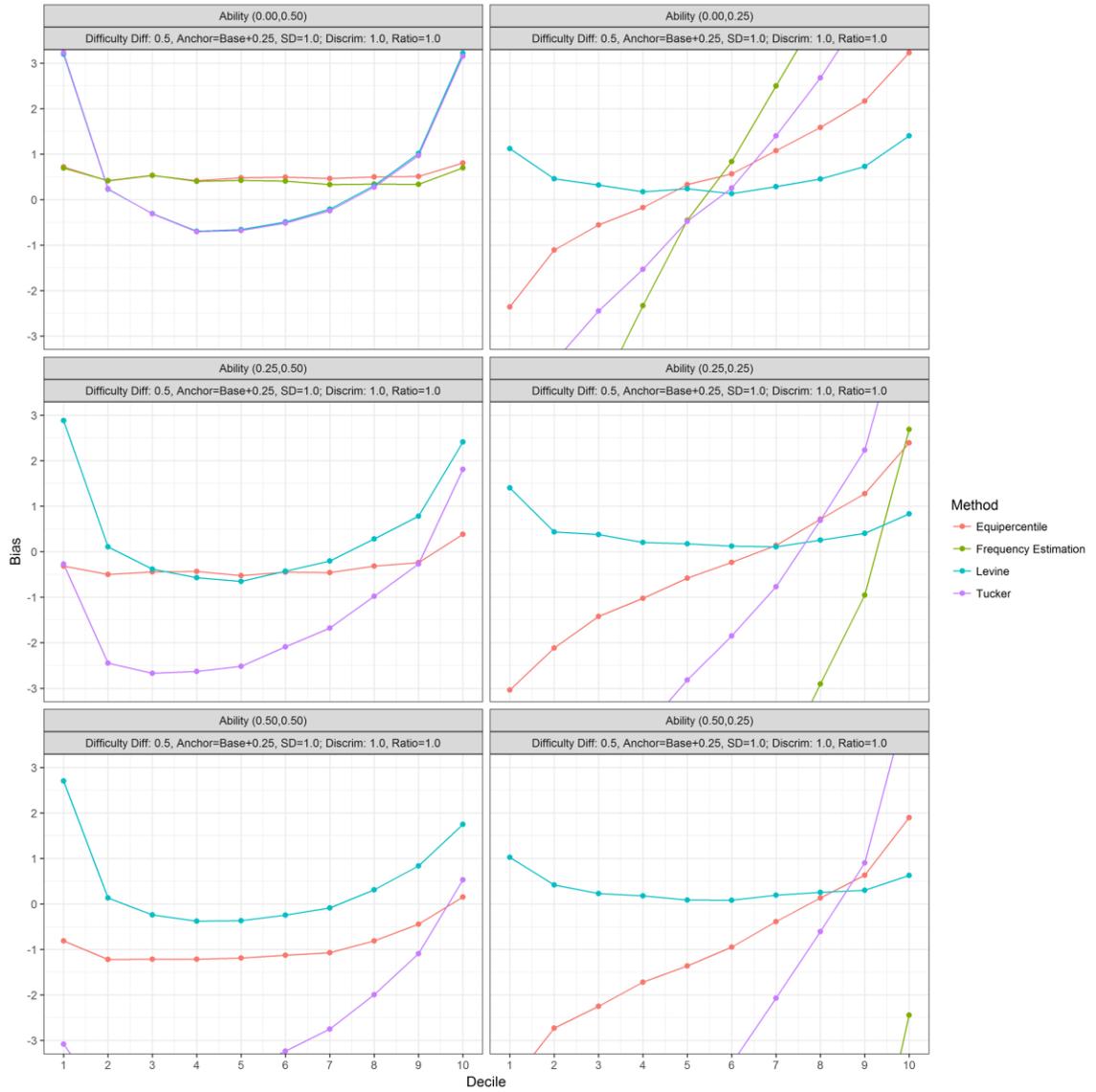


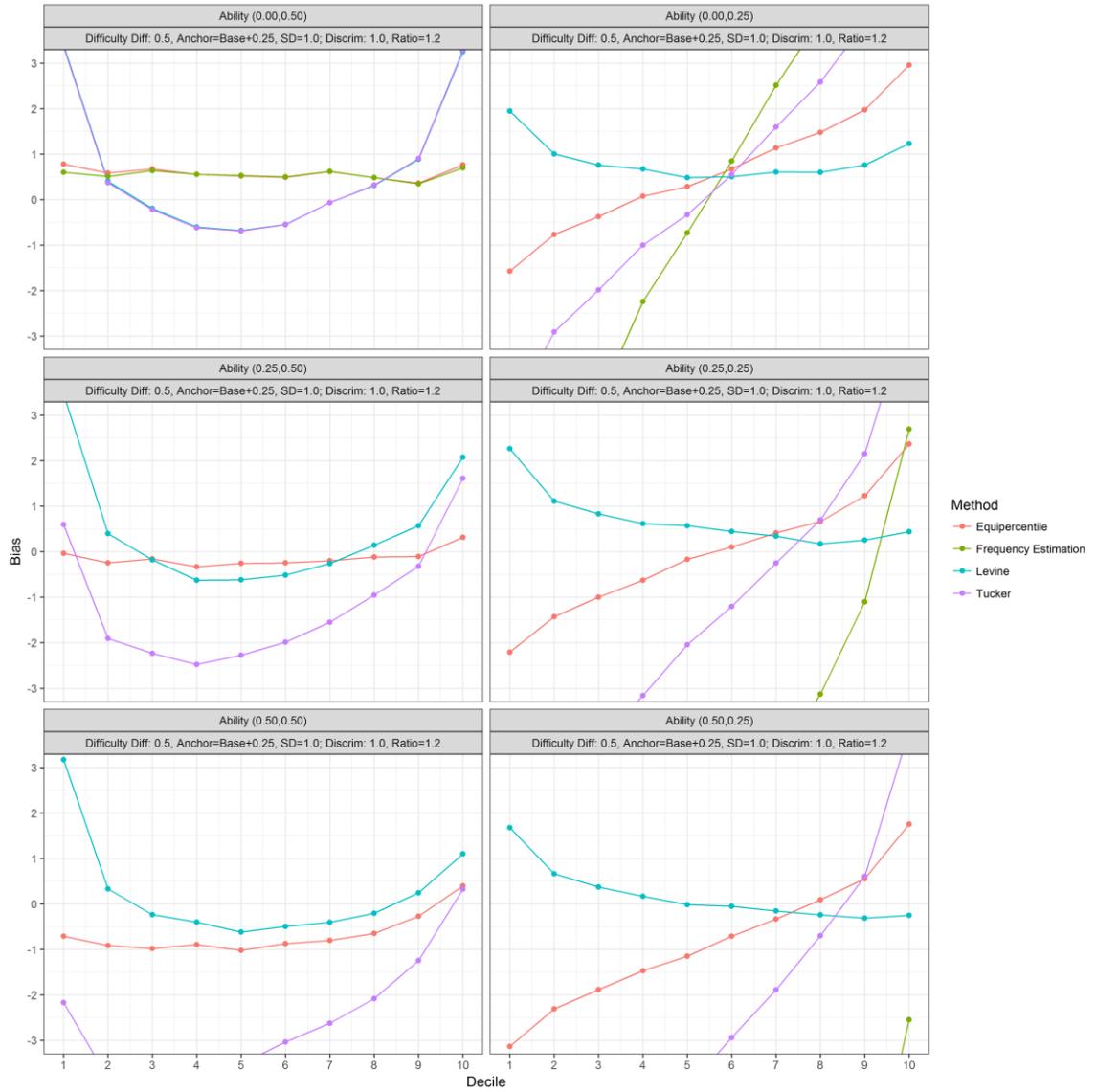


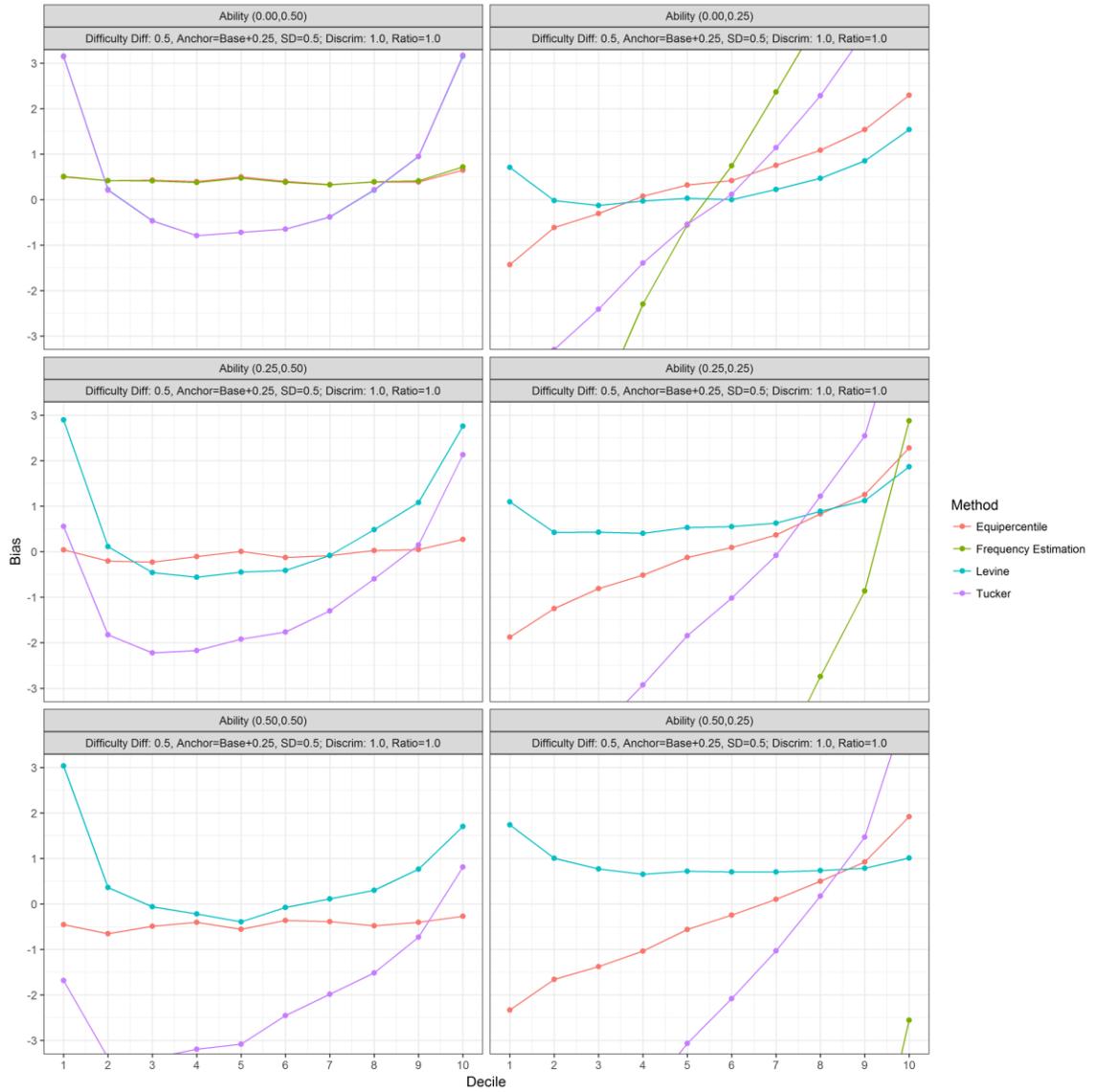


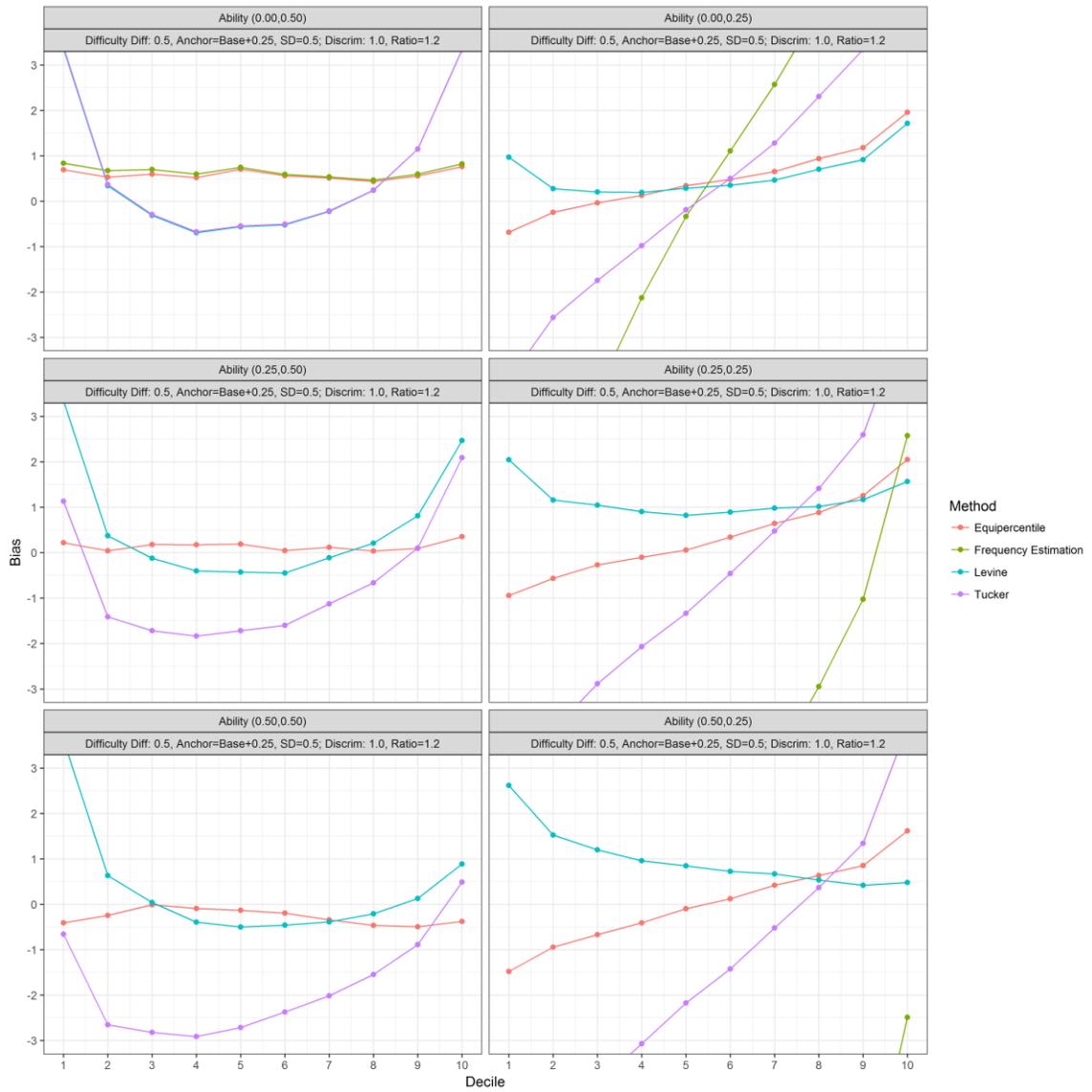




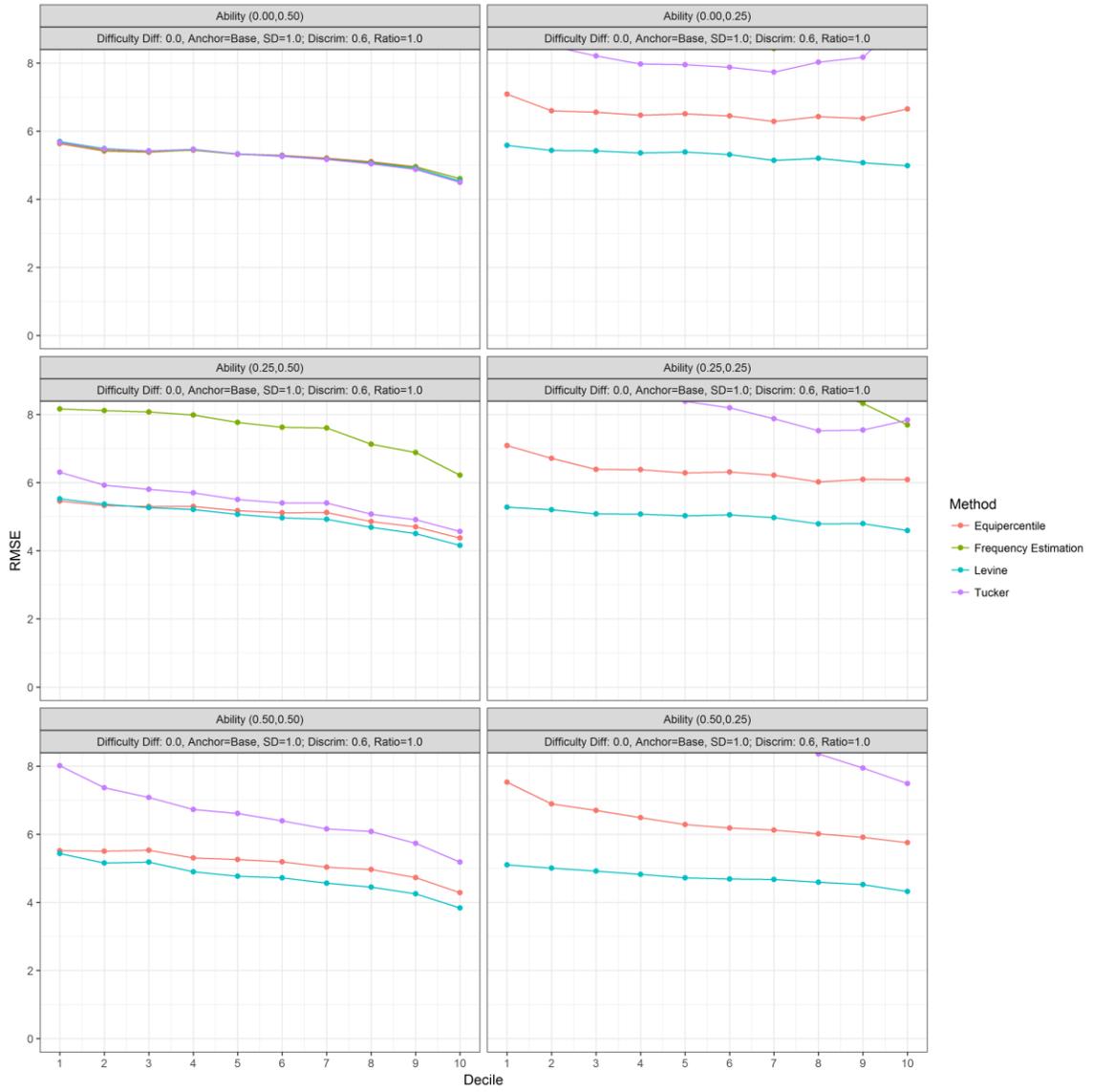


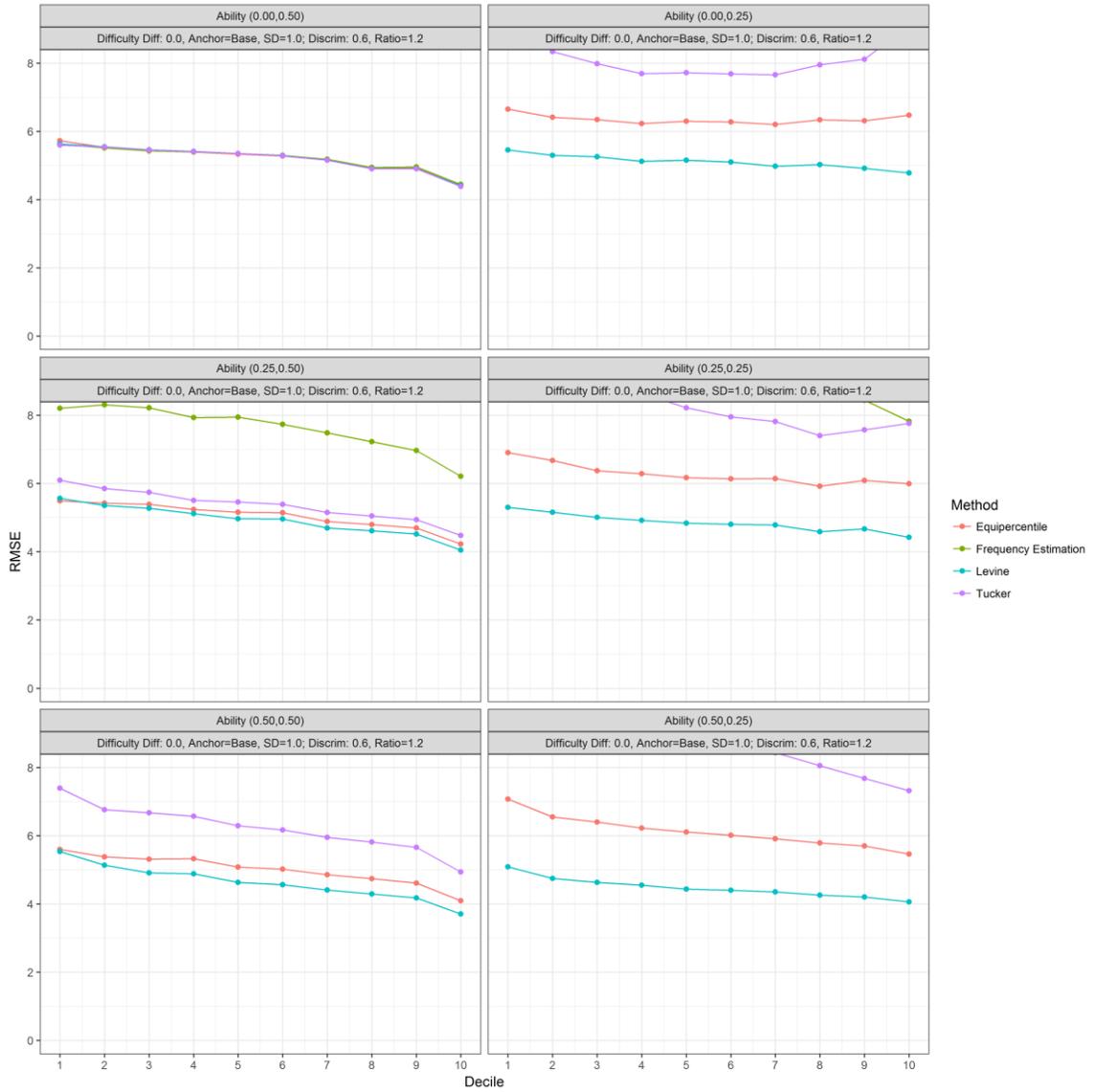


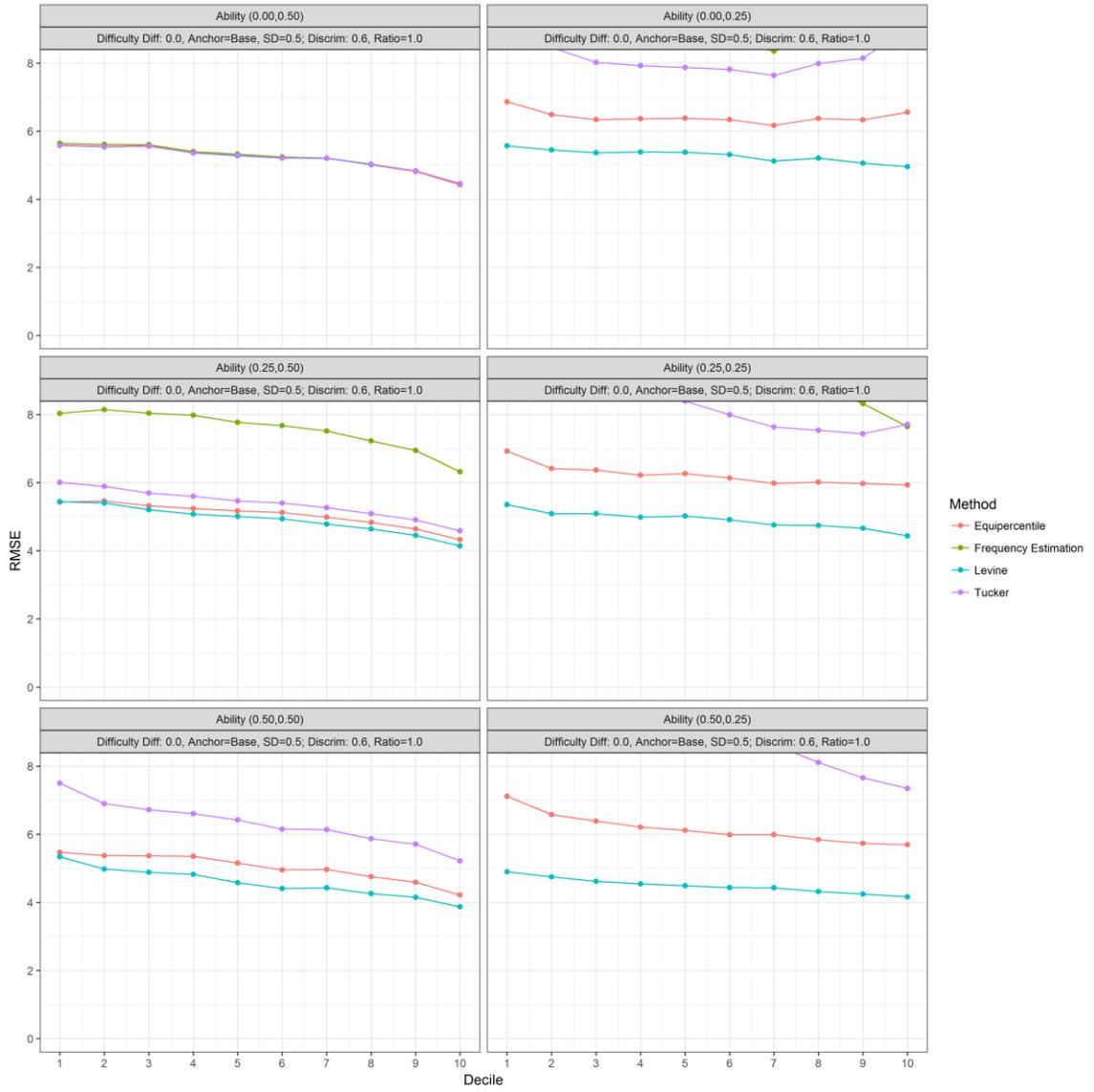


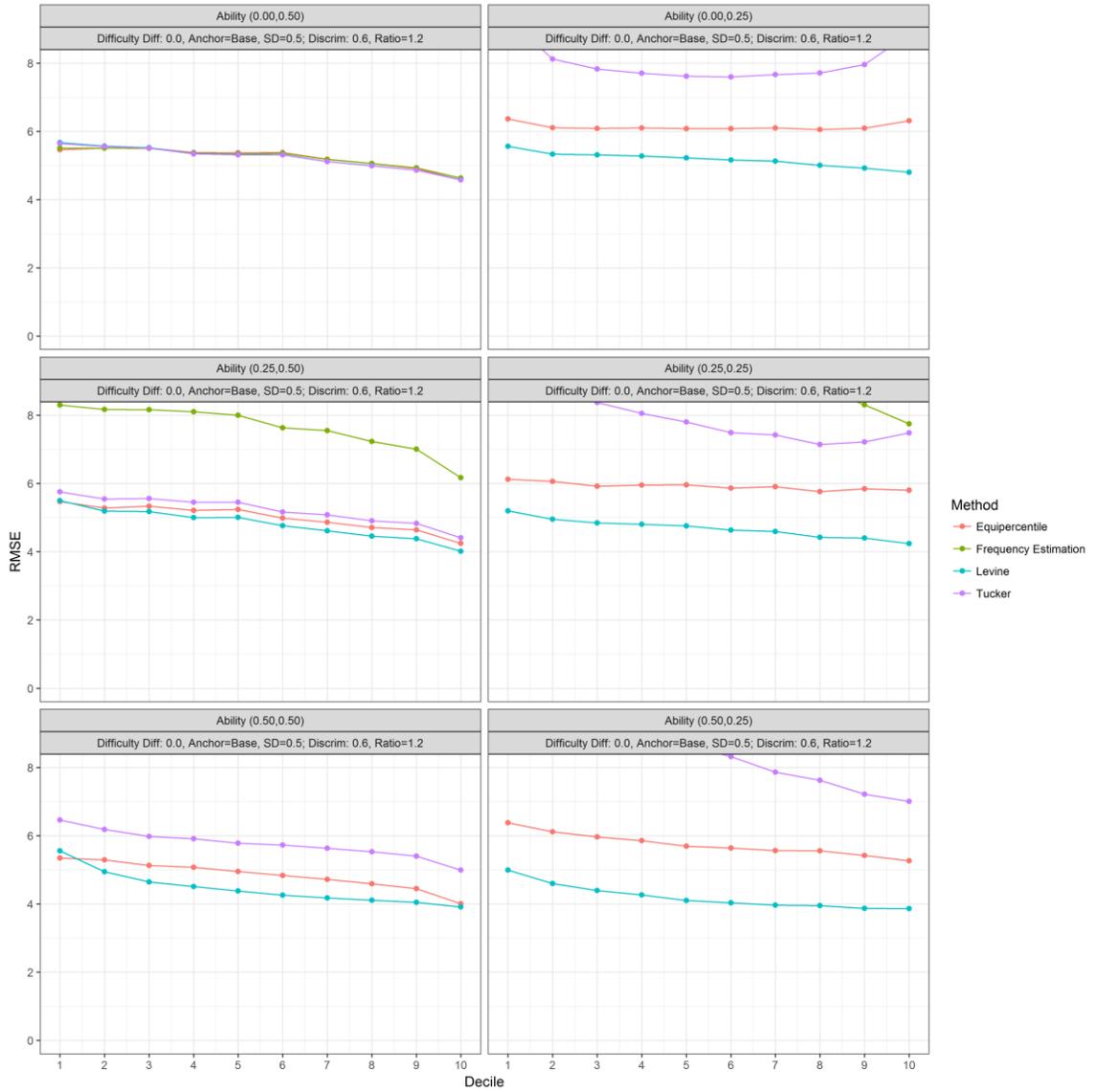


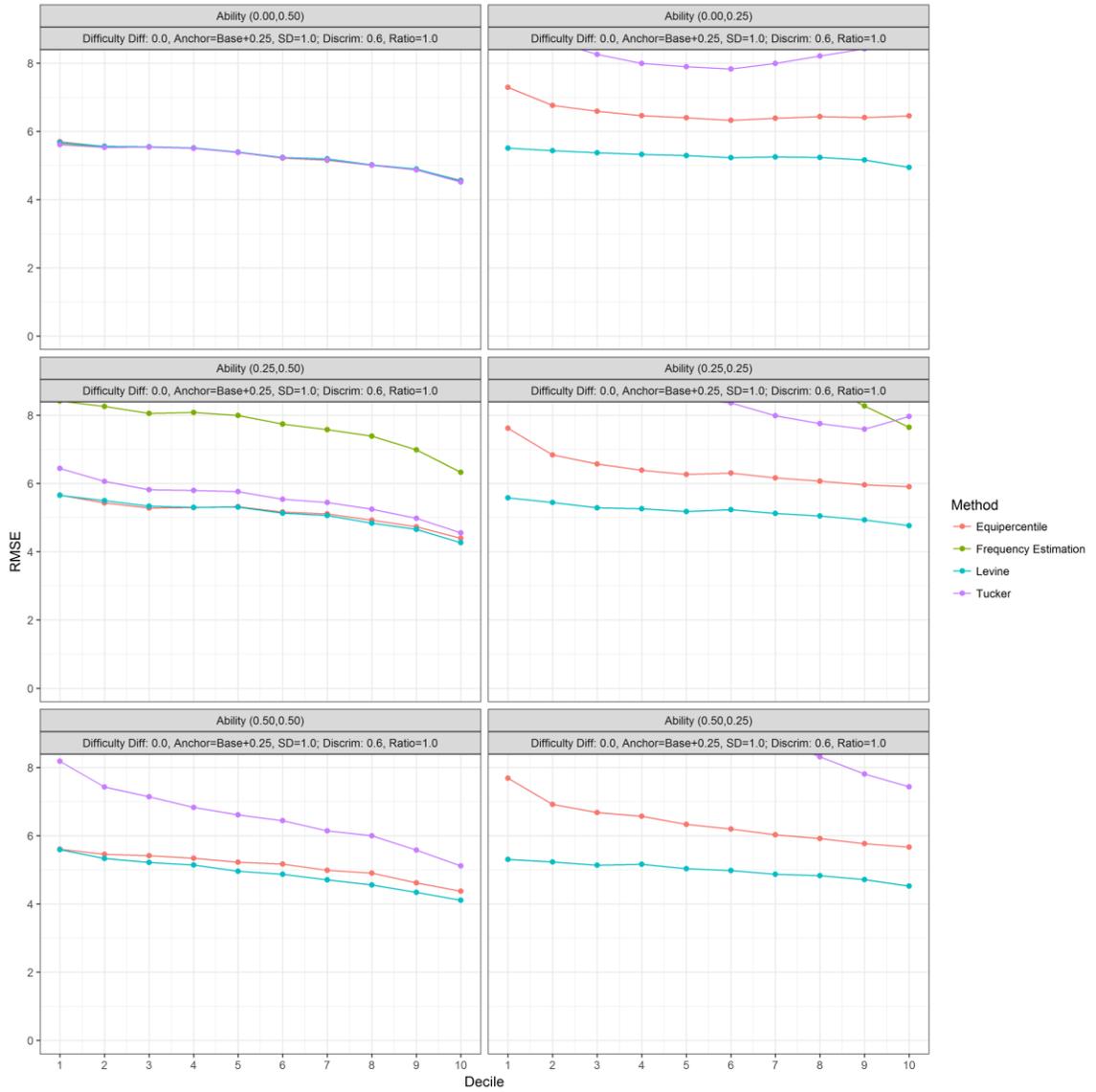
APPENDIX D  
CERTIFICATION RMSE RESULTS

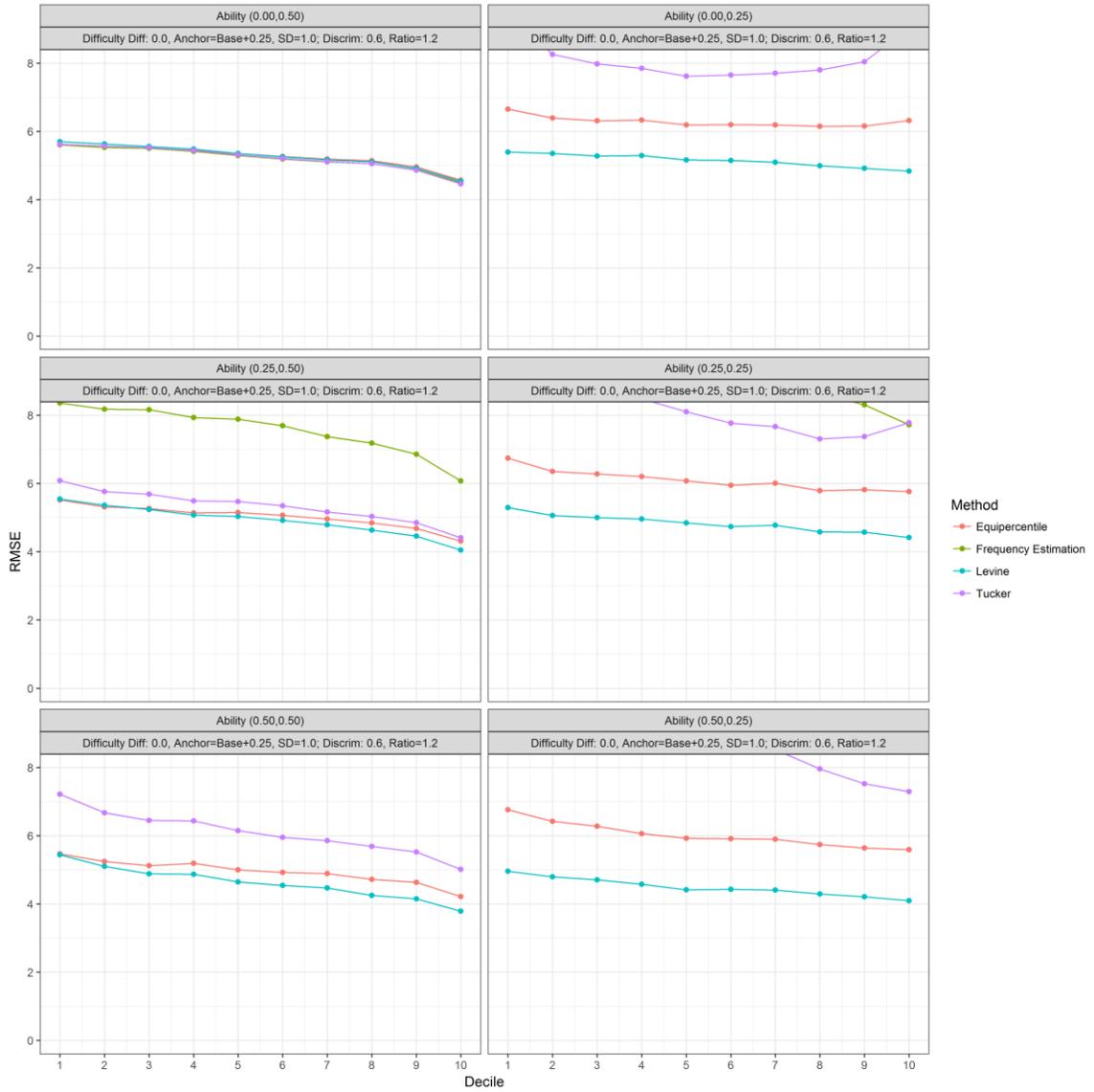


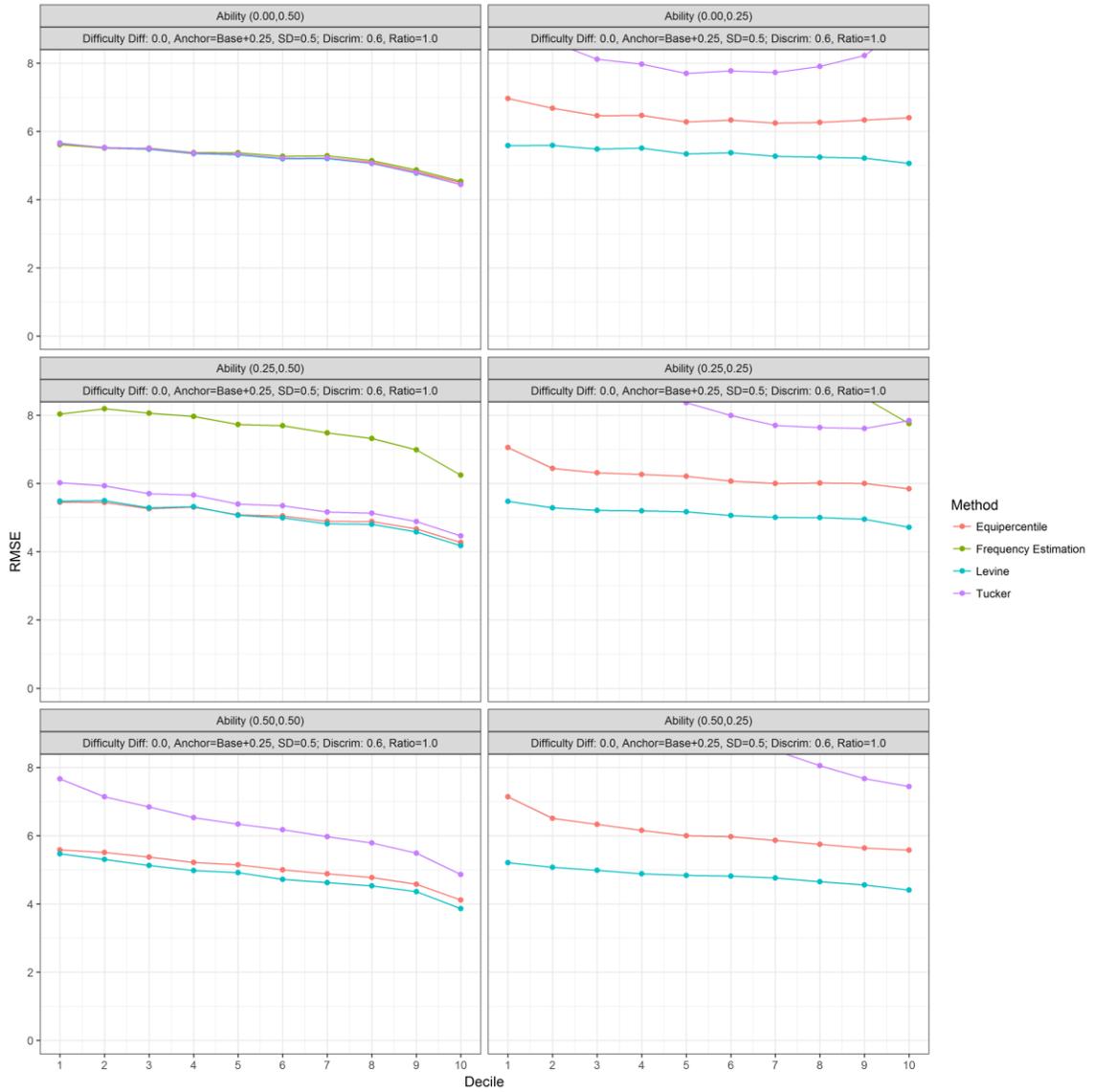


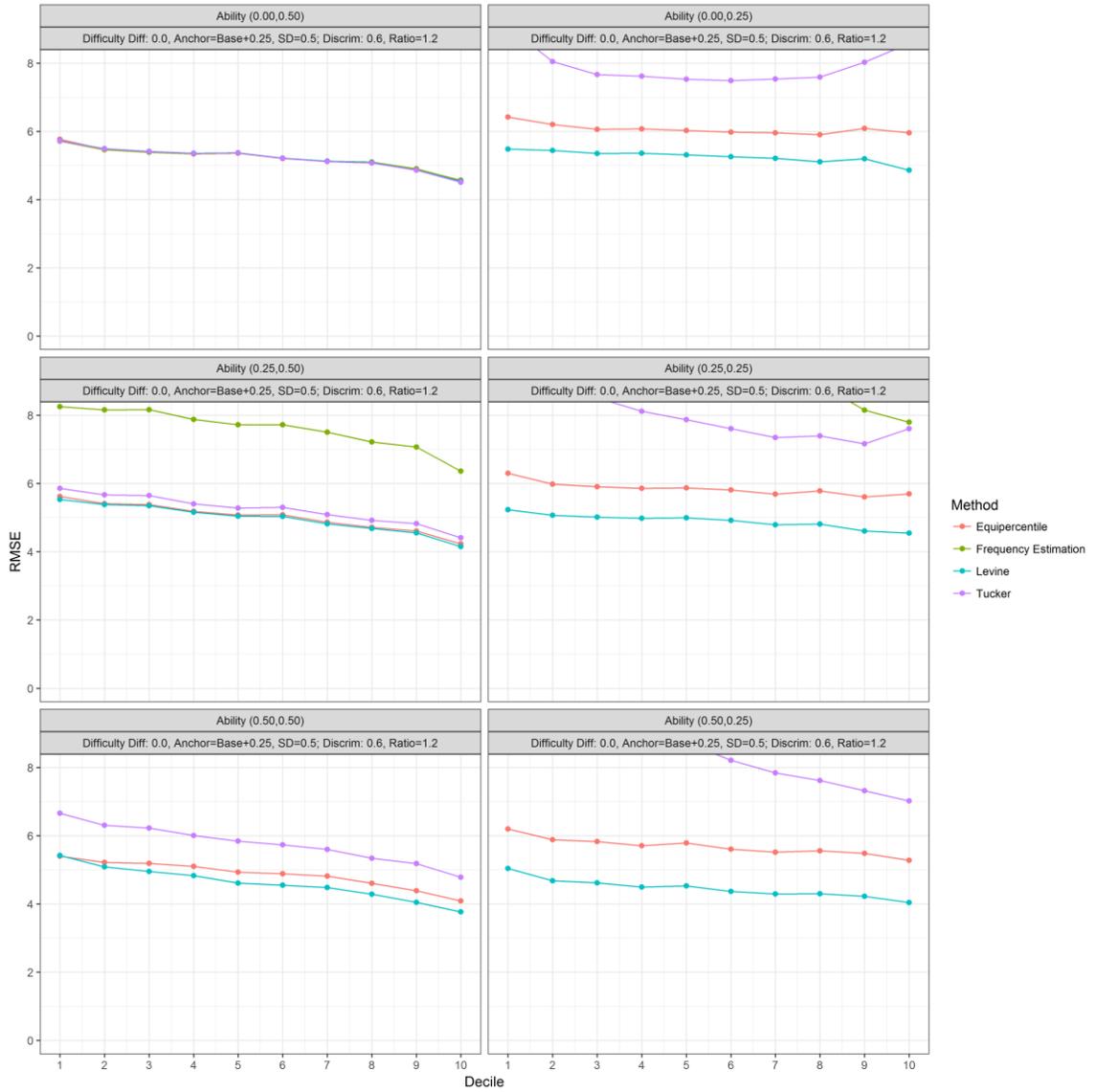


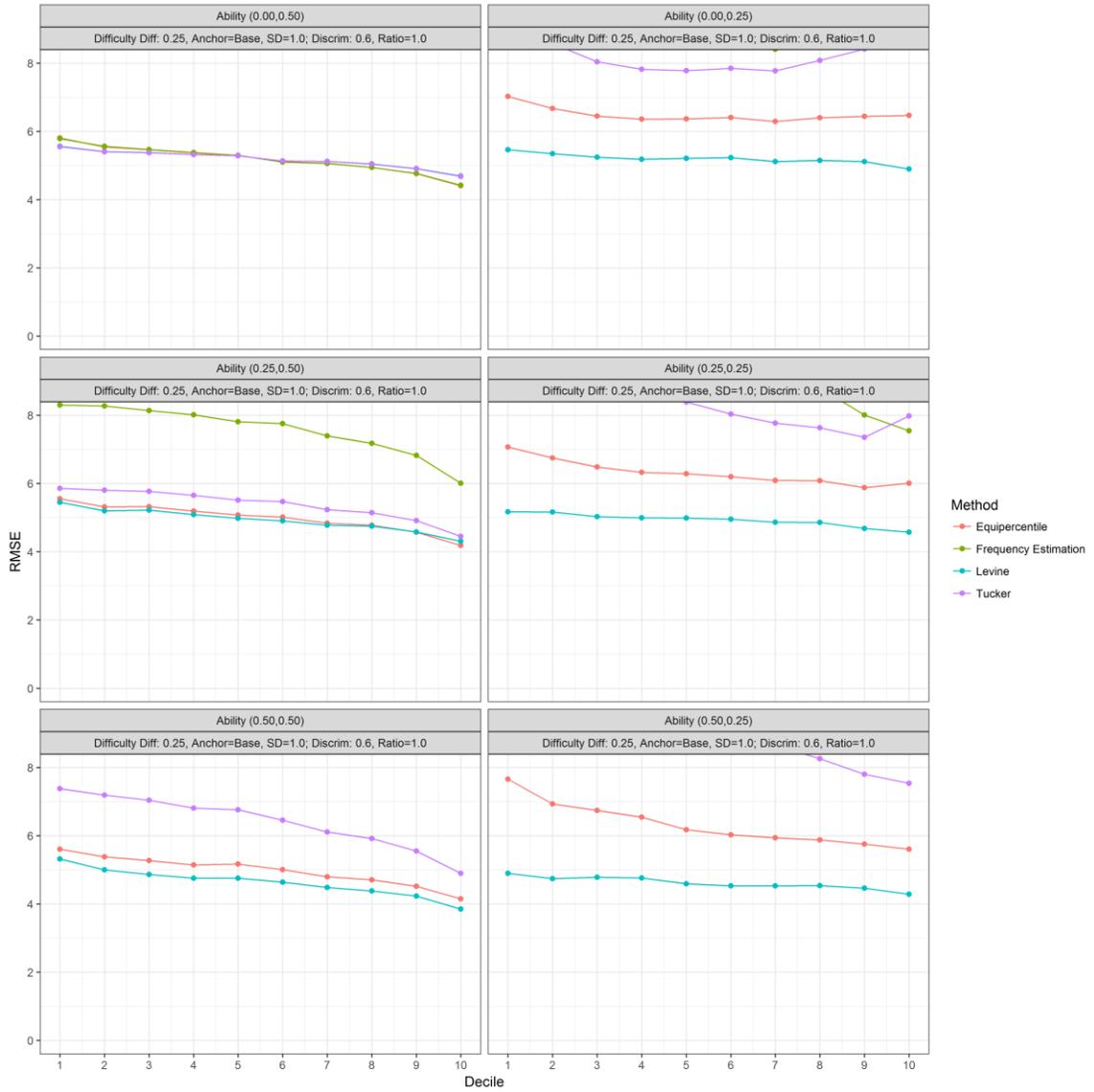


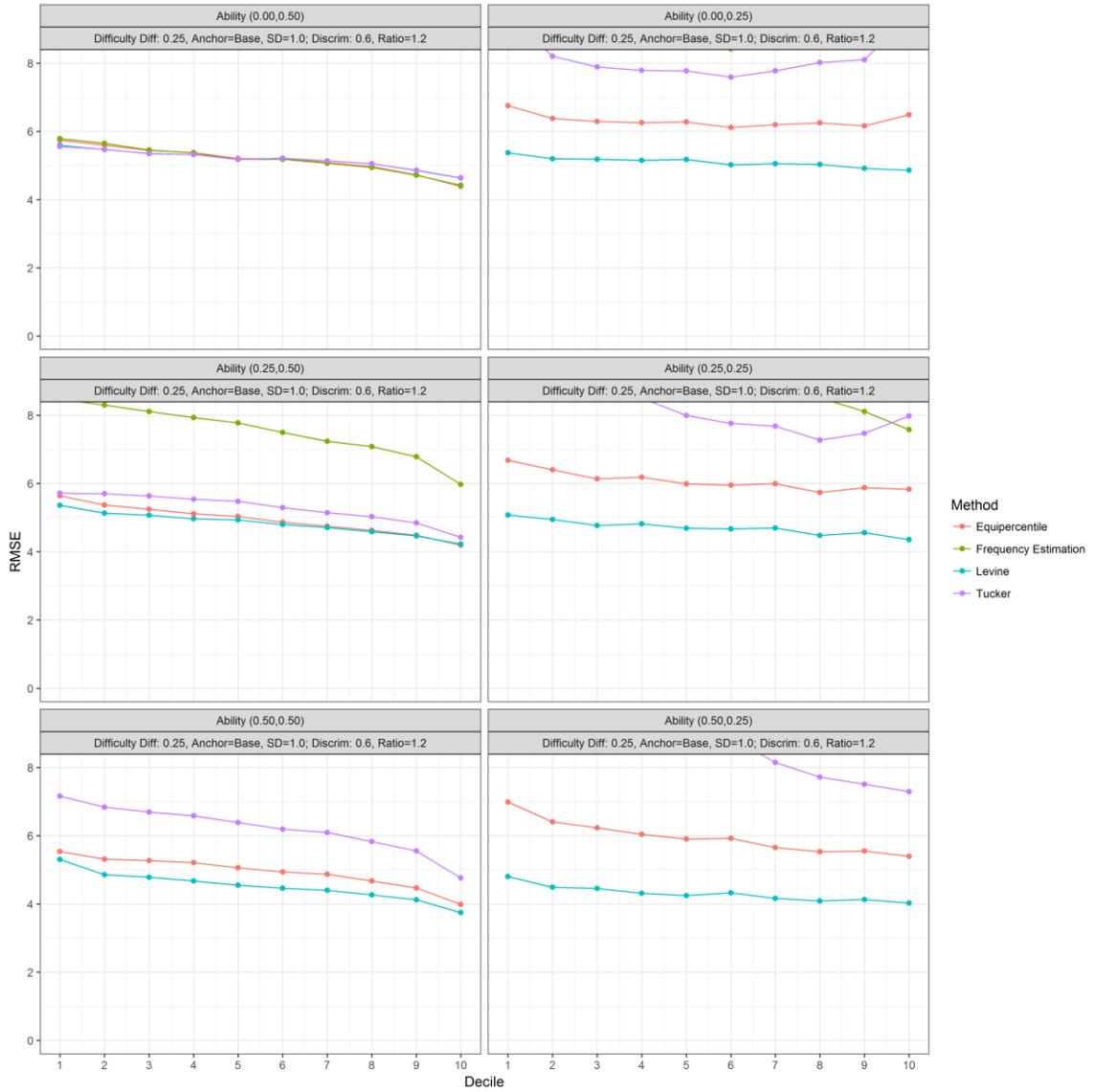


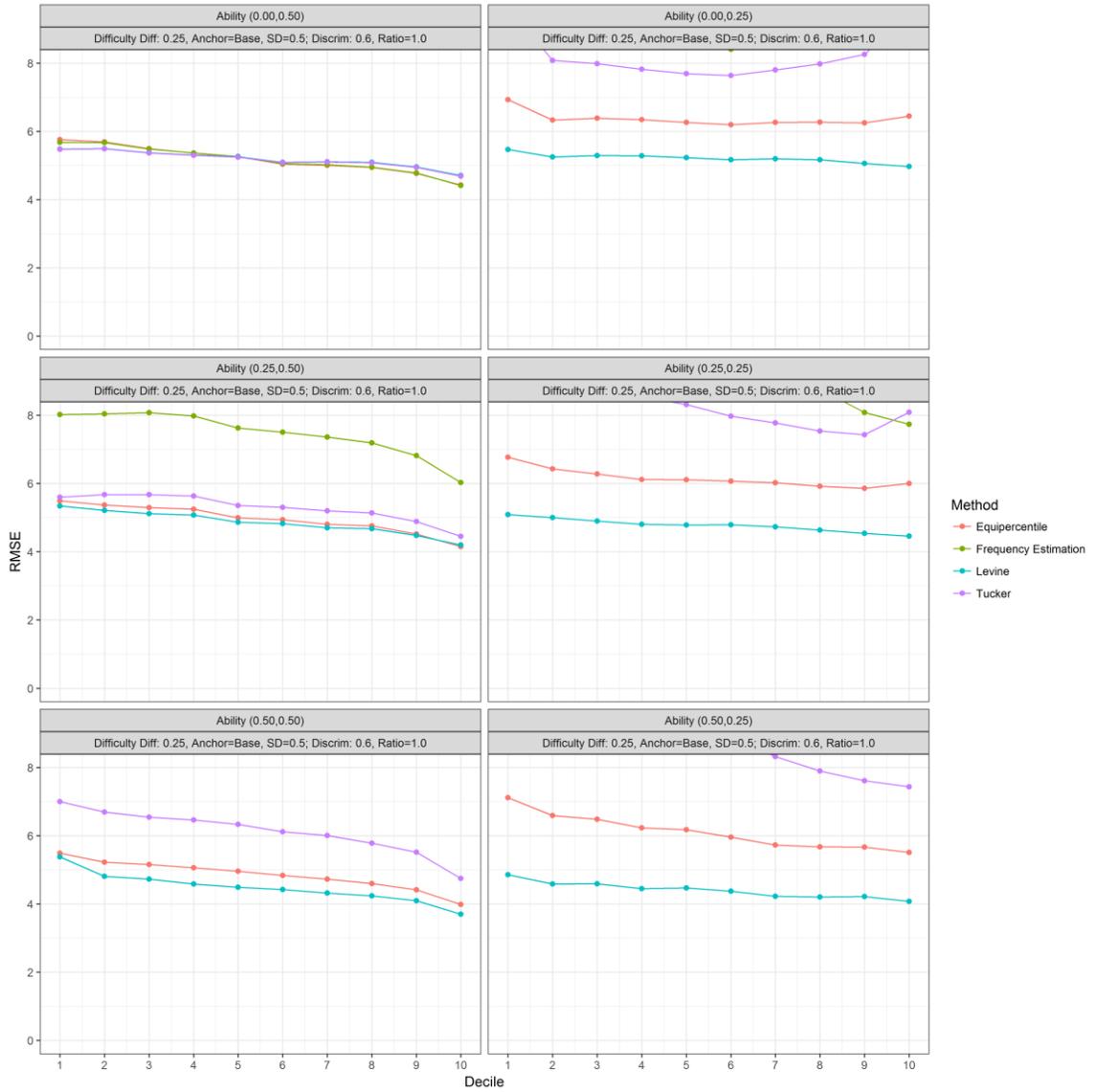


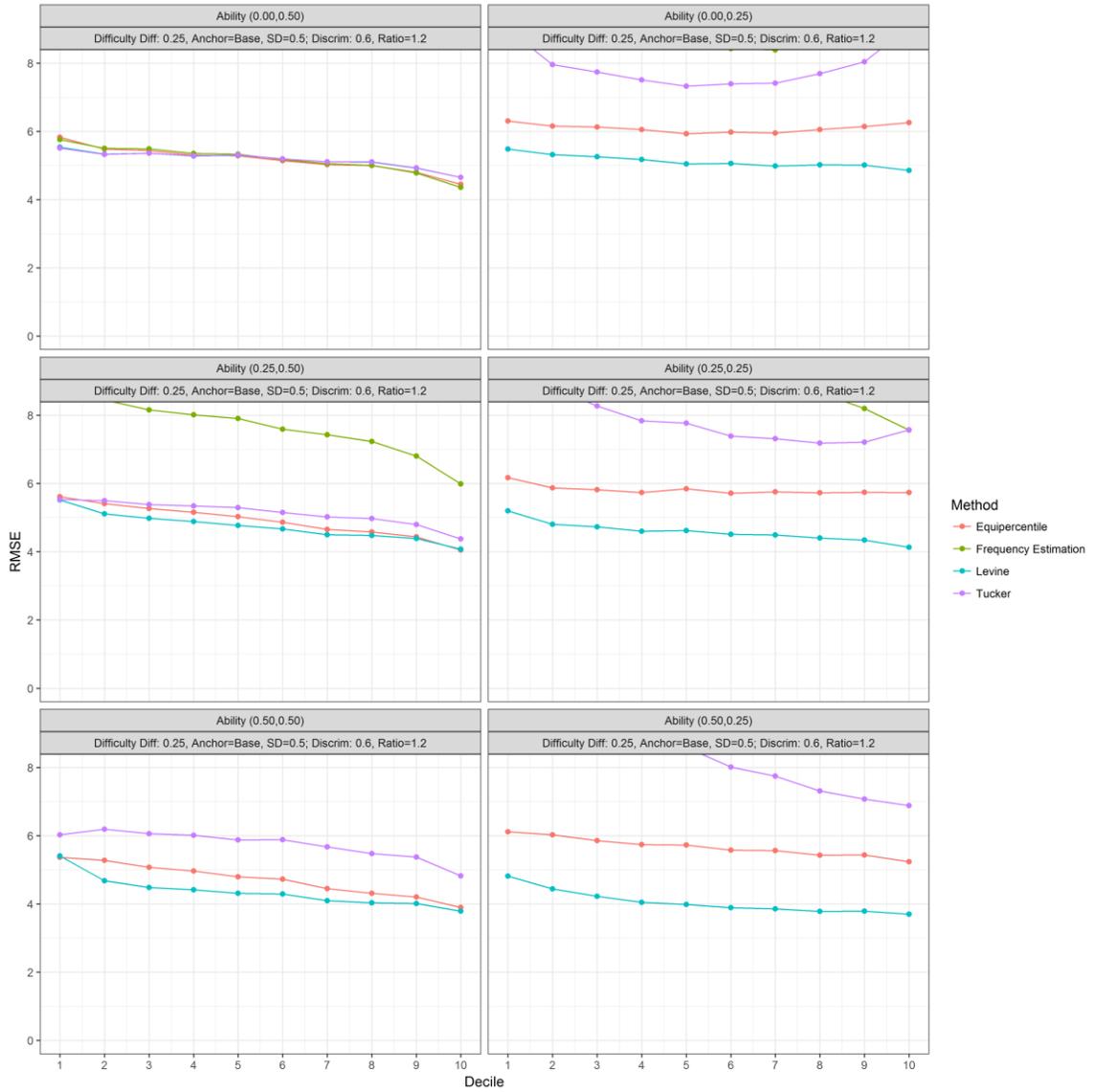


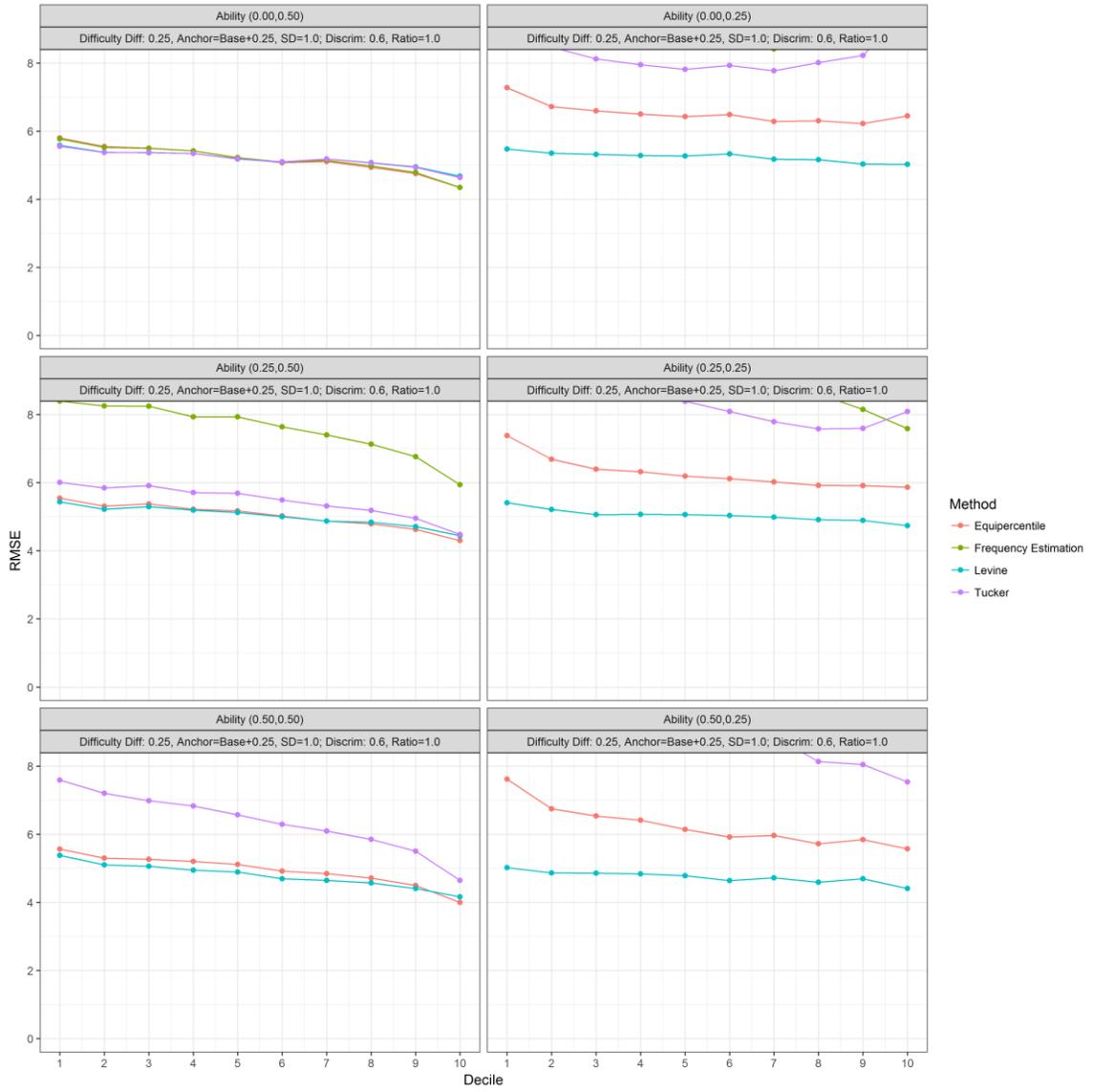


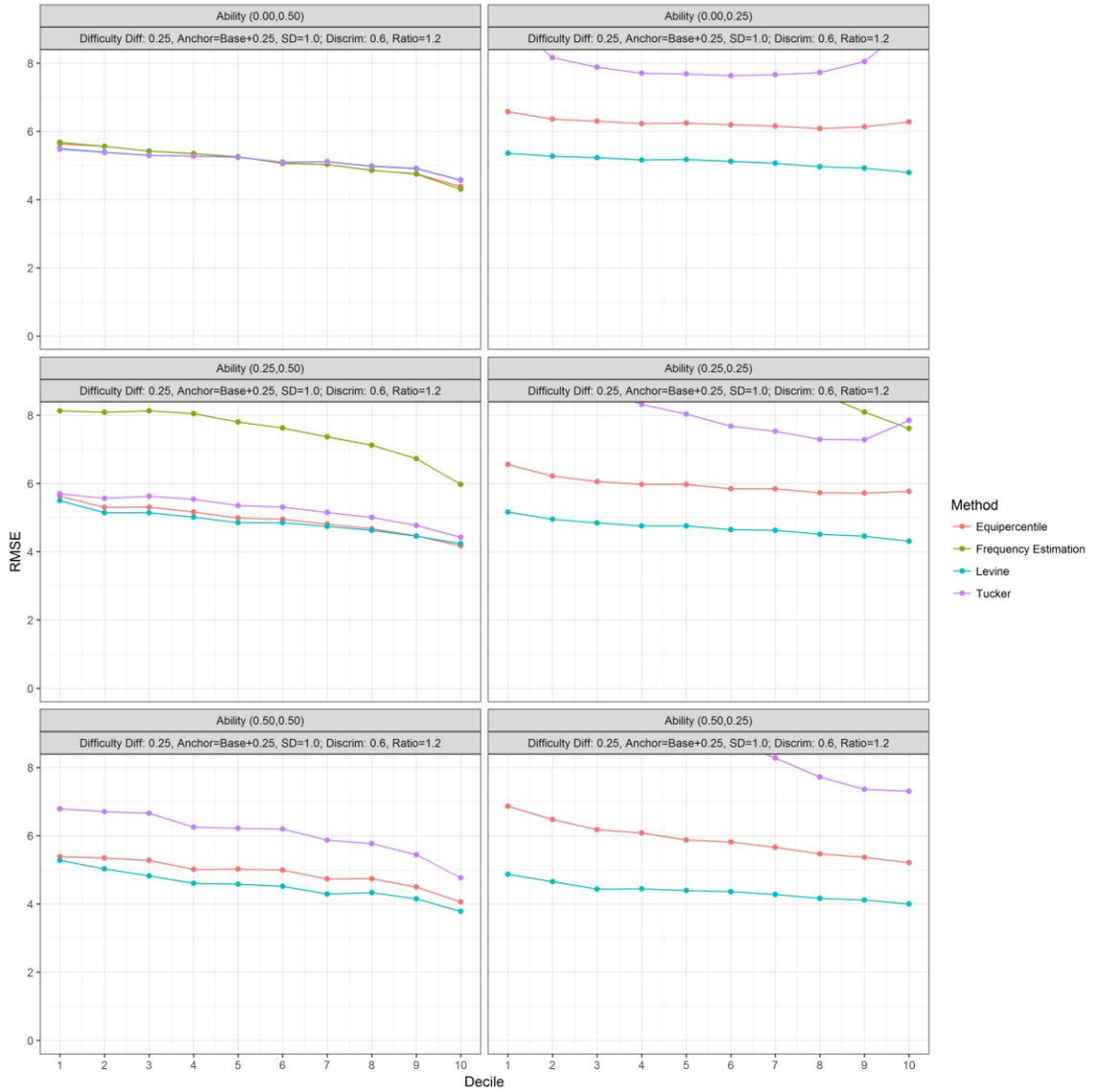


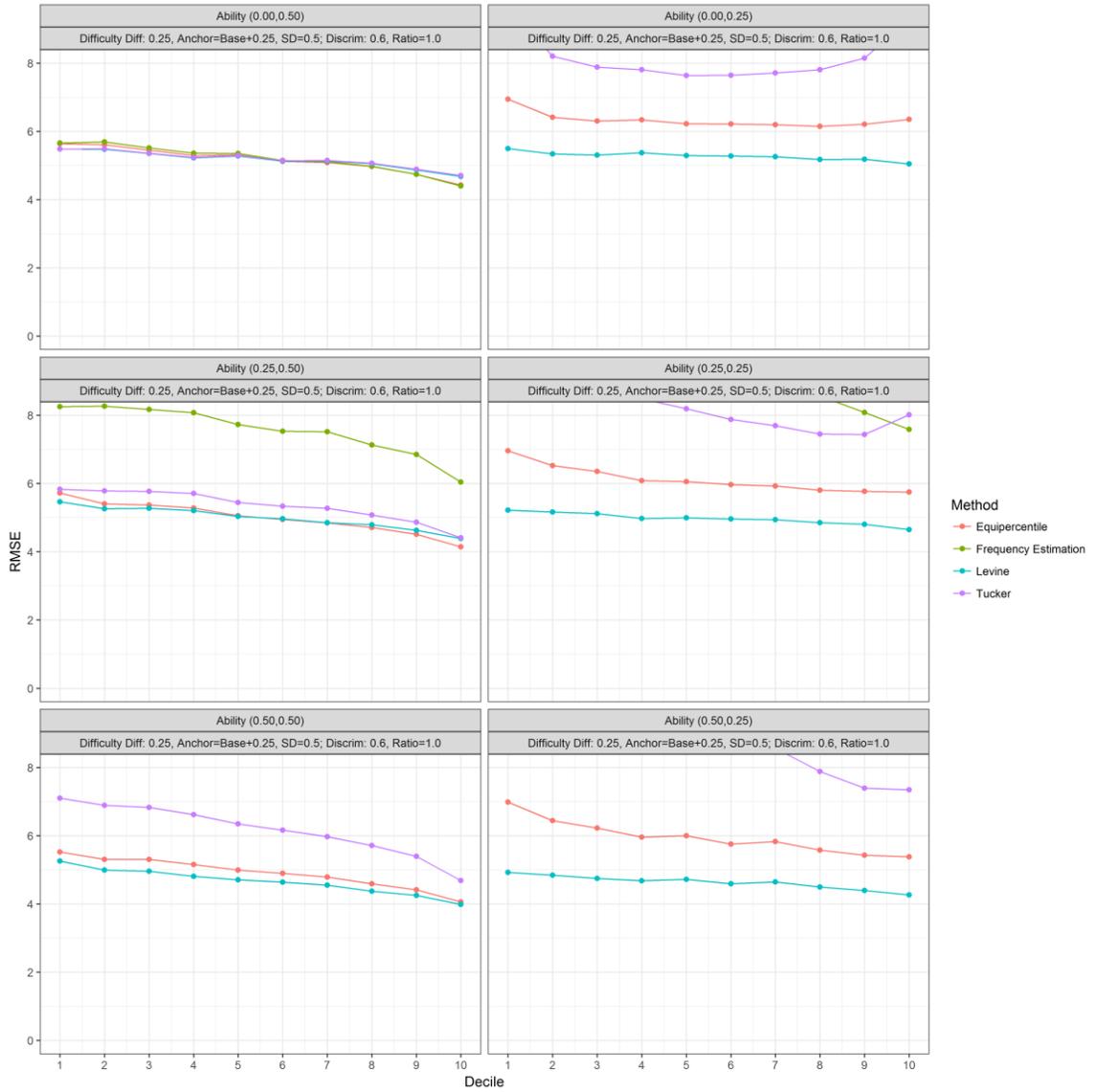


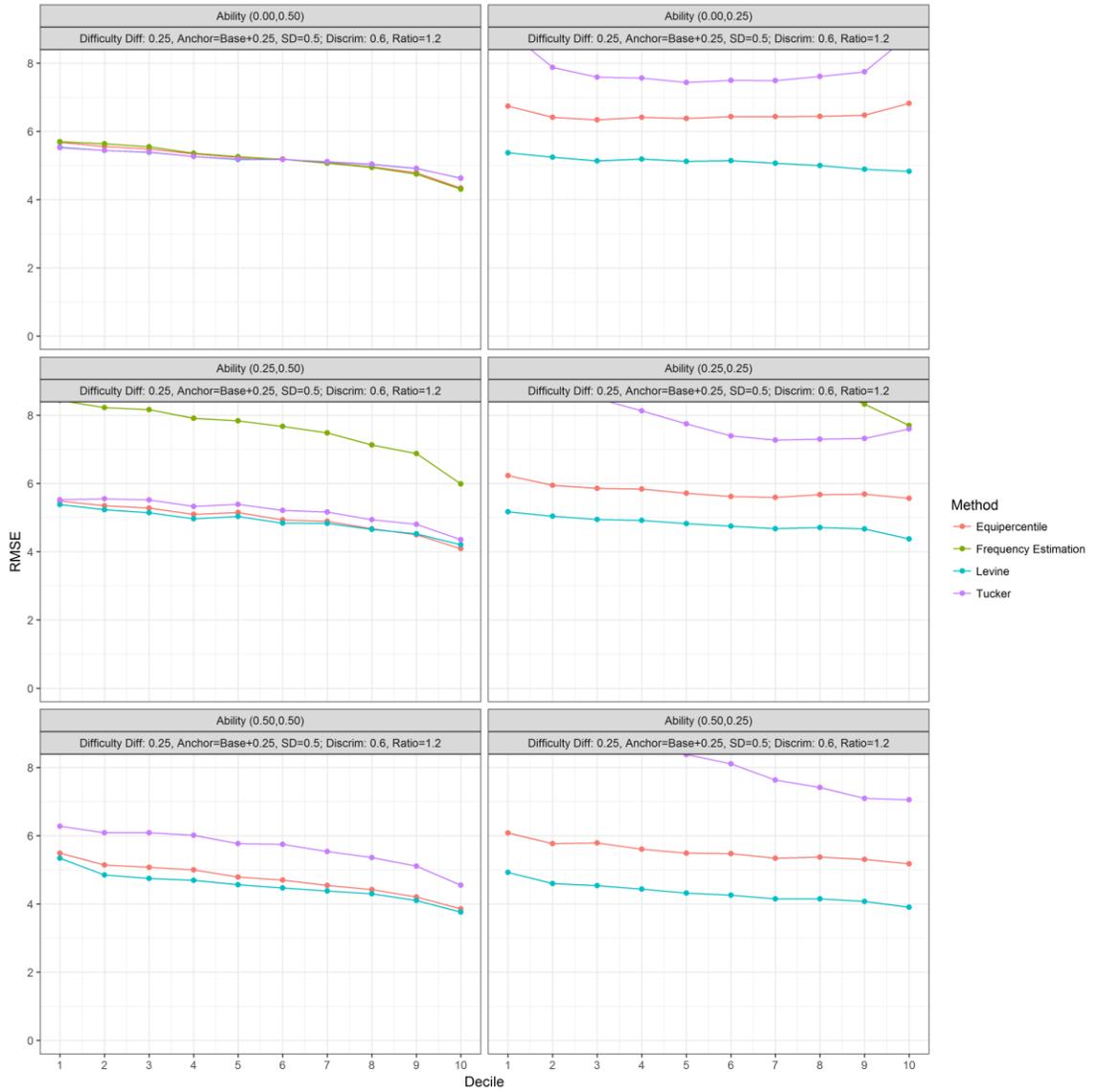


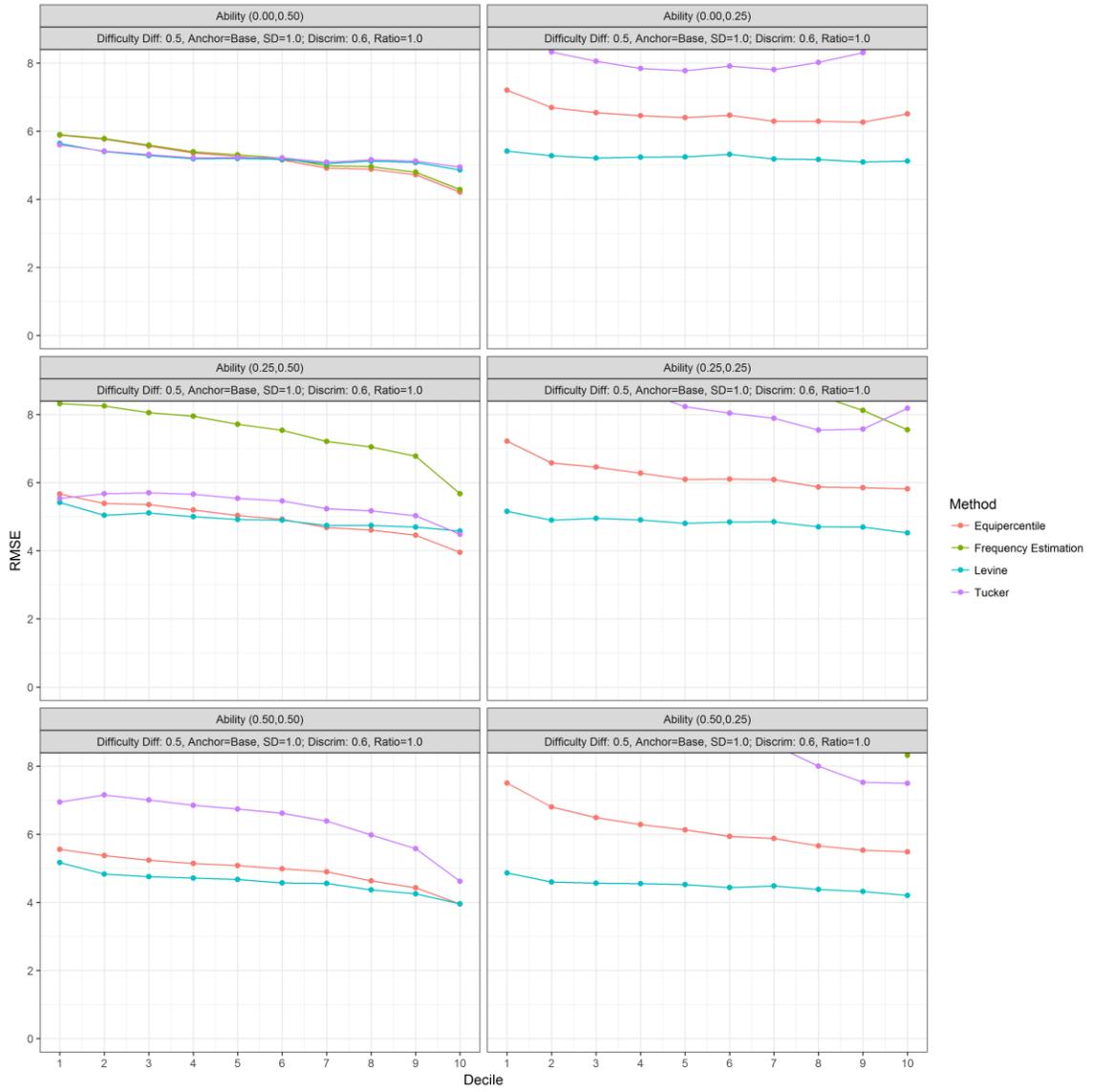


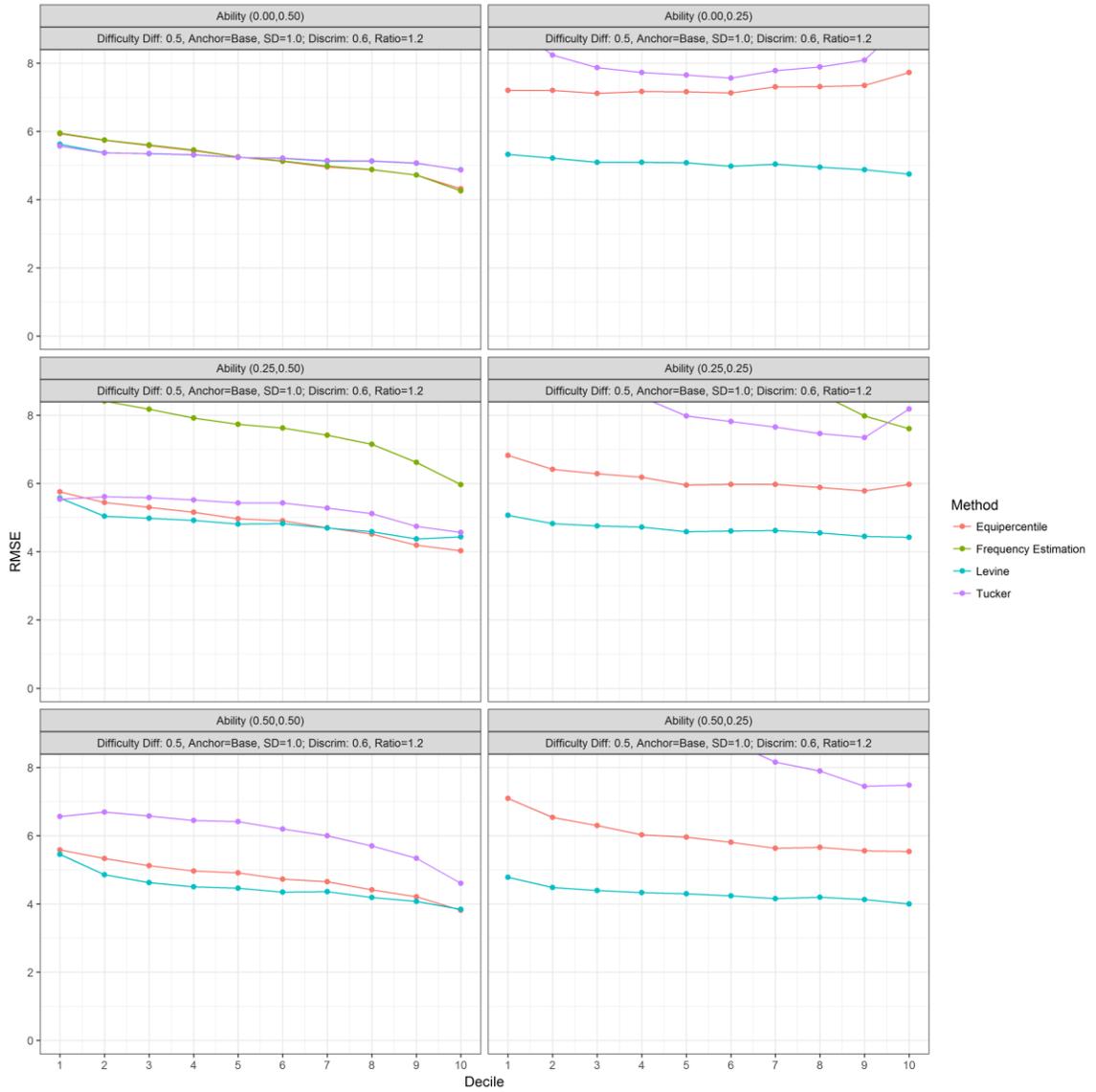


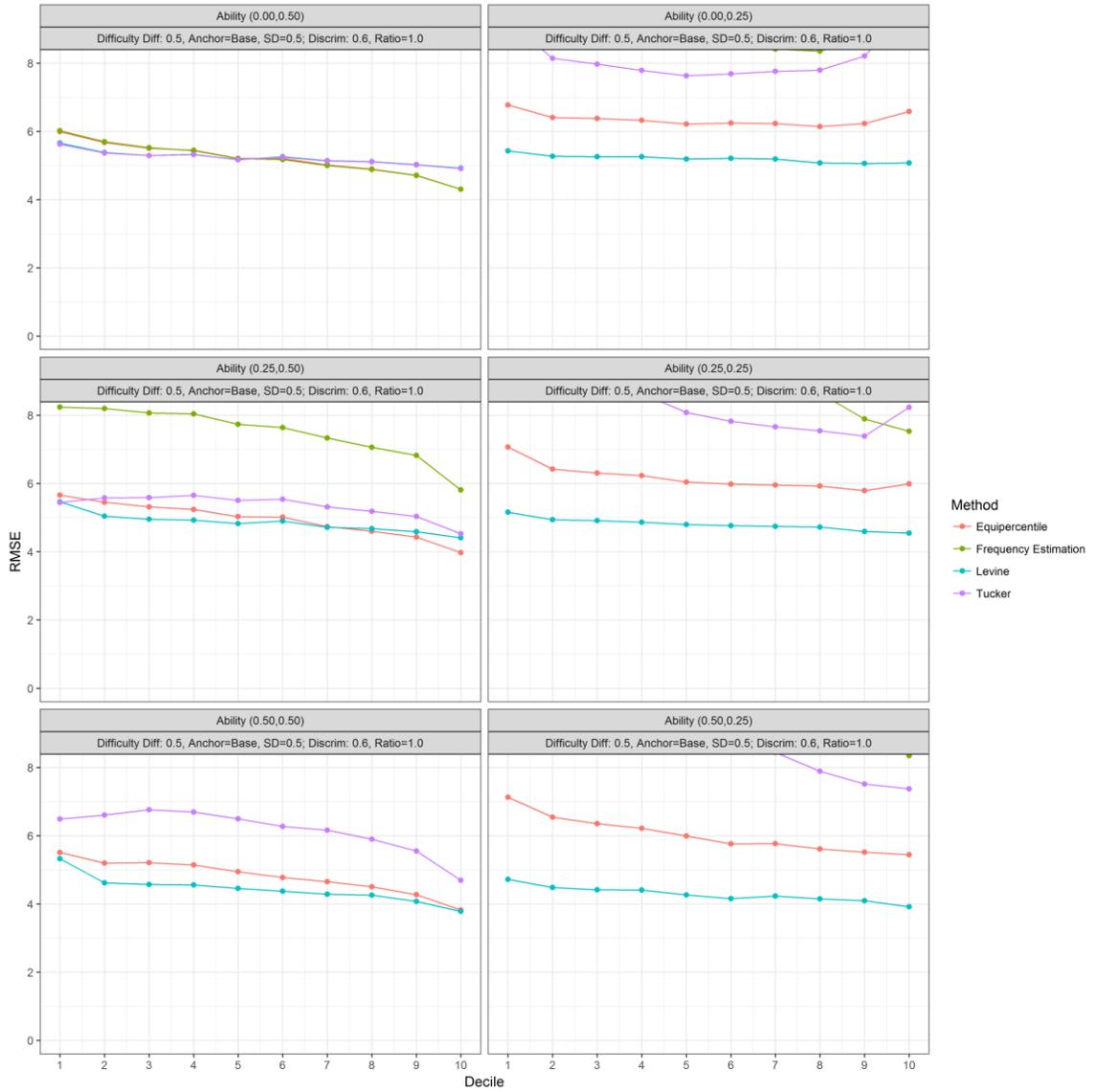


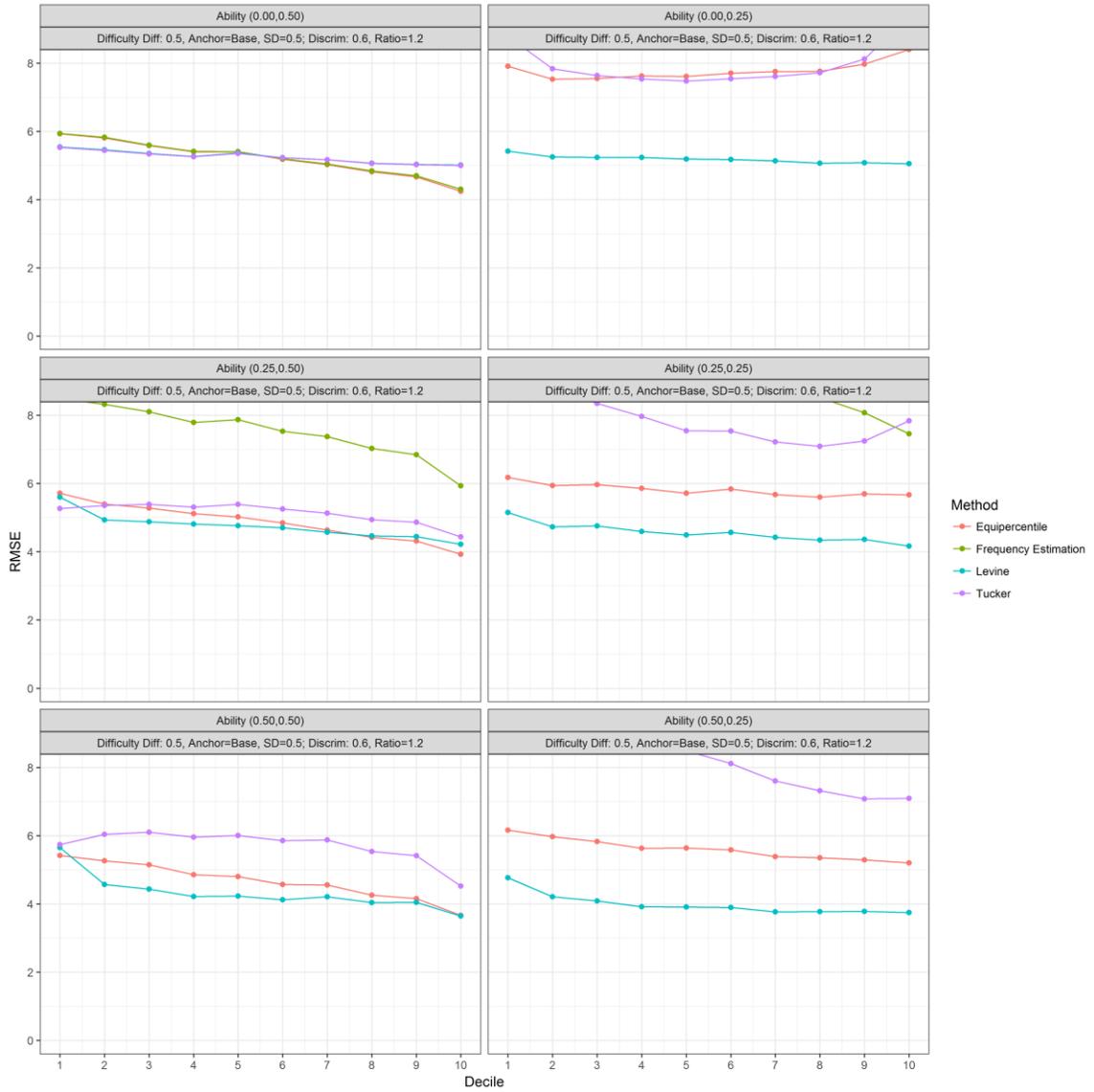


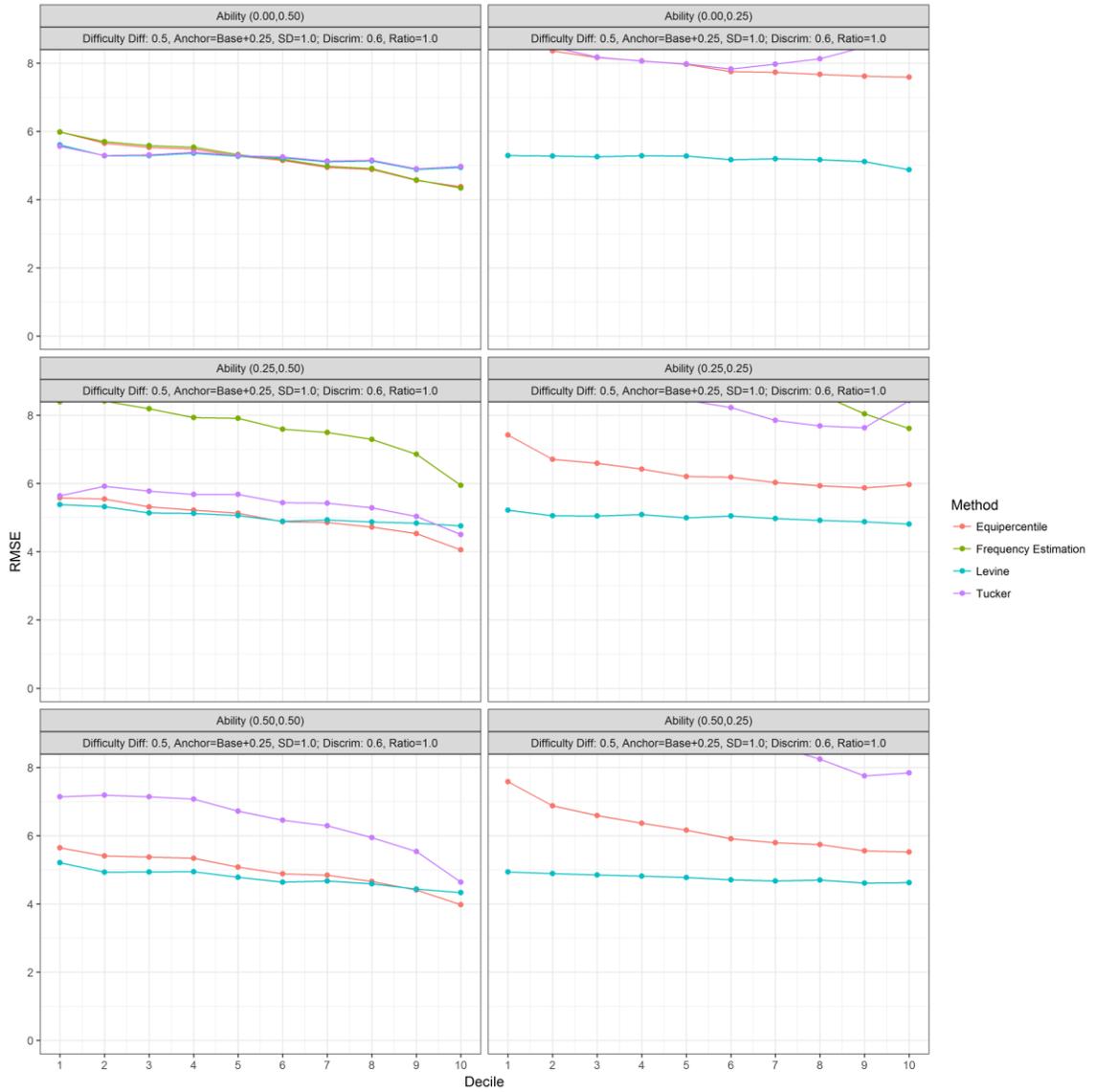


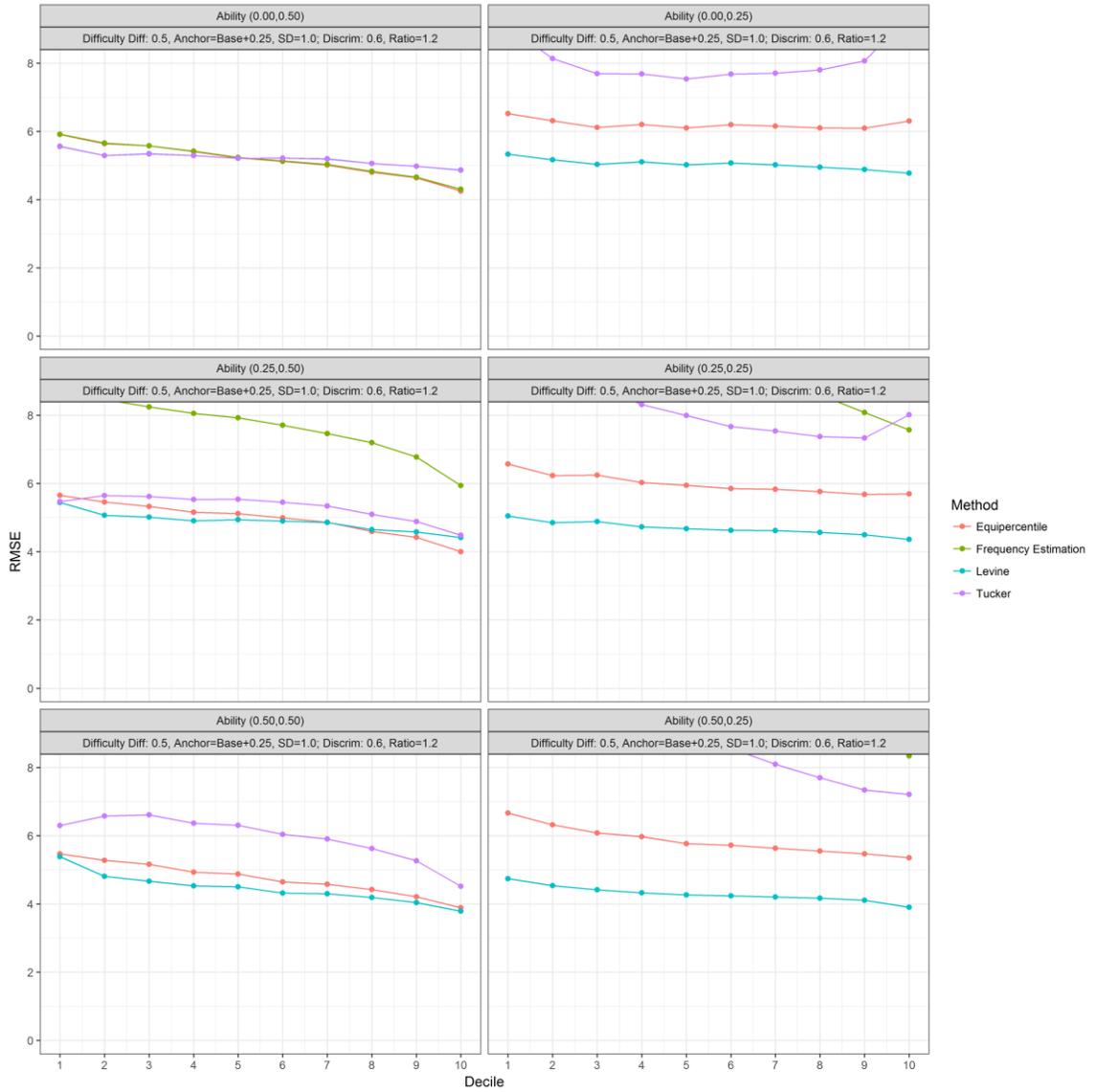


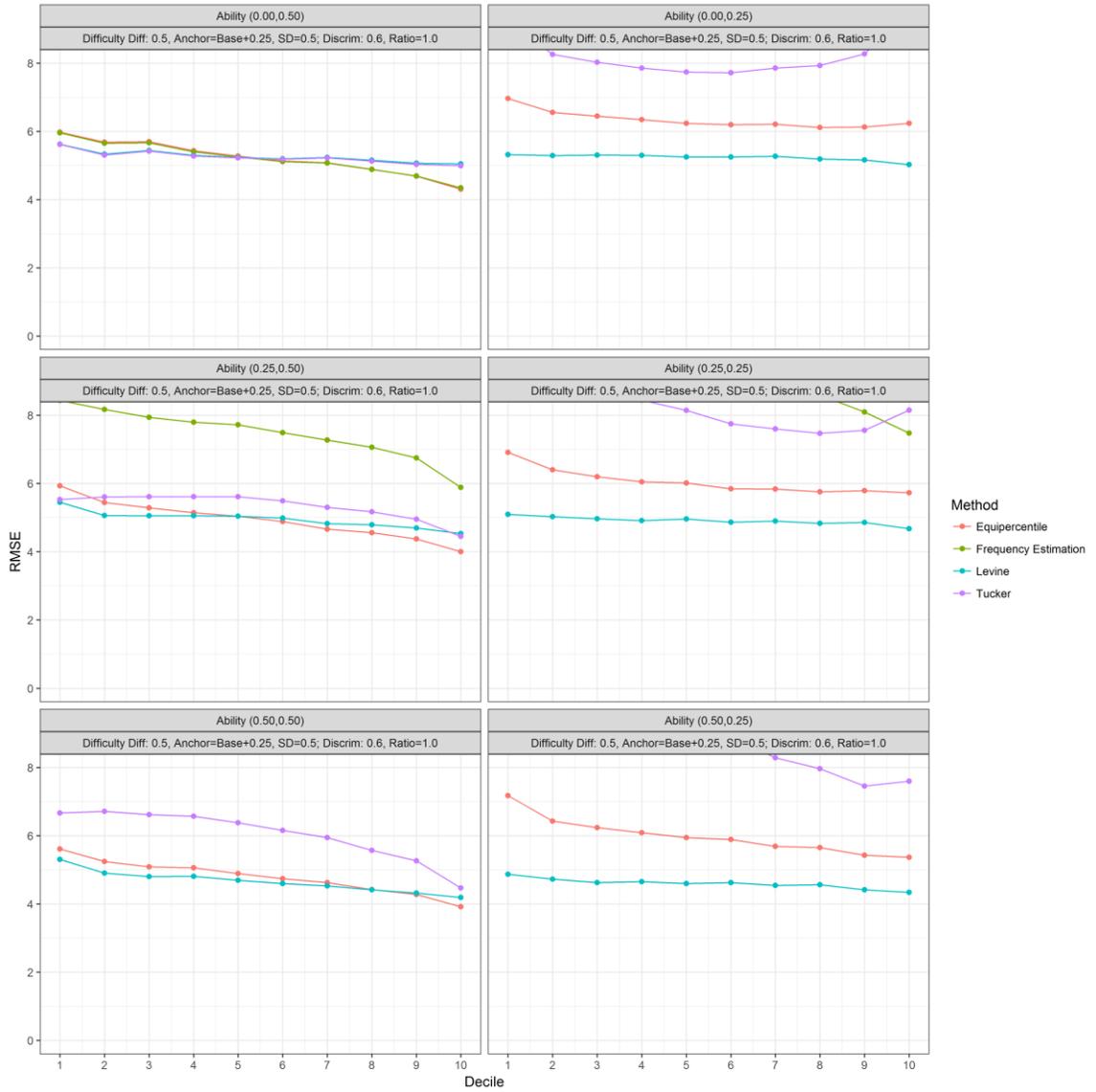


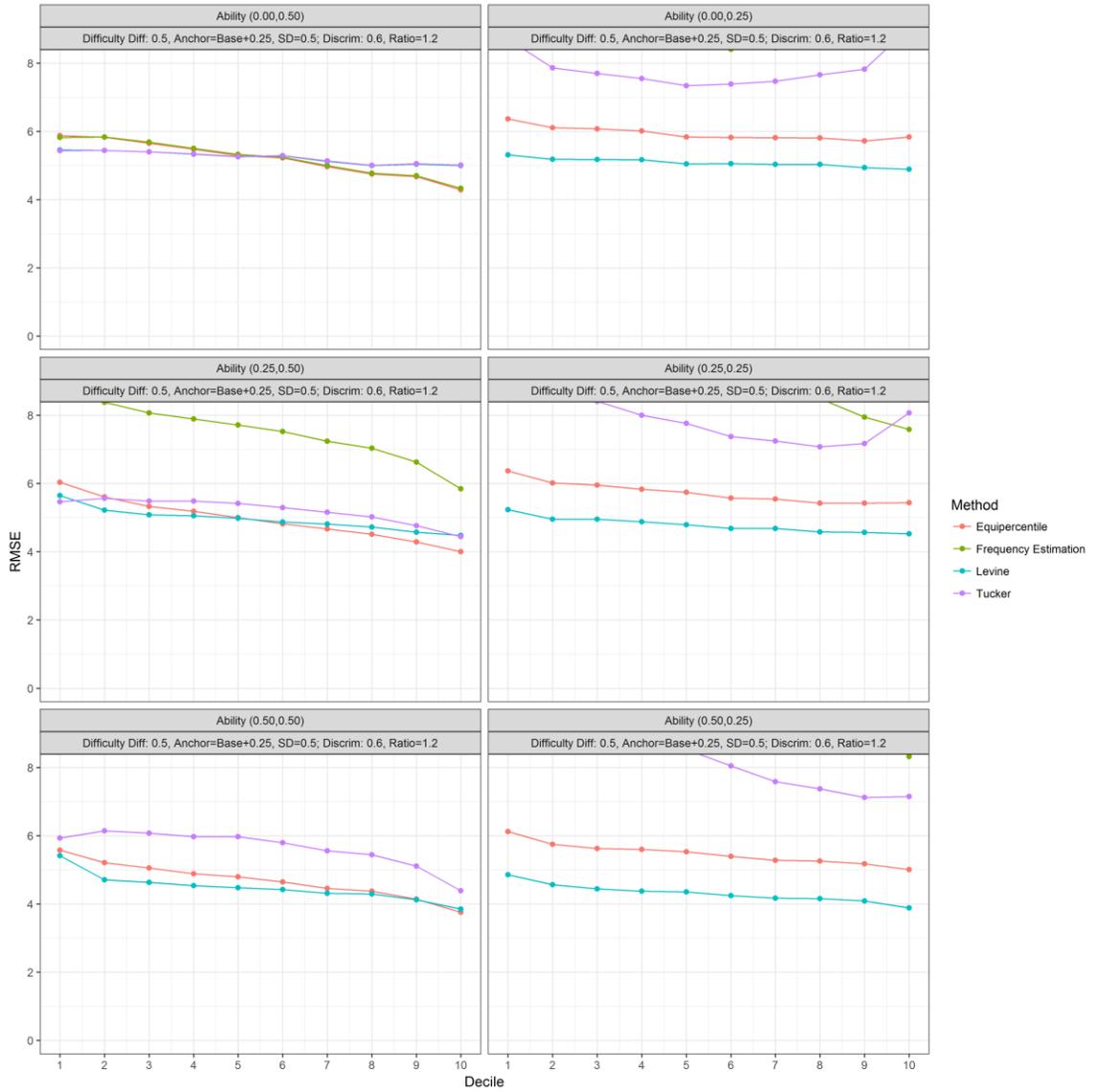


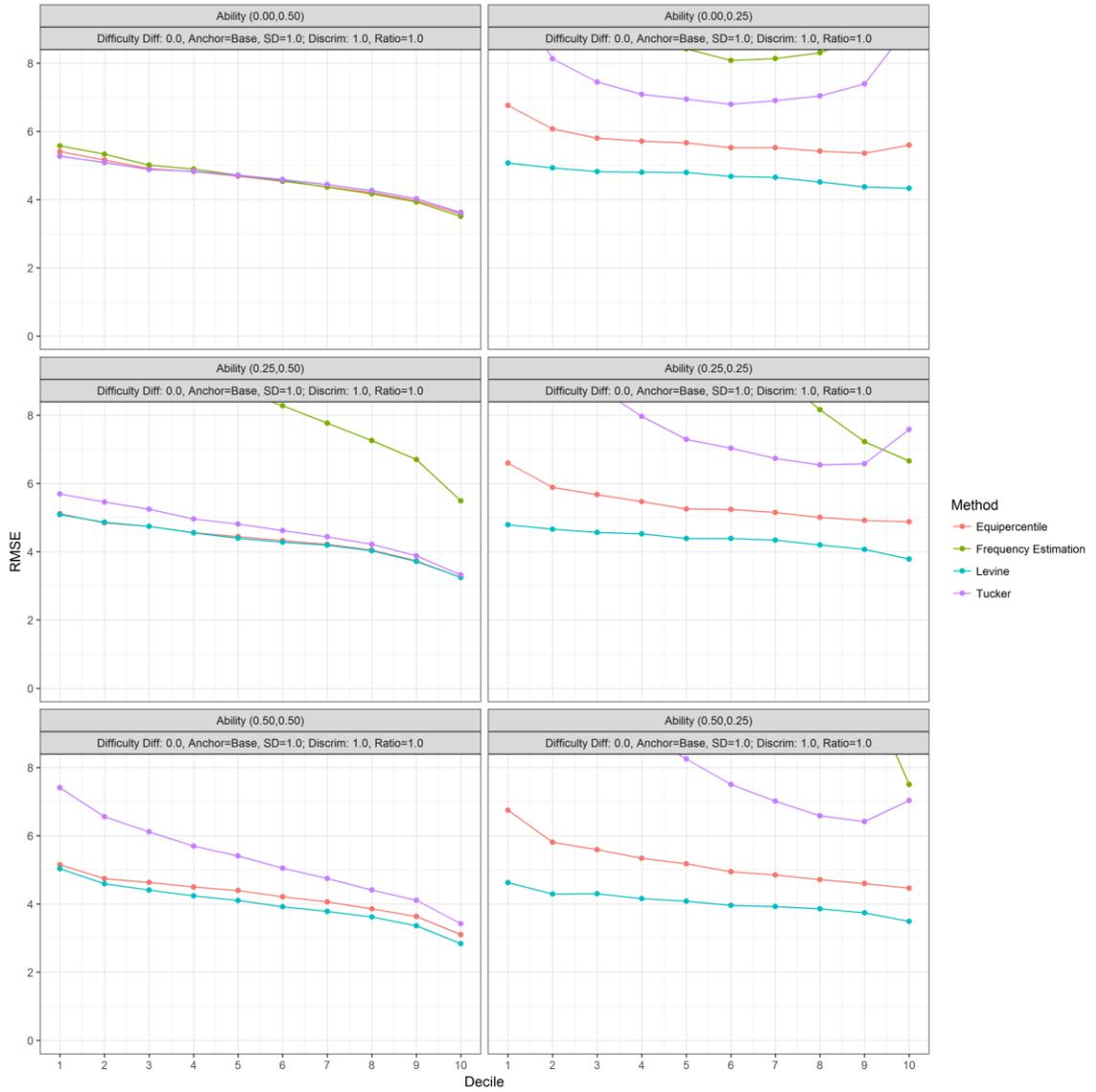


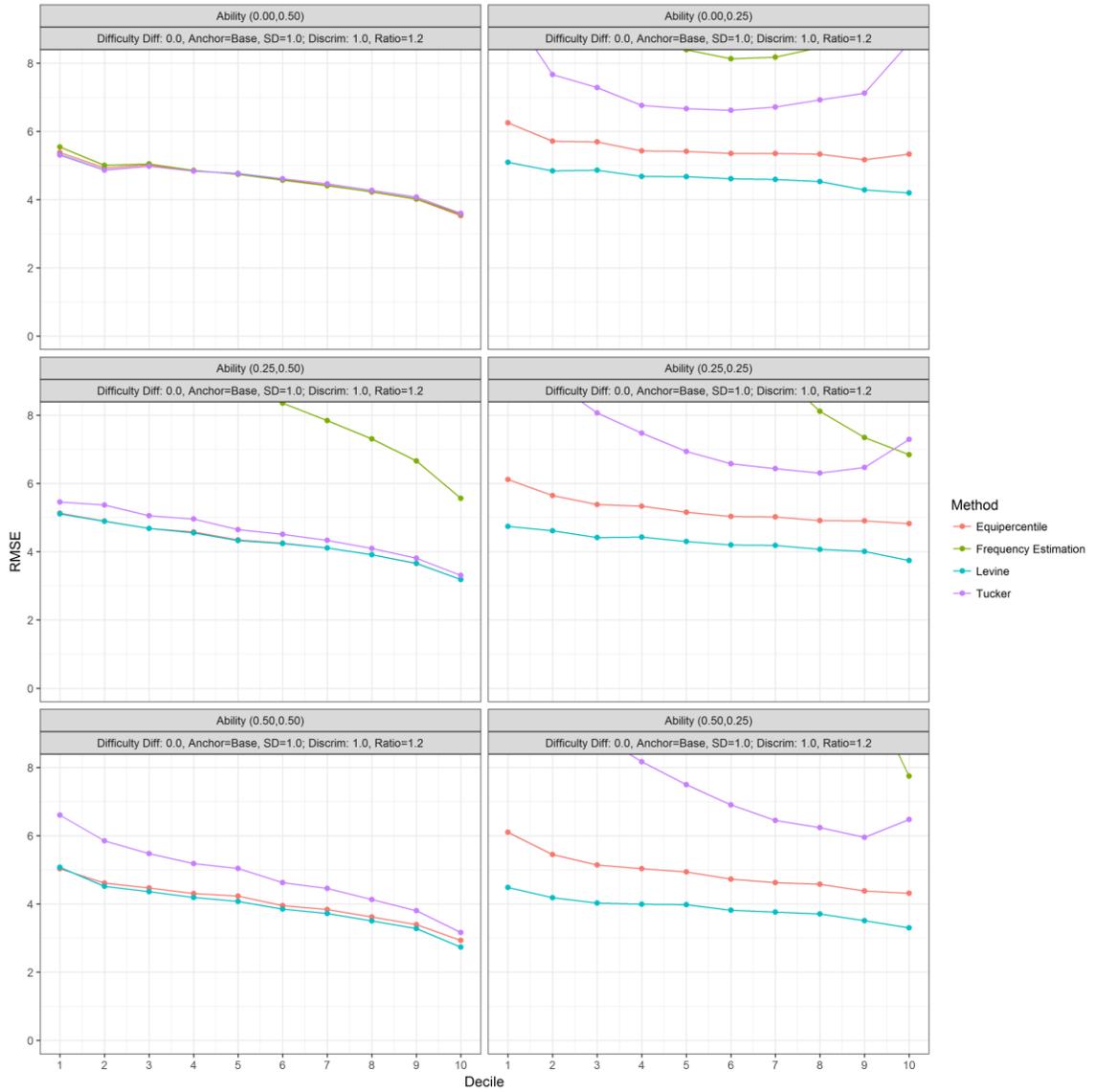


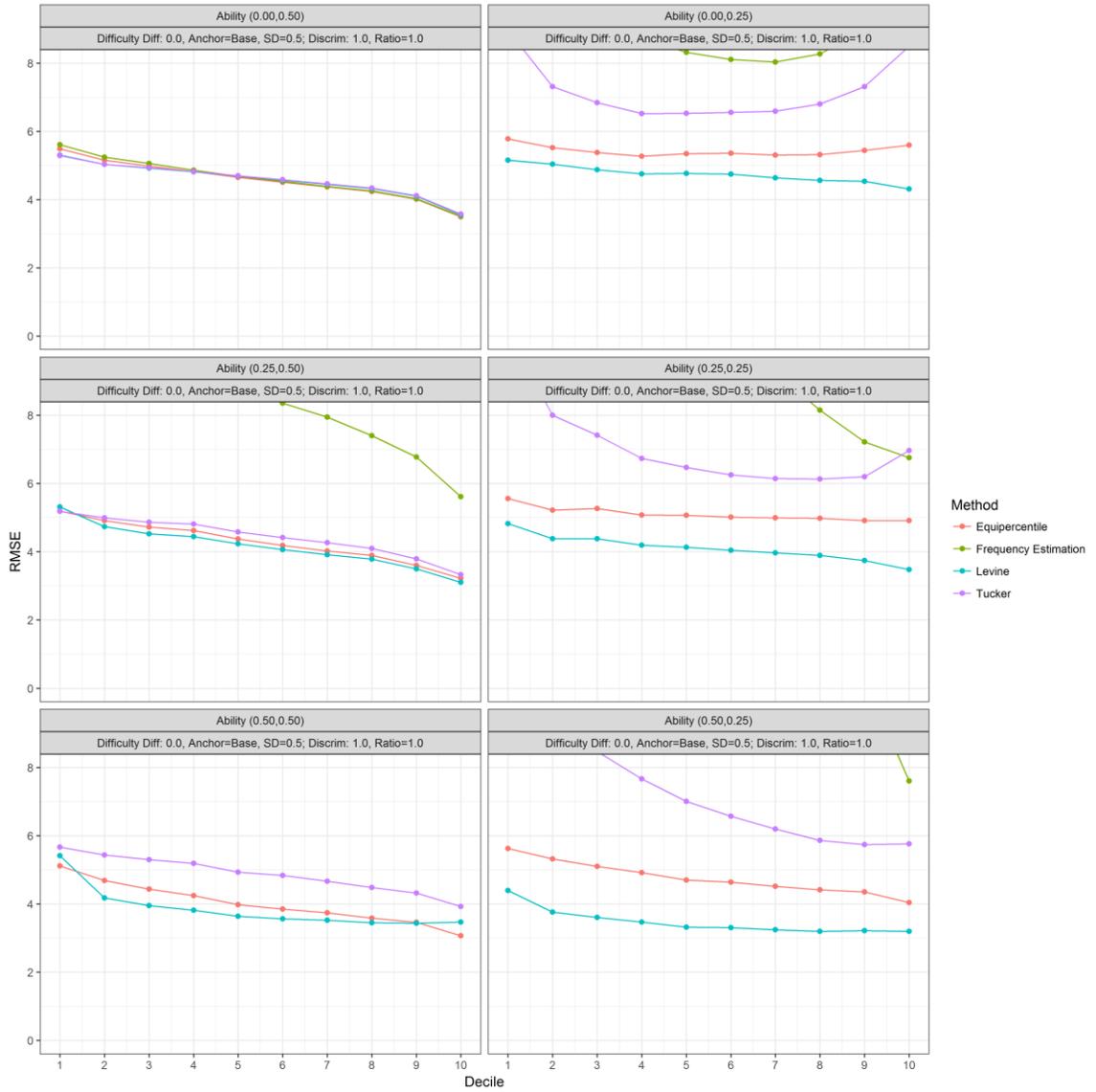


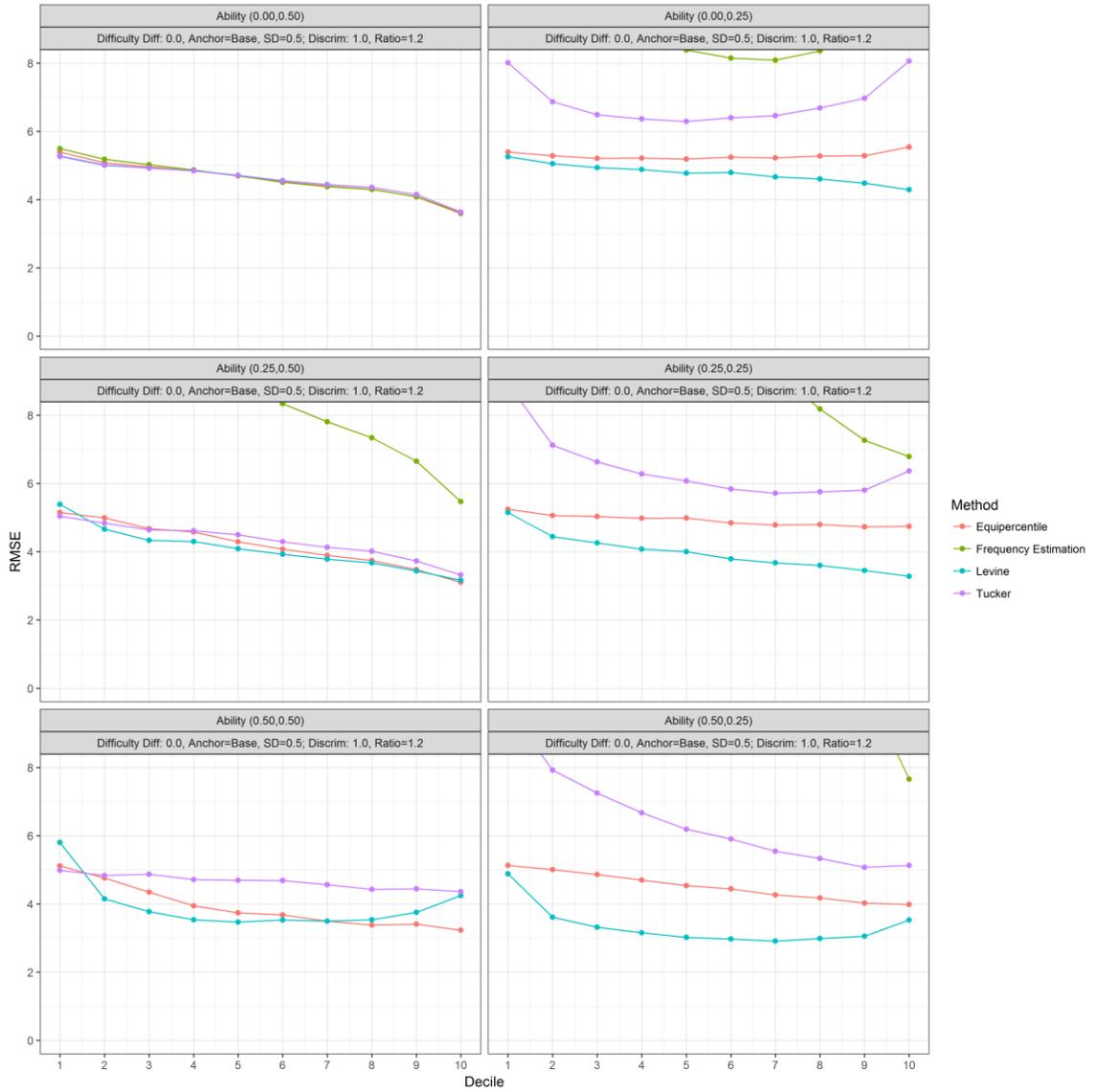


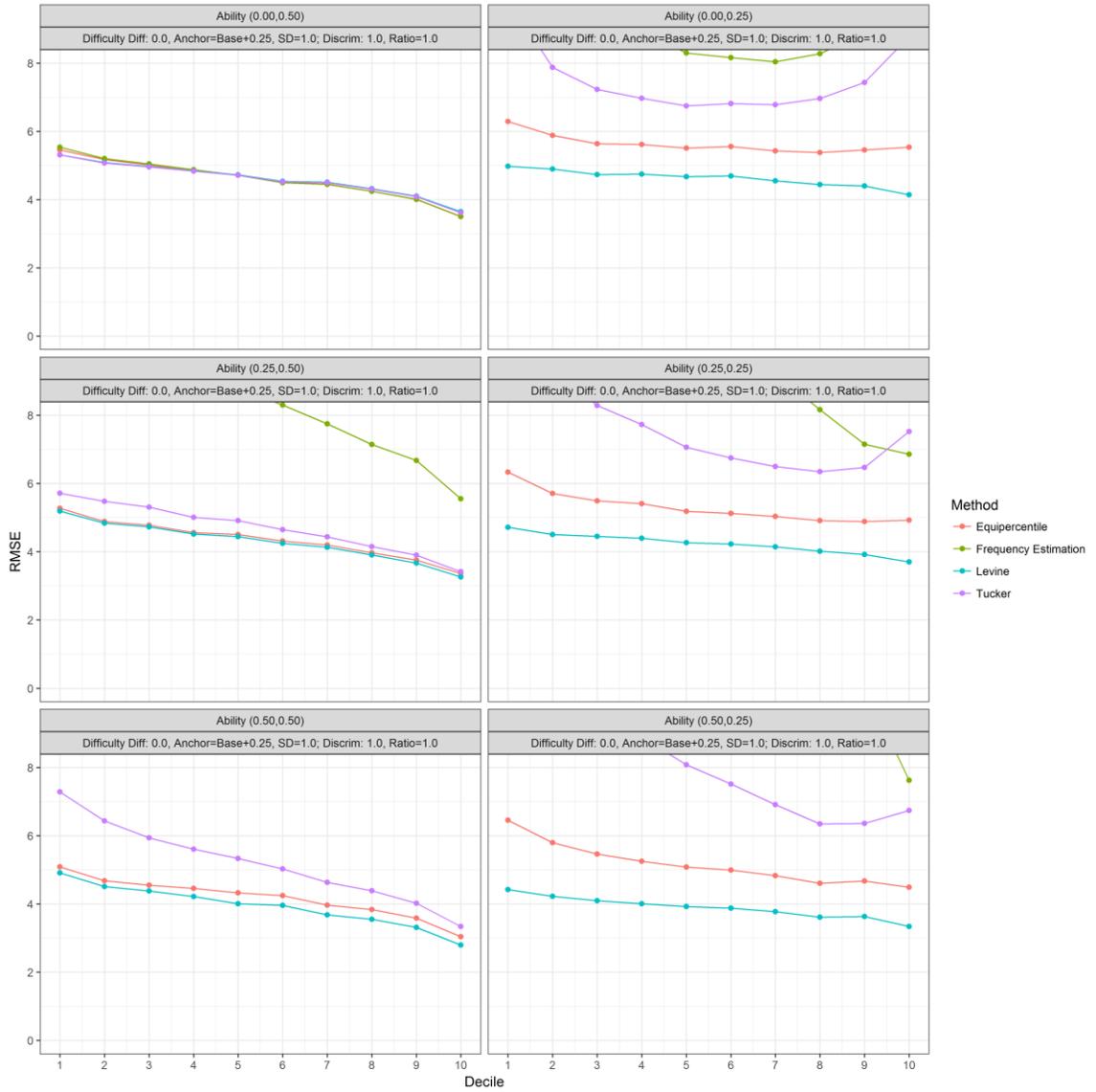


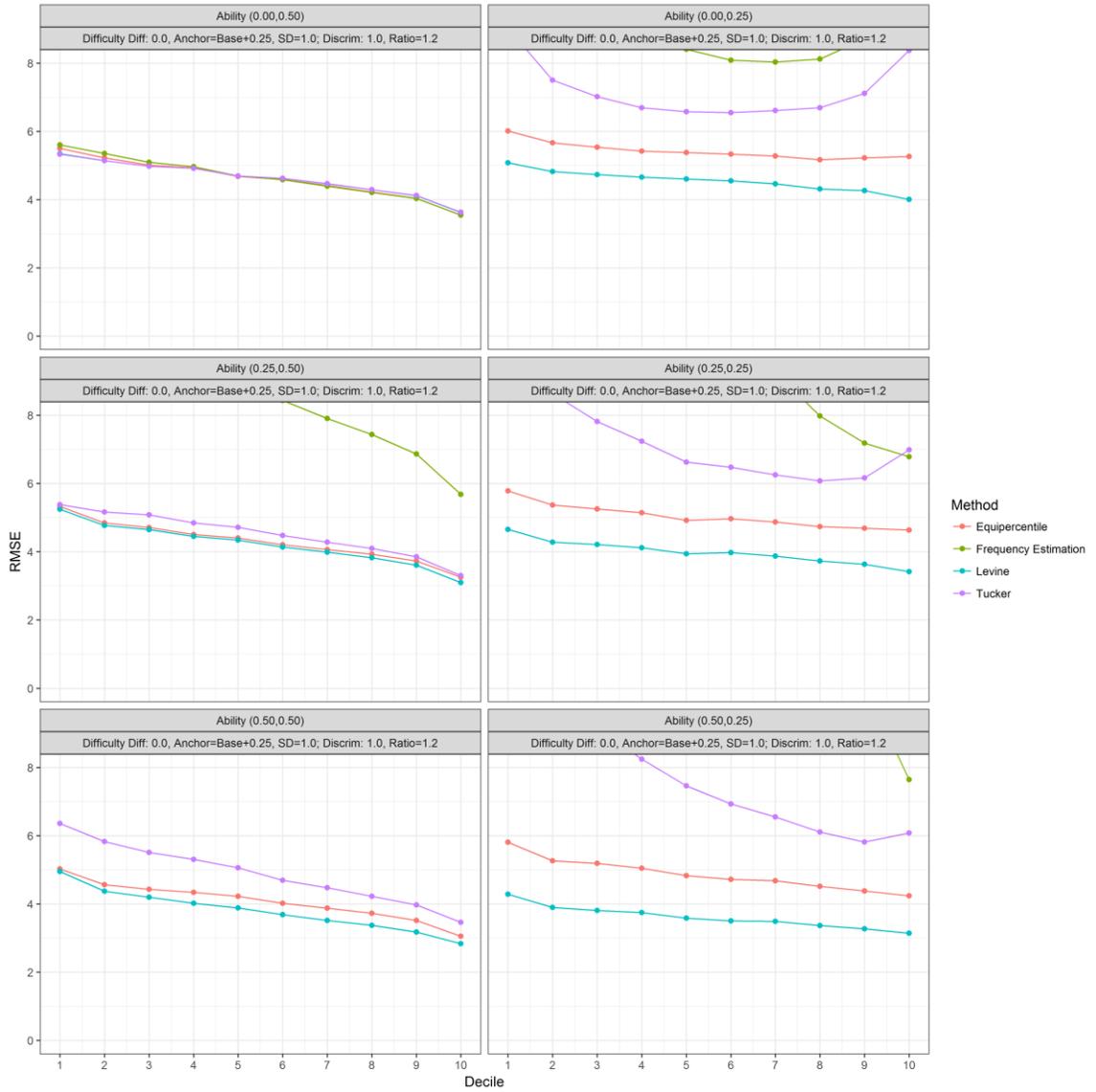


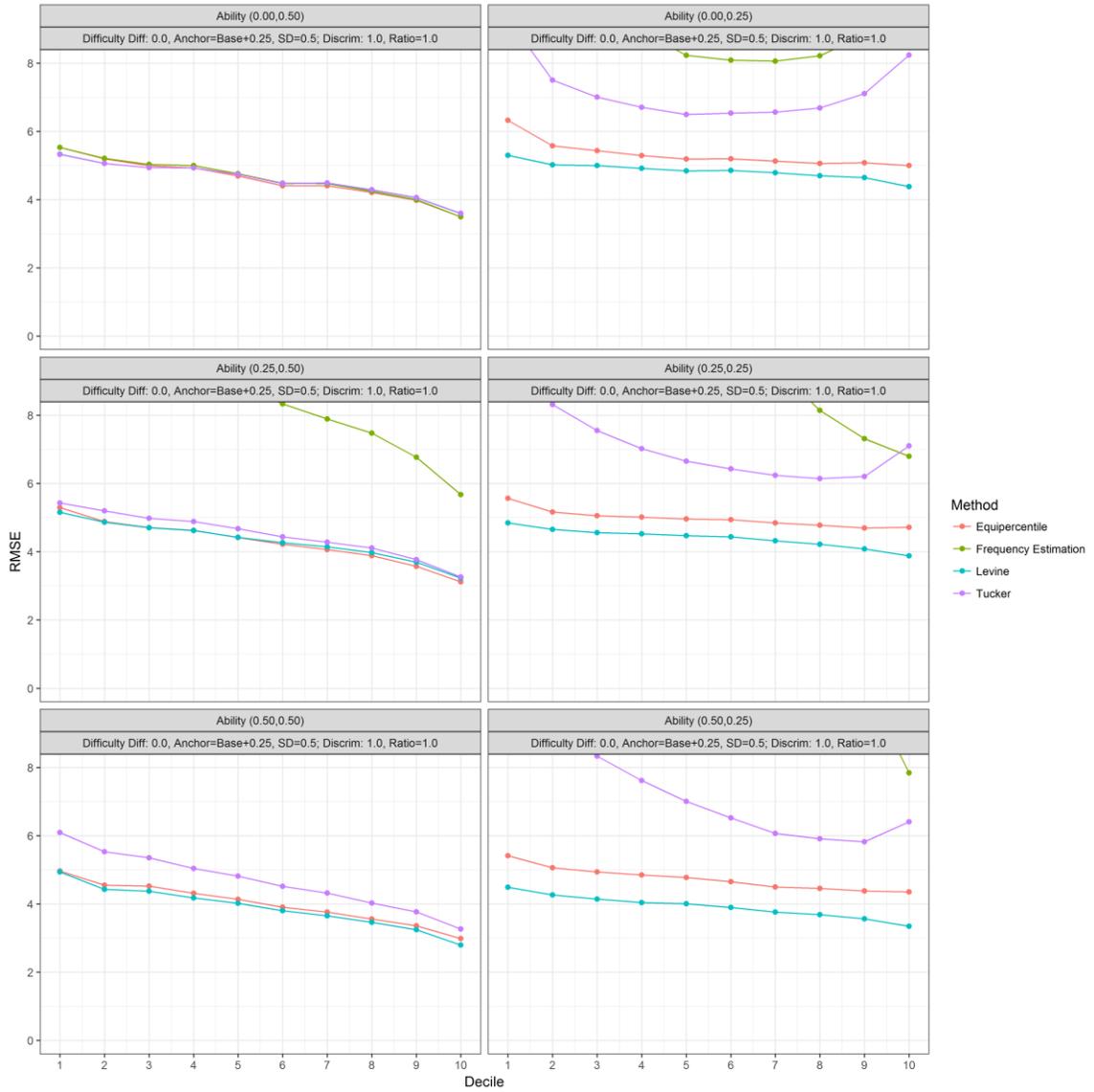


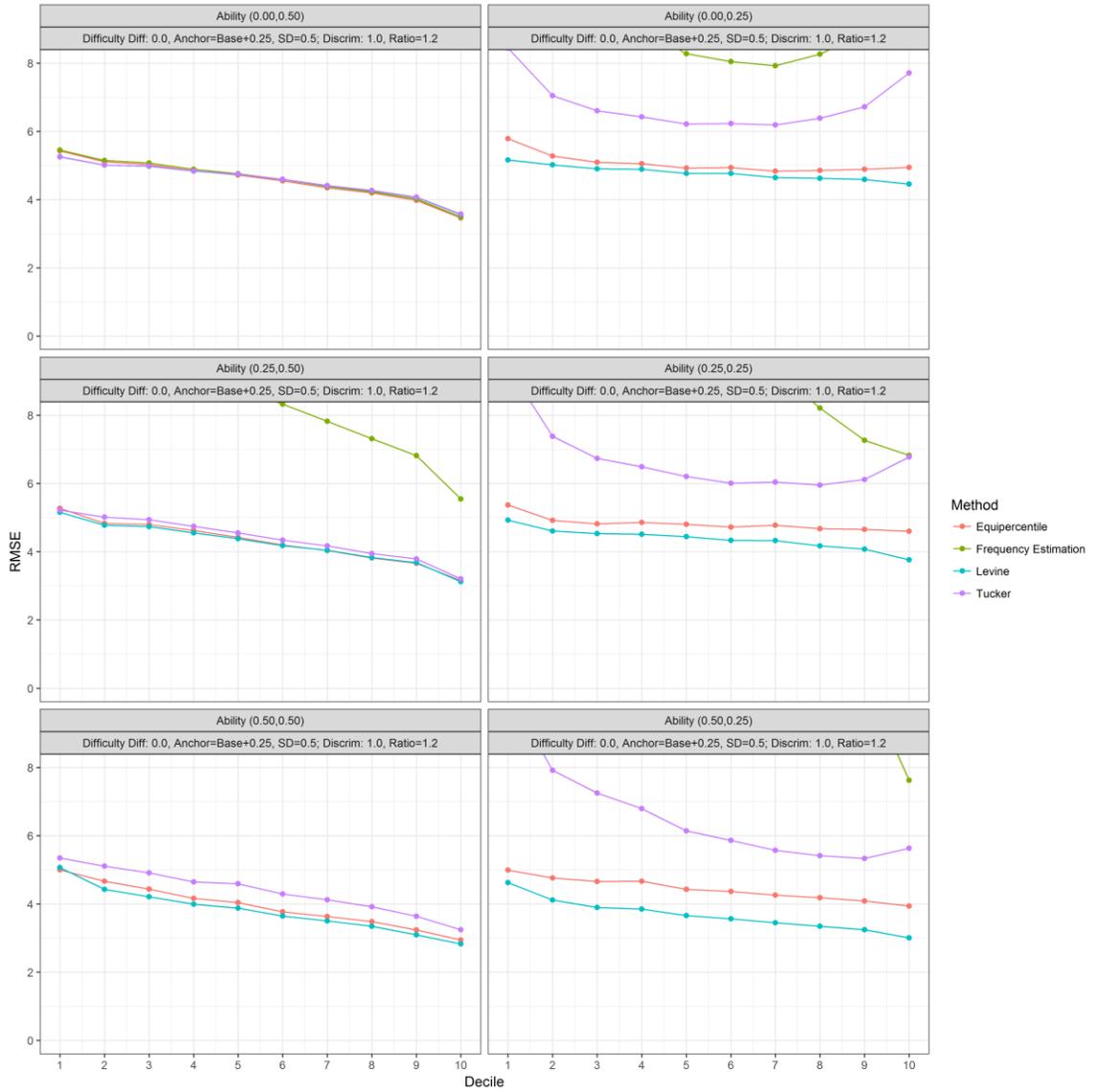


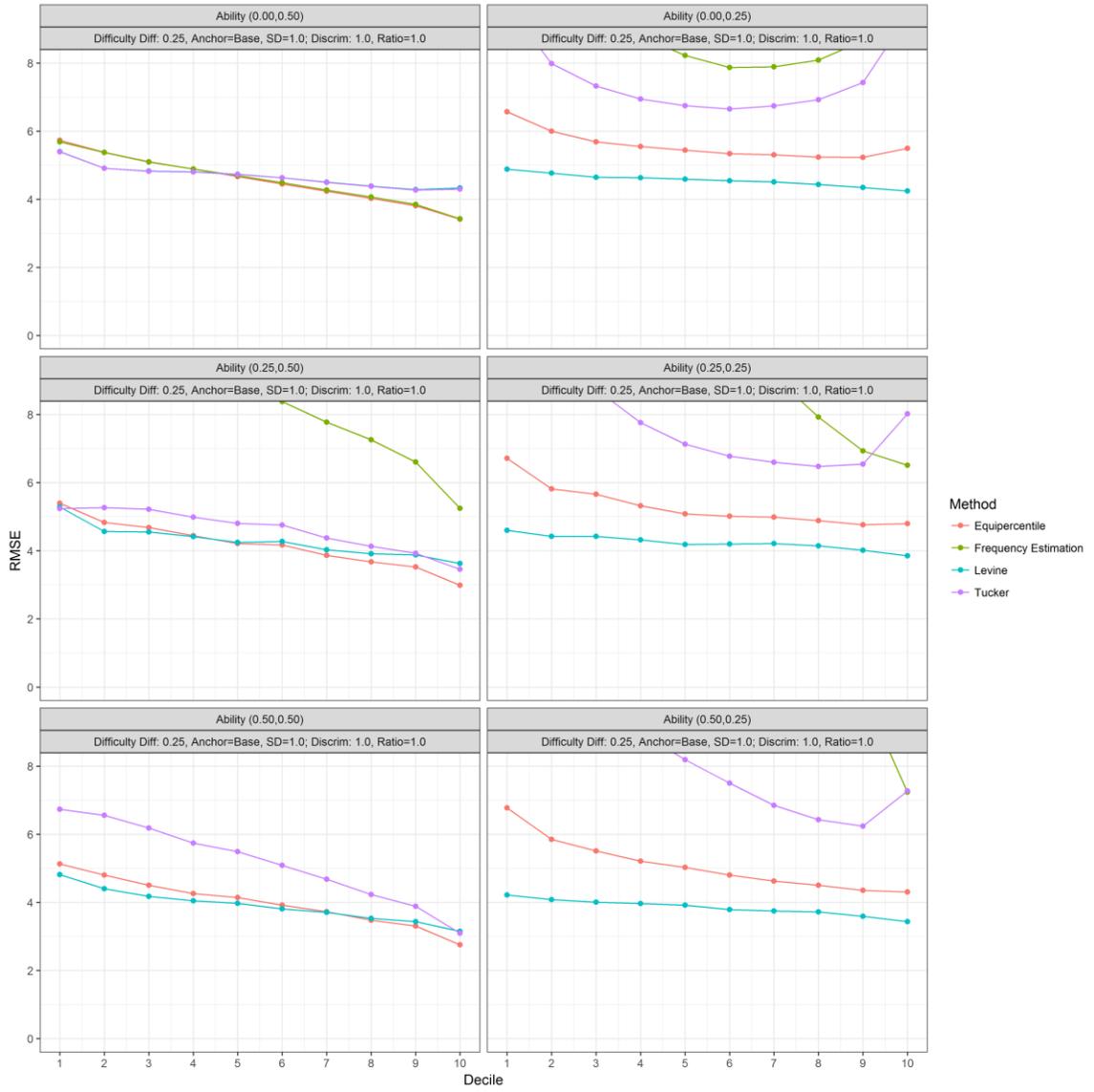


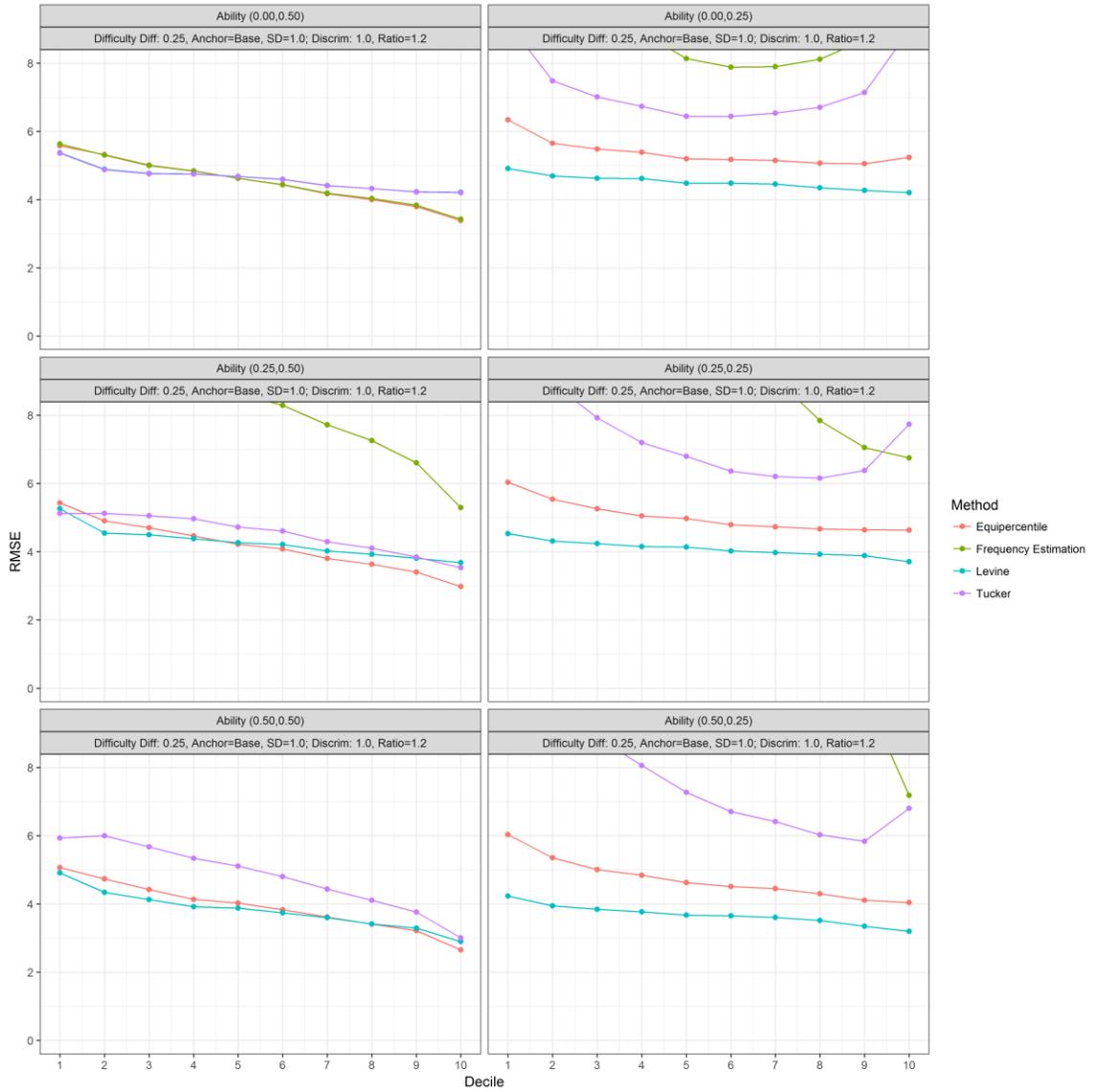


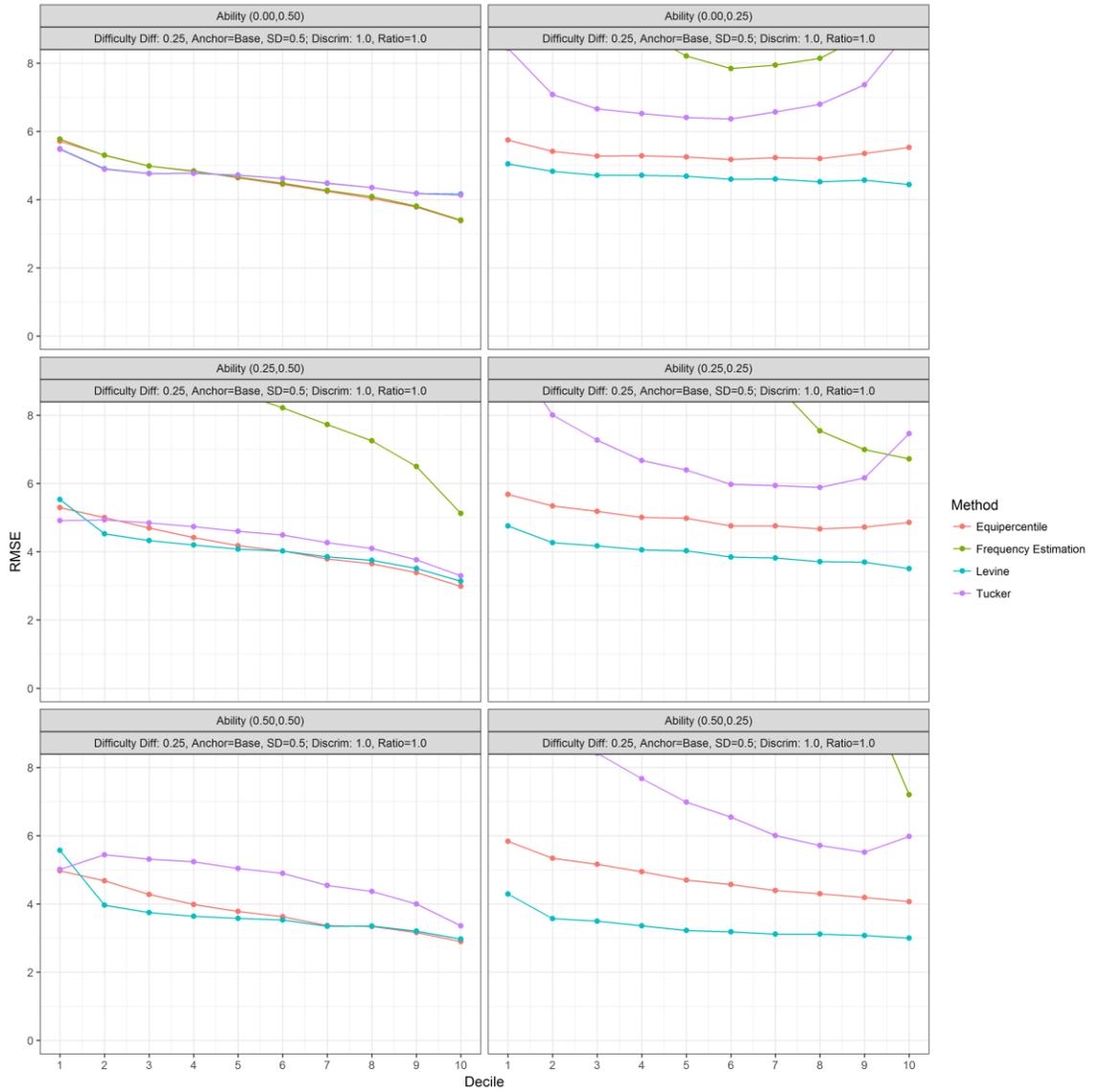


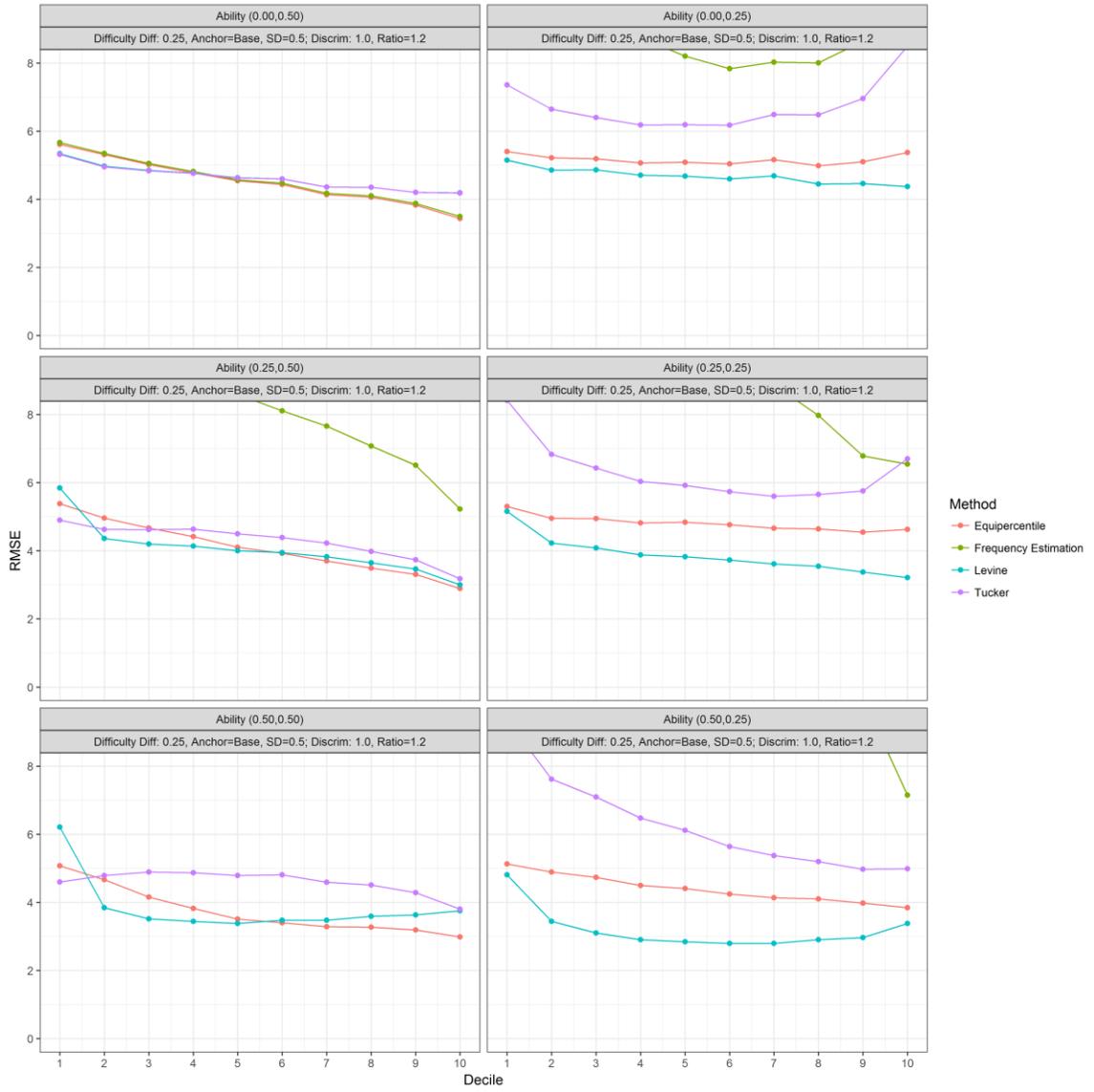


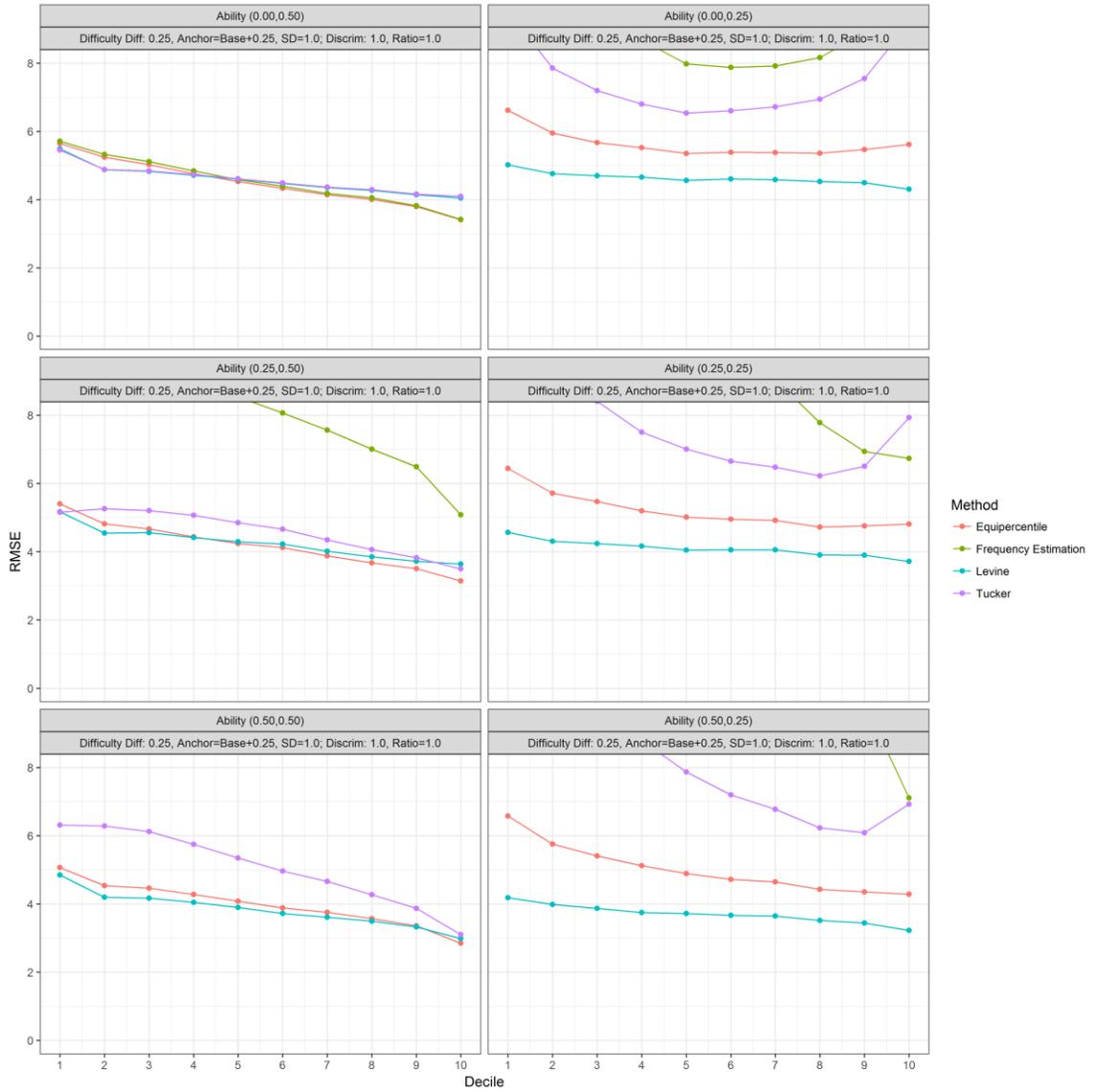


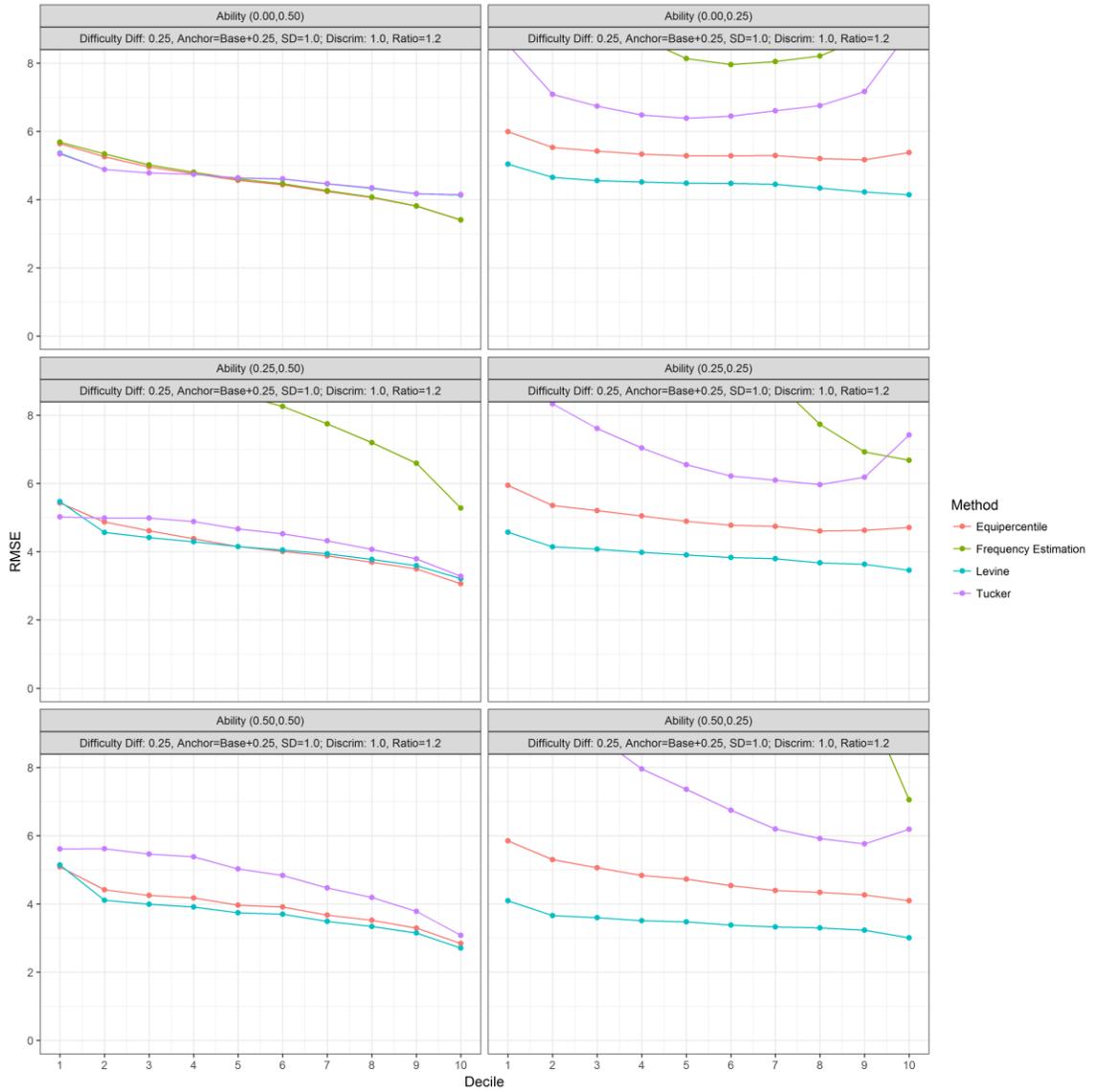


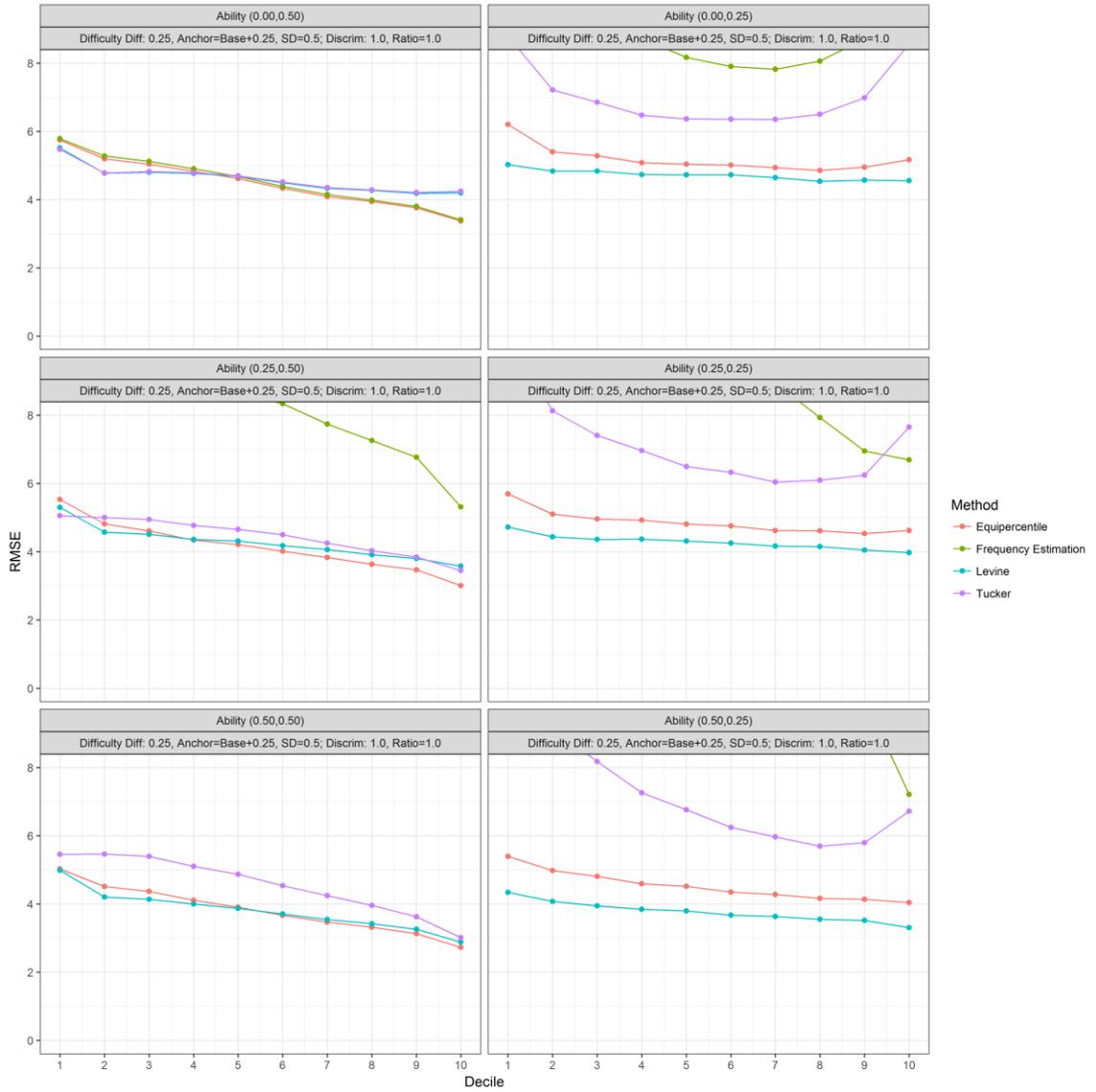


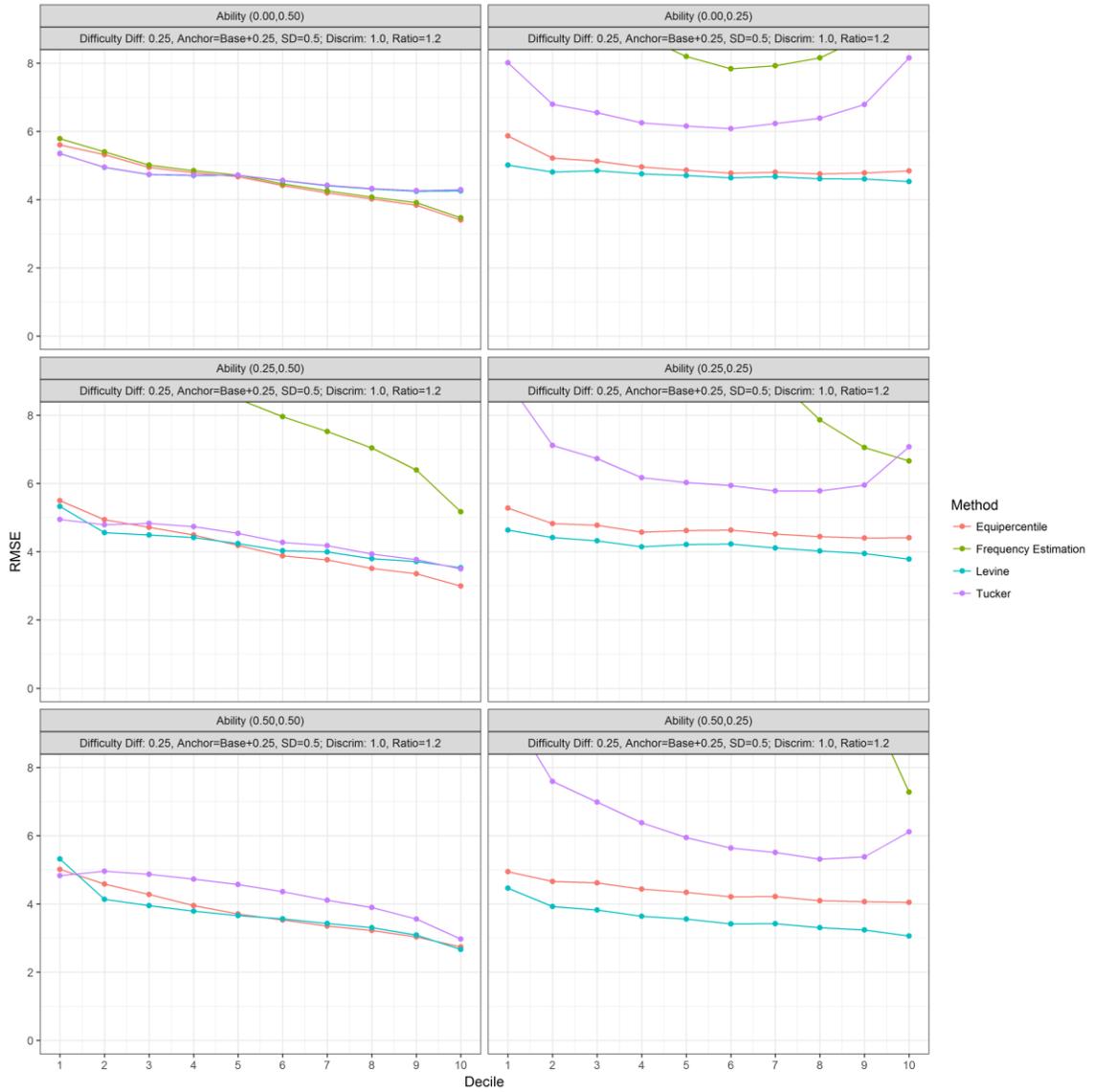


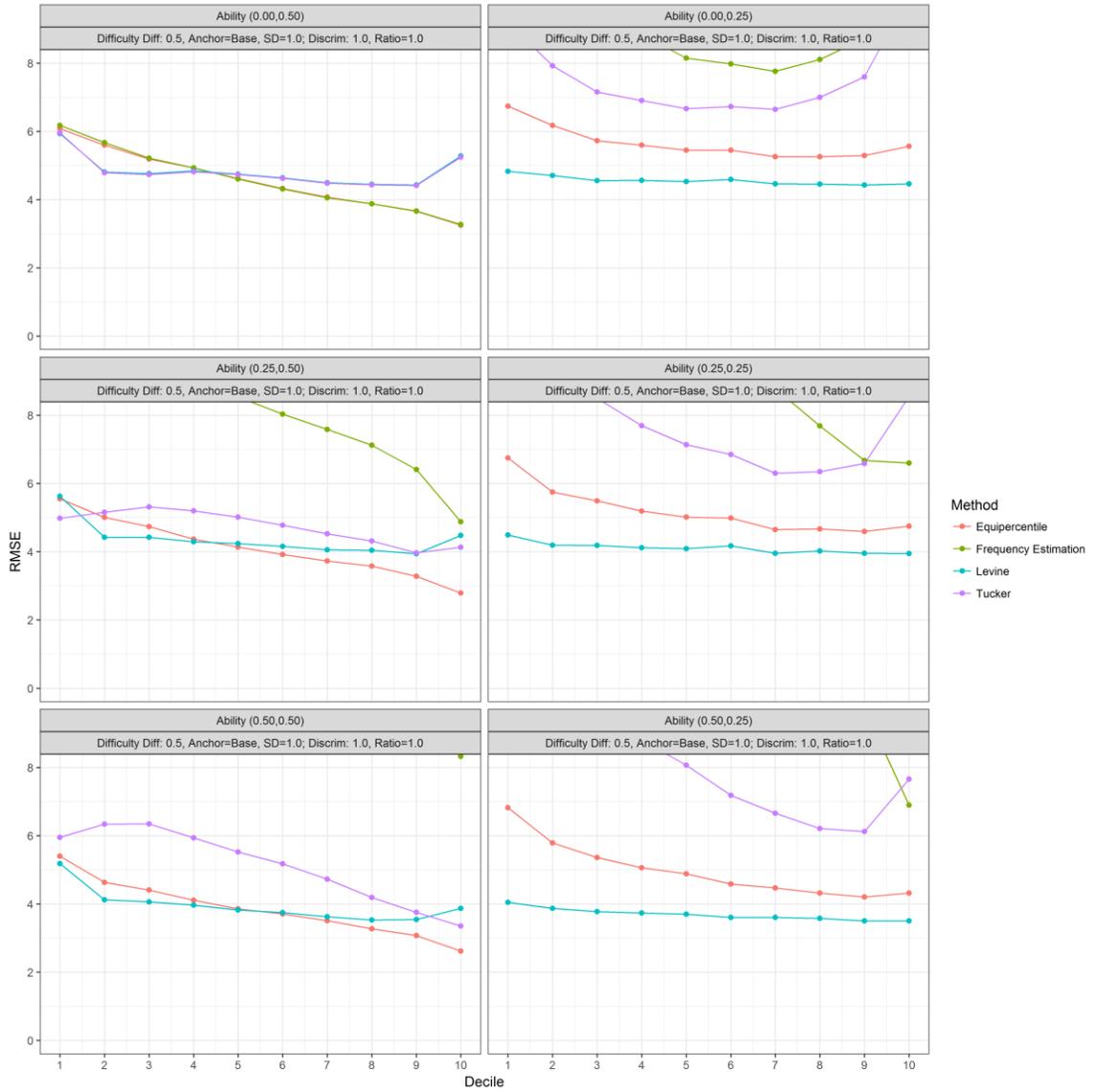


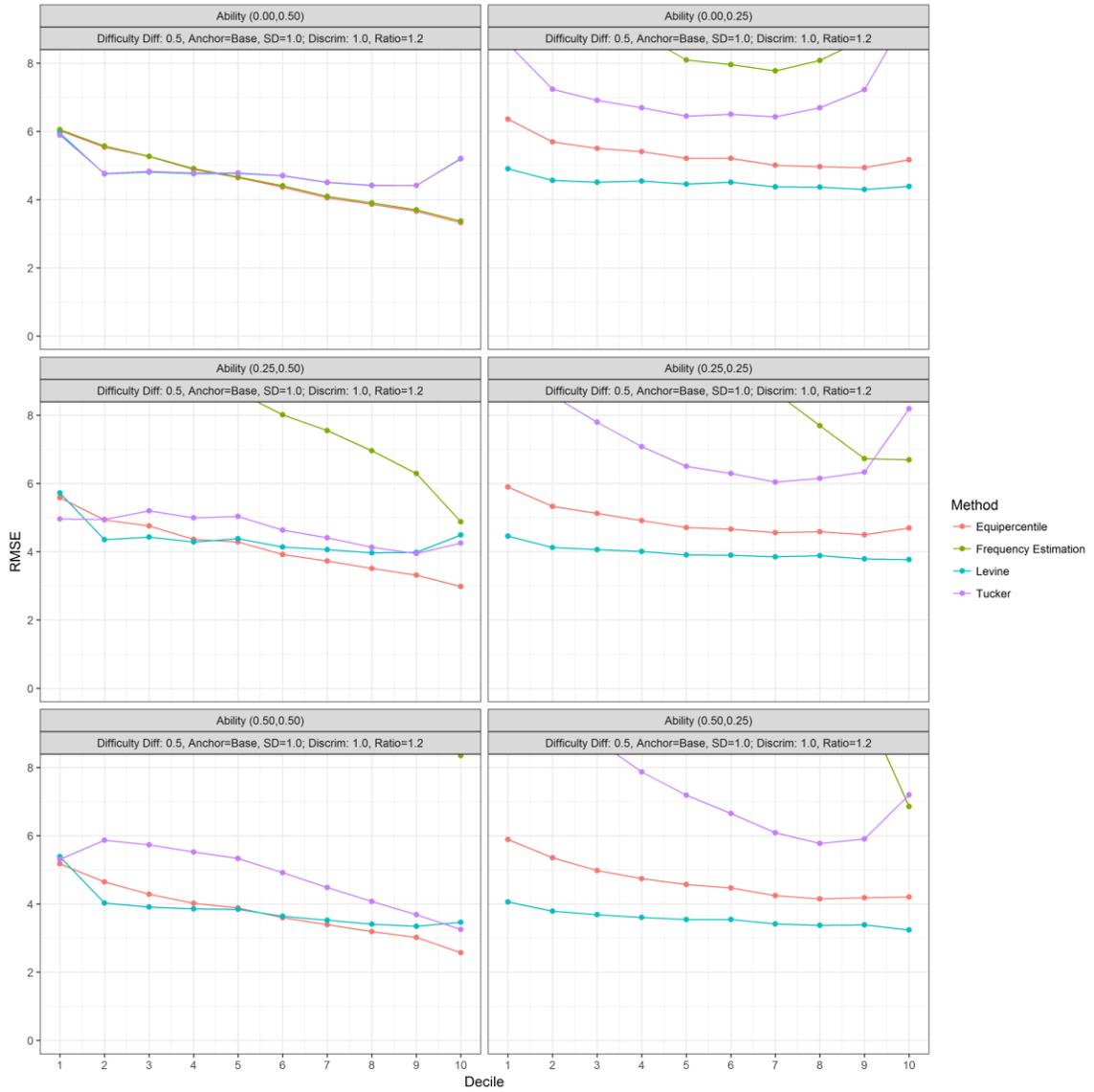


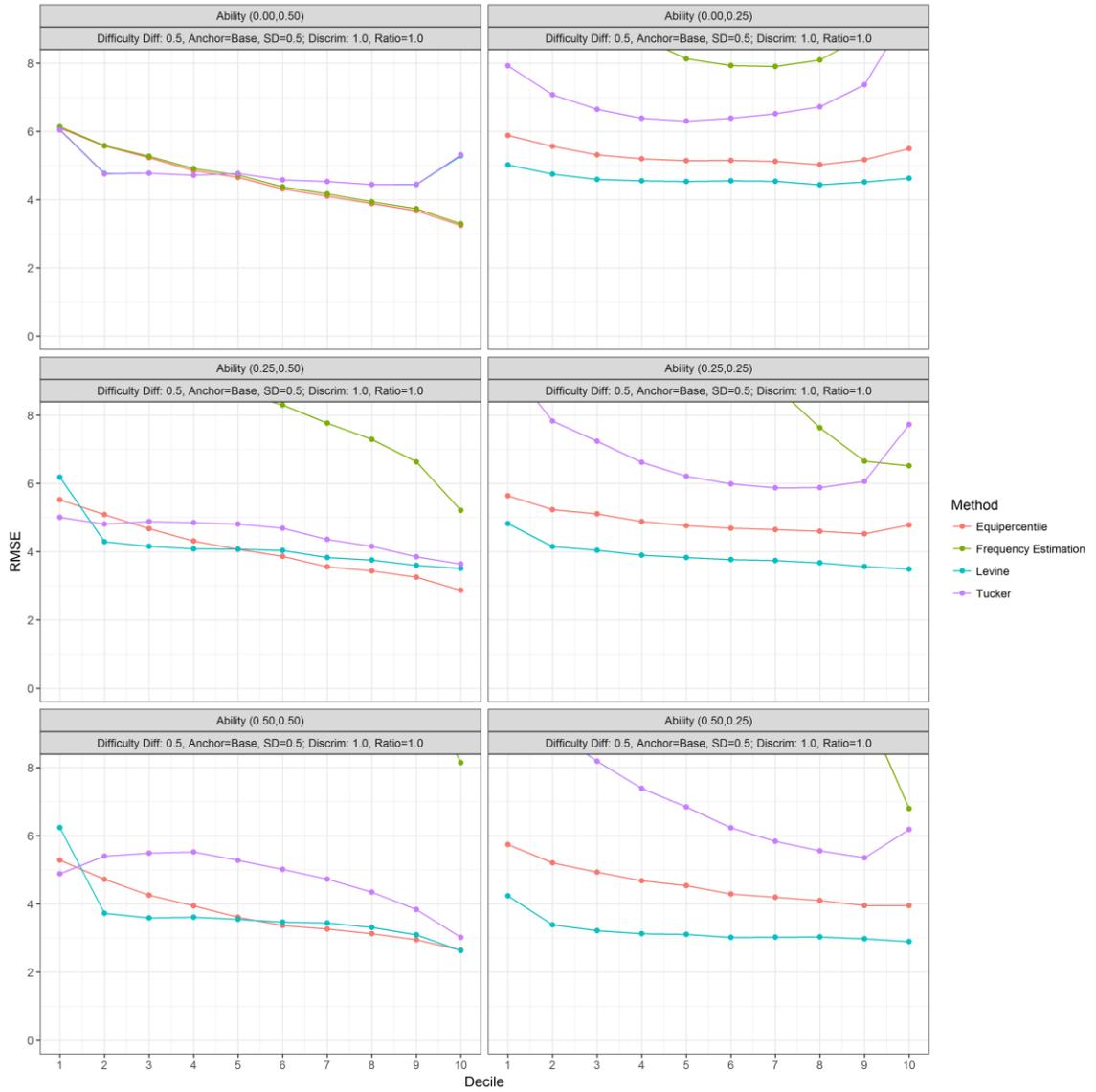


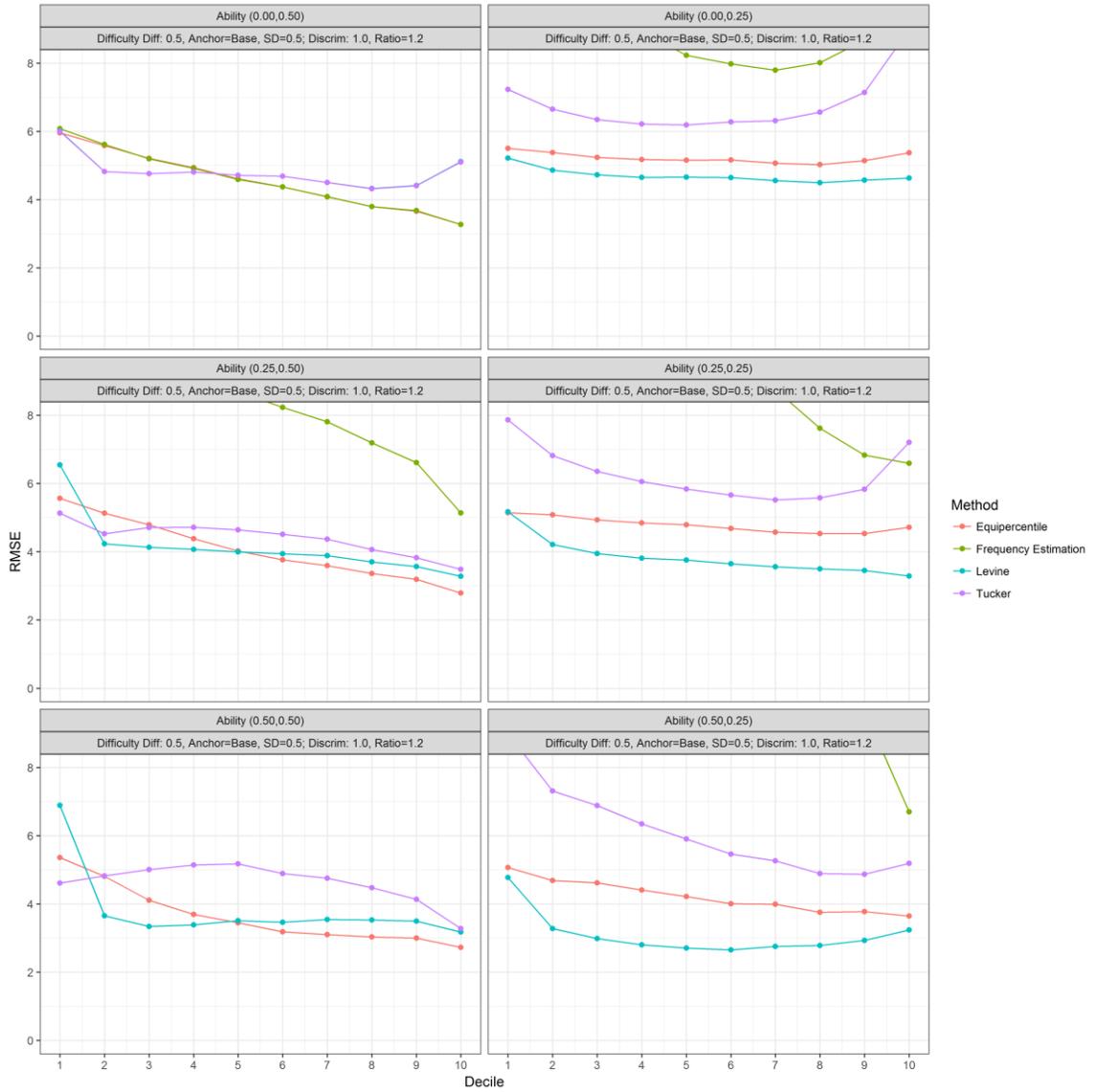


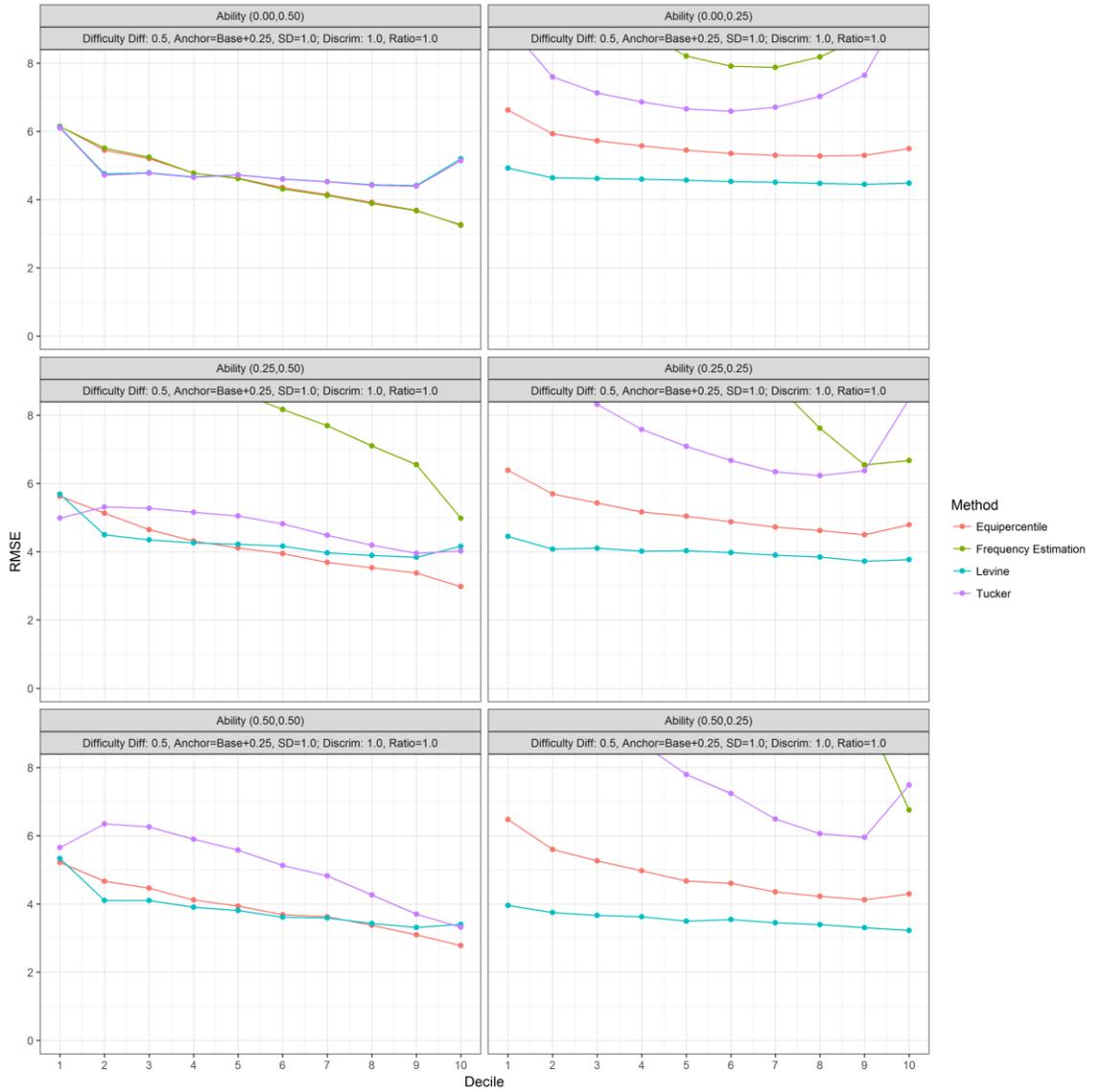


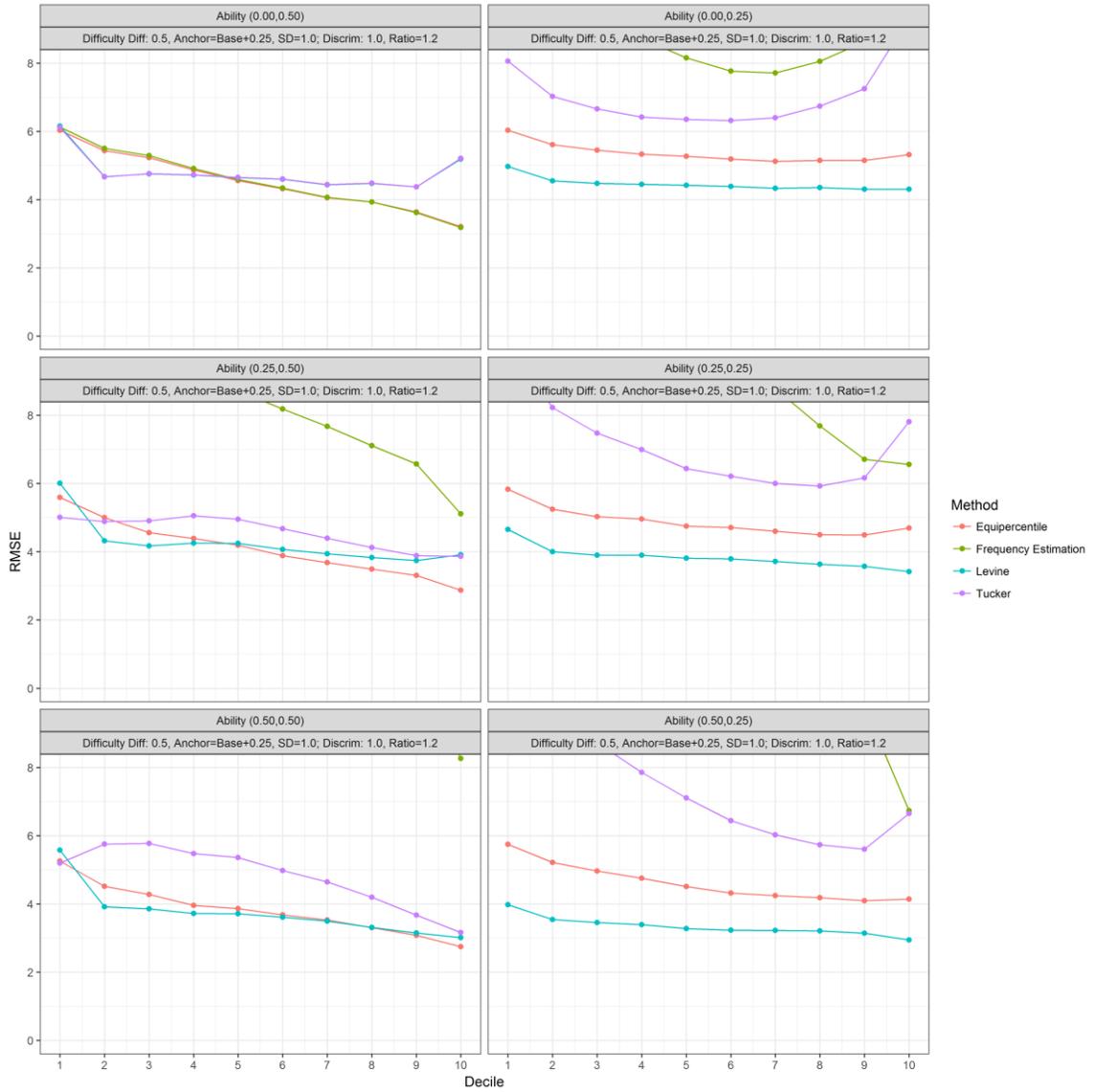


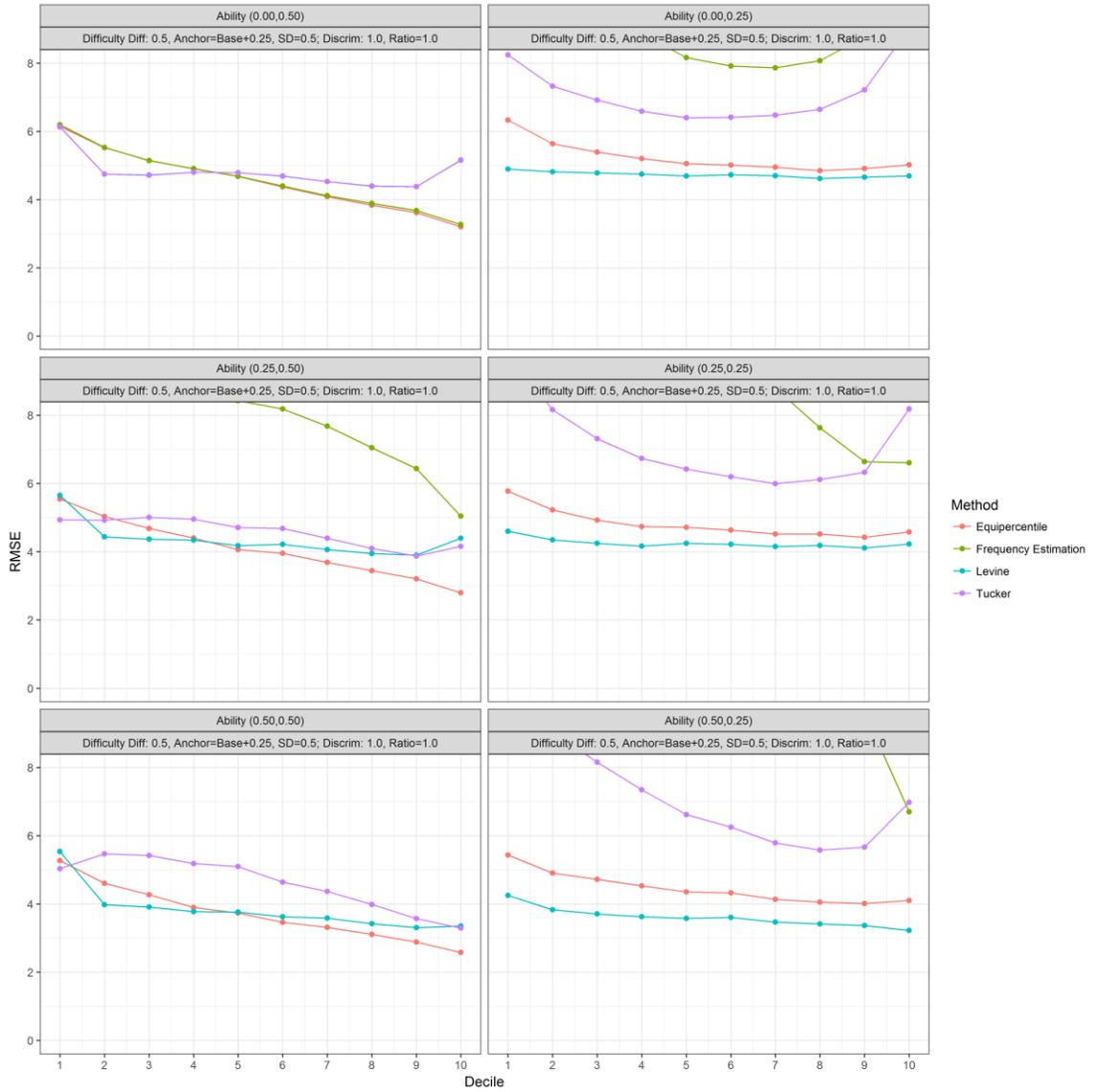


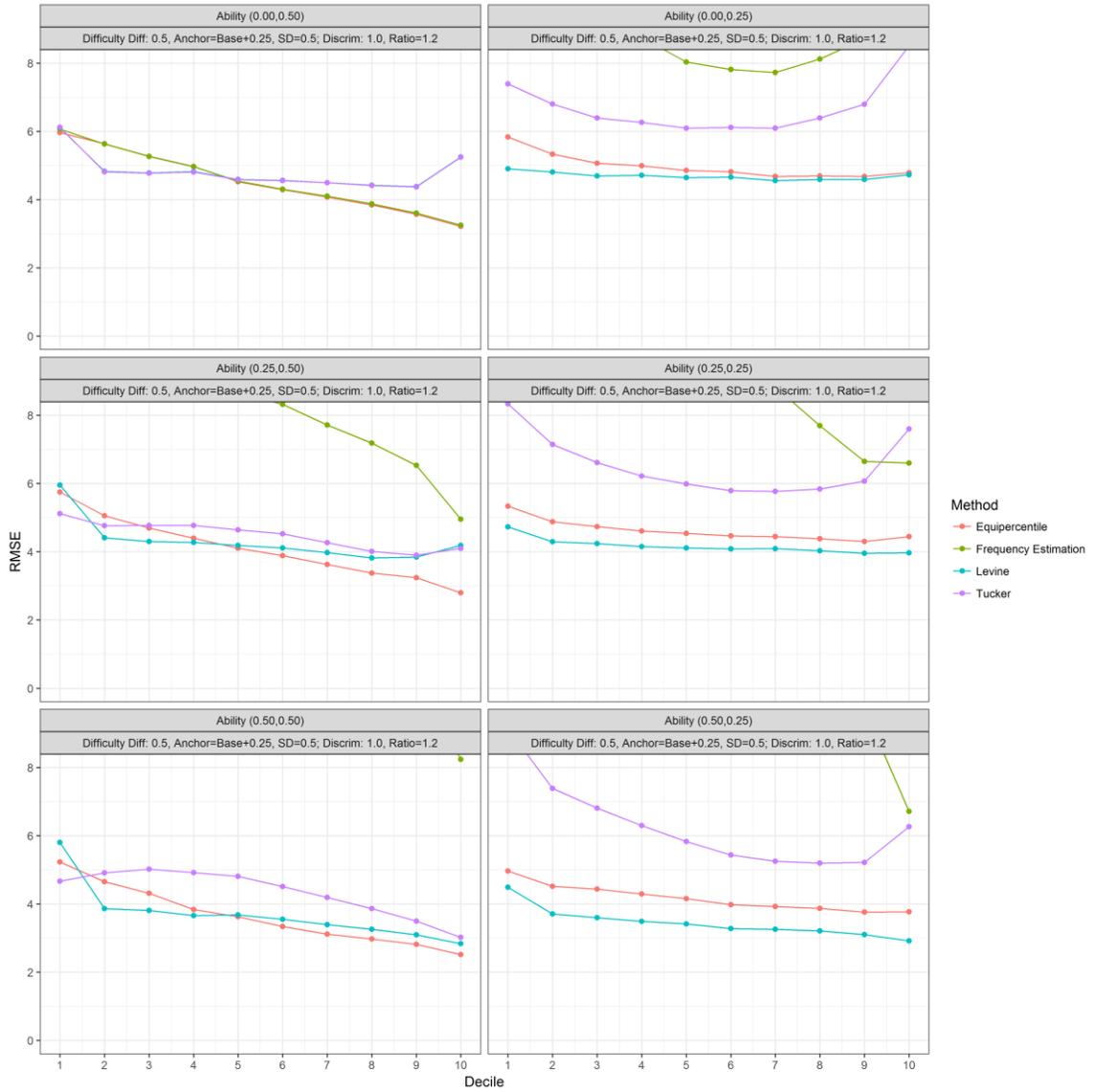












## APPENDIX E

### ACHIEVEMENT TOTAL AND ANCHOR SCORE CORRELATIONS

Base	Alt	Ability Conditions	Test Conditions
0.7658	0.7667	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7660	0.7709	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7651	0.7722	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7632	0.4977	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7686	0.5088	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7681	0.5087	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.8071	0.8074	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8063	0.8098	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8061	0.8102	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8055	0.5541	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8058	0.5587	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8075	0.5625	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7895	0.7920	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7949	0.7902	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7891	0.7903	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7928	0.5333	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7906	0.5386	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7913	0.5350	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.8230	0.8245	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8246	0.8253	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8248	0.8217	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8240	0.5866	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8230	0.5873	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8244	0.5867	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7714	0.7764	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7723	0.7776	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7745	0.7810	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7735	0.4997	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7728	0.5160	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7709	0.5212	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7955	0.7977	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7976	0.8019	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8003	0.8054	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8003	0.5429	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7991	0.5530	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2

Base	Alt	Ability Conditions	Test Conditions
0.7962	0.5502	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7866	0.7885	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7865	0.7917	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7864	0.7933	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7875	0.5290	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7858	0.5355	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7873	0.5426	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.8168	0.8185	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8193	0.8258	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8157	0.8273	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8185	0.5811	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8194	0.5856	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8179	0.5880	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7696	0.7639	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7666	0.7703	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7679	0.7730	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7694	0.4964	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7659	0.5009	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7660	0.5022	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.8071	0.8018	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8053	0.8086	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8050	0.8062	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8049	0.5476	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8039	0.5562	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8041	0.5602	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7930	0.7877	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7895	0.7898	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7935	0.7918	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7922	0.5288	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7912	0.5334	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7915	0.5333	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.8257	0.8204	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8232	0.8228	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8216	0.8229	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8249	0.5851	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8237	0.5871	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8233	0.5851	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7693	0.7734	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7762	0.7797	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0

Base	Alt	Ability Conditions	Test Conditions
0.7716	0.7796	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7705	0.4988	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7716	0.5162	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7714	0.5270	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7990	0.7964	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7978	0.8042	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8000	0.8051	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7963	0.5393	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7982	0.5579	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7982	0.5531	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7864	0.7870	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7861	0.7912	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7872	0.7939	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7861	0.5294	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7869	0.5372	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7872	0.5404	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.8177	0.8171	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8182	0.8232	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8183	0.8247	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8200	0.5800	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8186	0.5848	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8189	0.5926	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7693	0.7628	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7653	0.7655	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7663	0.7755	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7673	0.4909	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7668	0.5023	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7654	0.4980	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.8070	0.7981	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8055	0.8020	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8063	0.8081	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8055	0.5427	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8063	0.5518	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.8050	0.5612	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7922	0.7840	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7925	0.7853	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7901	0.7866	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7932	0.5235	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7922	0.5305	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0

Base	Alt	Ability Conditions	Test Conditions
0.7911	0.5327	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.8233	0.8137	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8241	0.8158	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8258	0.8177	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8227	0.5741	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8228	0.5815	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8228	0.5834	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7711	0.7702	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7722	0.7784	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7703	0.7800	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7724	0.5016	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7742	0.5101	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7745	0.5173	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.8003	0.7934	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7988	0.8002	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7975	0.8024	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7976	0.5345	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7969	0.5463	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7968	0.5501	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7866	0.7836	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7896	0.7857	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7899	0.7913	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7834	0.5202	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7862	0.5355	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7890	0.5342	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.8182	0.8175	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8182	0.8213	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8199	0.8220	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8207	0.5734	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8193	0.5854	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8178	0.5848	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8686	0.8681	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8667	0.8712	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8658	0.8735	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8680	0.6792	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8658	0.6830	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8667	0.6862	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8841	0.8823	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8829	0.8881	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2

Base	Alt	Ability Conditions	Test Conditions
0.8809	0.8896	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8839	0.7072	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8828	0.7157	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8825	0.7167	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8866	0.8849	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8860	0.8862	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8867	0.8863	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8865	0.7151	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8852	0.7237	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8866	0.7171	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8976	0.8971	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8986	0.9004	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8992	0.8999	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8995	0.7519	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8981	0.7545	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8985	0.7486	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8598	0.8593	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8614	0.8655	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8599	0.8677	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8603	0.6611	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8595	0.6719	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8608	0.6744	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8758	0.8770	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8747	0.8822	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8766	0.8839	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8748	0.6961	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8769	0.7079	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8755	0.7060	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8762	0.8803	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8774	0.8852	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8777	0.8872	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8754	0.7064	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8773	0.7197	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8763	0.7217	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8877	0.8923	Base (0,1) Alt (0,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8888	0.8975	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8882	0.9006	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8888	0.7406	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8889	0.7508	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2

Base	Alt	Ability Conditions	Test Conditions
0.8891	0.7550	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8683	0.8644	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8665	0.8703	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8672	0.8721	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8688	0.6719	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8687	0.6829	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8667	0.6811	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8835	0.8804	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8826	0.8855	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8824	0.8867	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8837	0.7053	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8833	0.7131	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8827	0.7130	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8854	0.8817	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8854	0.8828	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8849	0.8815	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8863	0.7136	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8853	0.7142	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8847	0.7142	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8974	0.8939	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8997	0.8962	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8987	0.8929	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8972	0.7453	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8991	0.7520	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8995	0.7458	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8598	0.8606	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8598	0.8644	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8595	0.8683	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8595	0.6619	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8592	0.6708	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8586	0.6775	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8742	0.8743	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8767	0.8817	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8762	0.8839	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8756	0.6927	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8755	0.7048	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8751	0.7117	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8769	0.8799	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8774	0.8842	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0

Base	Alt	Ability Conditions	Test Conditions
0.8779	0.8873	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8761	0.7023	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8772	0.7228	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8778	0.7197	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8896	0.8935	Base (0,1) Alt (0,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8876	0.8988	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8893	0.9019	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8892	0.7377	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8888	0.7491	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8872	0.7544	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8692	0.8592	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8681	0.8640	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8675	0.8659	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8674	0.6685	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8667	0.6731	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8655	0.6789	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8835	0.8730	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8830	0.8776	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8831	0.8805	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8821	0.6981	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8830	0.7120	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8823	0.7125	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8868	0.8741	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8856	0.8750	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8859	0.8728	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8856	0.7036	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8842	0.7128	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8867	0.7085	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8983	0.8853	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8985	0.8855	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8991	0.8836	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8985	0.7365	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8976	0.7405	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8987	0.7367	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8574	0.8538	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8605	0.8633	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8579	0.8653	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8605	0.6544	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8584	0.6627	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0

Base	Alt	Ability Conditions	Test Conditions
0.8601	0.6684	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8757	0.8724	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8754	0.8763	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8757	0.8812	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8753	0.6867	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8745	0.6995	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8772	0.7031	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8792	0.8762	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8759	0.8793	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8758	0.8831	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8780	0.7001	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8779	0.7150	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8747	0.7161	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8887	0.8884	Base (0,1) Alt (0,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8895	0.8933	Base (0,1) Alt (0.25,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8891	0.8948	Base (0,1) Alt (0.5,1)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8889	0.7311	Base (0,1) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8885	0.7481	Base (0,1) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8910	0.7506	Base (0,1) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2

## APPENDIX F

### CERTIFICATION TOTAL AND ANCHOR SCORE CORRELATIONS

Base	Alt	Ability Conditions	Test Conditions
0.7156	0.7126	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7190	0.7098	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7146	0.6970	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7145	0.4114	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7182	0.4017	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7177	0.3852	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7377	0.7381	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7375	0.7312	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7383	0.7205	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7327	0.4279	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7353	0.4209	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7387	0.4073	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7311	0.7332	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7305	0.7228	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7325	0.7103	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7277	0.4255	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7296	0.4121	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7323	0.3915	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7679	0.7687	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7679	0.7570	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7712	0.7445	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7729	0.4697	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7709	0.4550	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7687	0.4306	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7118	0.7110	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7133	0.7078	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7109	0.6963	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7112	0.3990	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7114	0.4057	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7109	0.3908	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7436	0.7468	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7457	0.7351	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7441	0.7278	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7464	0.4472	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7465	0.4329	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2

Base	Alt	Ability Conditions	Test Conditions
0.7454	0.4077	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7312	0.7327	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7322	0.7294	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7344	0.7184	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7342	0.4331	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7334	0.4163	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7310	0.4098	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7714	0.7732	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7720	0.7682	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7707	0.7582	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7719	0.4874	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7732	0.4638	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7717	0.4467	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7112	0.7173	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7139	0.7113	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7160	0.6998	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7171	0.4062	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7165	0.4109	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7187	0.3946	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7403	0.7376	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7358	0.7306	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7387	0.7184	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7382	0.4343	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7378	0.4270	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7374	0.4147	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7317	0.7311	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7277	0.7258	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7321	0.7126	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7261	0.4229	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7329	0.4113	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7288	0.4013	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7707	0.7690	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7680	0.7545	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7672	0.7456	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7692	0.4776	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7674	0.4610	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7651	0.4429	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7128	0.7145	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7145	0.7102	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0

Base	Alt	Ability Conditions	Test Conditions
0.7134	0.7049	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7080	0.4084	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7099	0.4081	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7109	0.3926	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7465	0.7483	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7460	0.7421	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7479	0.7331	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7439	0.4455	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7456	0.4351	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7463	0.4214	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7349	0.7337	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7317	0.7336	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7345	0.7245	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7337	0.4409	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7349	0.4206	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7328	0.4112	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7718	0.7740	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7730	0.7647	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7745	0.7579	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7711	0.4794	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7731	0.4733	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7742	0.4541	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7146	0.7148	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7171	0.7082	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7171	0.7043	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7140	0.4124	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7172	0.4048	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7147	0.4033	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7353	0.7371	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7363	0.7288	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7360	0.7200	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7387	0.4353	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7360	0.4249	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7382	0.4186	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7267	0.7295	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7300	0.7225	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7300	0.7140	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7274	0.4285	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7305	0.4182	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0

Base	Alt	Ability Conditions	Test Conditions
0.7289	0.4012	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7679	0.7669	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7658	0.7575	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7678	0.7444	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7696	0.4775	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7694	0.4631	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7710	0.4470	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7119	0.7121	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7083	0.7130	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7089	0.7099	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7091	0.3954	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7118	0.4009	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7156	0.3975	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.0
0.7456	0.7467	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7479	0.7390	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7440	0.7278	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7462	0.4430	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7469	0.4383	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7472	0.4313	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 0.6, Ratio=1.2
0.7319	0.7353	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7314	0.7326	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7346	0.7253	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7353	0.4309	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7345	0.4290	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7372	0.4221	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.0
0.7703	0.7706	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7719	0.7709	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7721	0.7588	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7752	0.4802	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7749	0.4764	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.7700	0.4577	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 0.6, Ratio=1.2
0.8220	0.8217	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8222	0.8148	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8228	0.8060	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8225	0.5580	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8231	0.5466	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8215	0.5244	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8436	0.8425	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8430	0.8353	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2

Base	Alt	Ability Conditions	Test Conditions
0.8458	0.8270	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8448	0.5942	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8442	0.5802	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8445	0.5616	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8544	0.8536	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8553	0.8420	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8540	0.8259	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8528	0.6141	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8508	0.5860	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8540	0.5457	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8720	0.8724	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8726	0.8601	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8736	0.8397	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8734	0.6544	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8728	0.6163	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8716	0.5758	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8291	0.8285	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8331	0.8167	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8290	0.8069	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8281	0.5664	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8302	0.5539	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8258	0.5190	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8477	0.8469	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8474	0.8376	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8503	0.8264	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8508	0.5965	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8491	0.5693	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8513	0.5414	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8594	0.8601	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8574	0.8528	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8568	0.8426	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8582	0.6234	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8571	0.6093	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8579	0.5803	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8796	0.8792	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8796	0.8733	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8797	0.8611	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8808	0.6720	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8796	0.6493	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2

Base	Alt	Ability Conditions	Test Conditions
0.8785	0.6118	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.0, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8241	0.8213	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8235	0.8192	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8202	0.8052	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8224	0.5589	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8219	0.5541	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8227	0.5273	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8431	0.8457	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8454	0.8382	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8450	0.8309	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8436	0.5986	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8460	0.5848	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8425	0.5624	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8539	0.8549	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8528	0.8384	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8545	0.8244	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8556	0.6161	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8529	0.5962	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8521	0.5557	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8731	0.8710	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8719	0.8548	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8728	0.8345	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8722	0.6576	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8718	0.6246	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8719	0.5824	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8281	0.8289	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8279	0.8186	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8270	0.8120	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8299	0.5784	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8256	0.5537	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8286	0.5319	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8485	0.8458	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8488	0.8377	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8487	0.8291	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8481	0.6013	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8485	0.5816	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8501	0.5563	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8602	0.8591	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8584	0.8553	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0

Base	Alt	Ability Conditions	Test Conditions
0.8577	0.8452	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8585	0.6399	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8595	0.6131	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8585	0.5899	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8799	0.8770	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8787	0.8717	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8789	0.8588	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8798	0.6767	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8787	0.6545	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8802	0.6211	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.25, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8218	0.8195	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8219	0.8139	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8208	0.8092	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8206	0.5603	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8218	0.5579	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8221	0.5374	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8447	0.8408	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8429	0.8354	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8439	0.8277	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8437	0.5994	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8456	0.5881	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8427	0.5705	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8538	0.8452	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8533	0.8379	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8545	0.8172	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8527	0.6170	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8535	0.5955	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8552	0.5640	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8722	0.8605	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8719	0.8496	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8727	0.8300	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8727	0.6552	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8729	0.6292	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8734	0.5898	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8285	0.8268	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8302	0.8177	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8300	0.8087	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8314	0.5688	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8264	0.5532	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0

Base	Alt	Ability Conditions	Test Conditions
0.8286	0.5343	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.0
0.8482	0.8419	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8467	0.8380	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8472	0.8243	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8479	0.6068	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8477	0.5848	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8478	0.5684	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=1.0; Discrim: 1.0, Ratio=1.2
0.8581	0.8570	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8590	0.8514	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8589	0.8400	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8573	0.6288	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8577	0.6223	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8577	0.5889	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.0
0.8775	0.8751	Base (0,0.5) Alt (0,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8785	0.8677	Base (0,0.5) Alt (0.25,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8787	0.8551	Base (0,0.5) Alt (0.5,0.5)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8770	0.6773	Base (0,0.5) Alt (0,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8784	0.6537	Base (0,0.5) Alt (0.25,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2
0.8800	0.6303	Base (0,0.5) Alt (0.5,0.25)	Difficulty Diff: 0.5, Anchor=Base+0.25, SD=0.5; Discrim: 1.0, Ratio=1.2