CHEN, DAVID FREDERICK, Ph.D. Impact of Item Parameter Drift on IRT Linking Methods. (2021)
Directed by Drs. Kyung Yong Kim and John Willse. 339 pp.

Item parameter drift is a severe threat to testing programs that need to ensure fair and comparable scores between different forms of the same test. This study examines the effect of drift on simulated and empirical data sets using the following five IRT linking methods: Stocking-Lord, Haebara, least absolute values, concurrent calibration, and fixed parameter calibration. Four factors were varied: the proportion of drifted items, the magnitude of drifted items, examinee ability distributions, and sample size. The least absolute values method was best at recovering linking constant B, difficulty estimates, and equated true and observed scores. Concurrent calibration and fixed parameter calibration most accurately recovered linking constant A and discrimination estimates. All linking methods provided similar classification accuracy and consistency rates. However, the profound impact of drift has the potential to affect equated scores even at lower magnitudes of drift because of its impact on the linking constants and item parameter estimates that precede equating. Practitioners should remove drifted items when possible and investigate the reason for drift to prevent future reoccurrences. Recommendations for identifying reasons for drift and accumulating evidence for validation when confronted with drift are discussed.

IMPACT OF ITEM PARAMETER DRIFT

ON IRT LINKING METHODS

by

David Frederick Chen

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
2021

Approved by

_____
Committee Co-Chair

_____
Committee Co-Chair

APPROVAL PAGE

This dissertation written by David Frederick Chen, has been approved by the following committee of the Faculty of The Graduate School at the University of North Carolina at Greensboro.

Committee Co-Chair_____

Committee Co-Chair_____

Committee Members_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

I want to thank my dissertation committee of Dr. Kyung Yong Kim, Dr. John Willse, Dr. Micheline Chalhoub-Deville, and Dr. Bob Henson for their time and help throughout this endeavor. To Dr. Kim, thank you for your patience, feedback, and responsiveness to all of my questions. To Dr. Willse, thank you for your guidance and mentorship during my time in the program. To Dr. Chalhoub-Deville, thank you for your invaluable professional advice, technical expertise, and encouragement. Lastly to Dr. Henson, no one is better at making dense subject matter seem easier than it really is, and I thank you for your calming demeanor. I am so lucky and grateful to have learned from such great minds in the field.

I want to thank my mother, who has been my biggest advocate and supporter my entire life. Your constant love and support have helped me get to where I am today. Thank you Katherine, for your unwavering help and support. I'm excited to see what comes next for us. Finally, thank you to everyone who I did not have a chance to acknowledge, but that have contributed greatly to my success.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER

I. INTRODUCTION 1

Overview of Item Response Theory 2
Unidimensional Item Response Theory 2
Item Parameter Drift 6
Purpose 9
Current Study and Research Questions 10

II. LITERATURE REVIEW 12

IRT Equating 12
True Score Equating 12
Observed Score Equating 13
Common-Item Nonequivalent Groups (CINEG) Design 14
Unidimensional IRT Linking Methods 15
Mean-Sigma 17
Mean-Mean 17
Haebara 18
Stocking-Lord 18
Least Absolute Values 19
Concurrent Calibration 20
Fixed Parameter Calibration 20
Comparison of Unidimensional IRT Linking Methods 21
Item Parameter Drift 27
Reasons for IPD 28
IPD Implications for Validity and Validation 32
Comparison of Unidimensional IRT Linking Methods under IPD 43
IRT Robustness to IPD? 43
Linking Method Studies under IPD 47

III. METHODS 66

Overview 66
Simulation Design 66

Data Generation 66
Linking Methods 68
Conditions 68
Evaluation Criteria 72
Empirical Data Analysis 77
Evaluation Criteria 78

IV. RESULTS 79

Simulation Study 79
Drift Detection 79
Linking Constants 83
Linked Item Parameter Estimates 105
Equated Scores 128
Classification Accuracy 173
Classification Consistency 183
Empirical Analysis 198
Drift Detection 198
Linking Constants 200
Linked Item Parameter Estimates 200
Equated Scores 202
Classification 206
Validation Implications 208
Addressing Drift Using Kane's Argument-Based Approach 214
Addressing Drift Using the Standards 219
Extending Drift to Different Testing Contexts 223

V. DISCUSSION 226

Study Findings and Conclusions 226
Drift Detection 229
Research Question 1: Linking Constants 230
Research Question 2: Linked Item Parameter Estimates 231
Research Question 3: Equated Scores 232
Research Question 4: Classification Accuracy 234
Research Question 5: Classification Consistency 234
Empirical Analysis 235
Implications for Validity and Validation 235
Limitations and Directions for Future Research 236

REFERENCES 239

APPENDIX A. GENERATING ITEM PARAMETERS 255

APPENDIX B. BIAS, SE, RMSE VALUES FOR LINKING CONSTANTS 259

APPENDIX C. BIAS, SE, RMSE VALUES FOR UNIQUE ITEM ESTIMATES 272

APPENDIX D. BIAS, SE, RMSE VALUES FOR ALL ITEM ESTIMATES 291

APPENDIX E. BIAS, SE, RMSE VALUES FOR EQUATED SCORES 310

APPENDIX F. BIAS, SE, RMSE VALUES FOR CLASSIFICATION RATES 323

APPENDIX G. EMPIRICAL ITEM ESTIMATES 336

LIST OF TABLES

Page

Table 1. Potential Reasons and Directionality of IPD 31

Table 2. Simulation Study Conditions 71

Table 3. Descriptive Statistics for Generating Item Parameters 80

Table 4. Drift Detection Results – 1,000 Examinees 82

Table 5. Drift Detection Results – 3,000 Examinees 82

Table 6. Estimated Linking Constant A – 1,000 Examinees 85

Table 7. Estimated Linking Constant A – 3,000 Examinees 90

Table 8. Estimated Linking Constant B – 1,000 Examinees 96

Table 9. Estimated Linking Constant B – 3,000 Examinees 101

Table 10. Classification Accuracy Rates – 1,000 Examinees 175

Table 11. Classification Accuracy Rates – 3,000 Examinees 176

Table 12. Classification Consistency Rates – 1,000 Examinees 185

Table 13. Classification Consistency Rates – 3,000 Examinees 186

Table 14. Summary of Simulation Results – 1,000 Examinees 194

Table 15. Summary of Simulation Results – 3,000 Examinees 196

Table 16. Descriptive Statistics for Test Forms 198

Table 17. Item Estimates for Test Forms 199

Table 18. Drift Detection for Real Data 200

Table 19. Empirical Analysis of Linking Constants 201

Table 20. Empirical Analysis of Linked Item Parameter Estimates 202

Table 21. Empirical Equating Results – Form X Converted to Form Y Scale 203

Table 22. Marginal Classification Accuracy and Consistency Rates 206

LIST OF FIGURES

Page

Figure 1. Kane's Argument Based Approach to Validation 38

Figure 2. Toulmin's Model of Inference Applied to Kane's Scoring Inference 41

Figure 3. Bias Values for Linking Constant A – 1,000 Examinees 86

Figure 4. SE Values for Linking Constant A – 1,000 Examinees 87

Figure 5. RMSE Values for Linking Constant A – 1,000 Examinees 88

Figure 6. Bias Values for Linking Constant A – 3,000 Examinees 91

Figure 7. SE Values for Linking Constant A – 3,000 Examinees 92

Figure 8. RMSE Values for Linking Constant A – 3,000 Examinees 93

Figure 9. Bias Values for Linking Constant B – 1,000 Examinees 97

Figure 10. SE Values for Linking Constant B – 1,000 Examinees 98

Figure 11. RMSE Values for Linking Constant B – 1,000 Examinees 99

Figure 12. Bias Values for Linking Constant B – 3,000 Examinees 102

Figure 13. SE Values for Linking Constant B – 3,000 Examinees 103

Figure 14. RMSE Values for Linking Constant B – 3,000 Examinees 104

Figure 15. Bias Values for Item Estimate a – 1,000 Examinees 106

Figure 16. SE Values for Item Estimate a – 1,000 Examinees 107

Figure 17. RMSE Values for Item Estimate a – 1,000 Examinees 108

Figure 18. Bias Values for Item Estimate a – 3,000 Examinees 110

Figure 19. SE Values for Item Estimate a – 3,000 Examinees 111

Figure 20. RMSE Values for Item Estimate a – 3,000 Examinees 112

Figure 21. Bias Values for Item Estimate b – 1,000 Examinees 115

Figure 22. SE Values for Item Estimate b – 1,000 Examinees.....116

Figure 23. RMSE Values for Item Estimate b – 1,000 Examinees.....117

Figure 24. Bias Values for Item Estimate b – 3,000 Examinees 118

Figure 25. SE Values for Item Estimate b – 3,000 Examinees.....119

Figure 26. RMSE Values for Item Estimate b – 3,000 Examinees.....120

Figure 27. Bias Values for Item Estimate c – 1,000 Examinees.....122

Figure 28. SE Values for Item Estimate c – 1,000 Examinees 123

Figure 29. RMSE Values for Item Estimate c – 1,000 Examinees.....124

Figure 30. Bias Values for Item Estimate c – 3,000 Examinees.....125

Figure 31. SE Values for Item Estimate c – 3,000 Examinees 126

Figure 32. RMSE Values for Item Estimate c – 3,000 Examinees.....127

Figure 33. Bias Values for True Scores – 1,000 Examinees 131

Figure 34. SE Values for True Scores – 1,000 Examinees.....132

Figure 35. RMSE Values for True Scores – 1,000 Examinees.....133

Figure 36. Conditional RMSE for SL True Scores – 1,000 Examinees 136

Figure 37. Conditional RMSE for HB True Scores – 1,000 Examinees 137

Figure 38. Conditional RMSE for LAV True Scores – 1,000 Examinees.....138

Figure 39. Conditional RMSE for CC True Scores – 1,000 Examinees.....139

Figure 40. Conditional RMSE for FPC True Scores – 1,000 Examinees.....140

Figure 41. Bias Values for True Scores – 3,000 Examinees 142

Figure 42. SE Values for True Scores – 3,000 Examinees.....143
Figure 43. RMSE Values for True Scores – 3,000 Examinees.....144
Figure 44. Conditional RMSE for SL True Scores – 3,000 Examinees 146
Figure 45. Conditional RMSE for HB True Scores – 3,000 Examinees 147
Figure 46. Conditional RMSE for LAV True Scores – 3,000 Examinees.....148
Figure 47. Conditional RMSE for CC True Scores – 3,000 Examinees.....149
Figure 48. Conditional RMSE for FPC True Scores – 3,000 Examinees.....150
Figure 49. Bias Values for Observed Scores – 1,000 Examinees.....153
Figure 50. SE Values for Observed Scores – 1,000 Examinees 154
Figure 51. RMSE Values for Observed Scores – 1,000 Examinees.....155
Figure 52. Conditional RMSE for SL Observed Scores – 1,000 Examinees 157
Figure 53. Conditional RMSE for HB Observed Scores – 1,000 Examinees.....158
Figure 54. Conditional RMSE for LAV Observed Scores – 1,000 Examinees.....159
Figure 55. Conditional RMSE for CC Observed Scores – 1,000 Examinees.....160
Figure 56. Conditional RMSE for FPC Observed Scores – 1,000 Examinees.....161
Figure 57. Bias Values for Observed Scores – 3,000 Examinees.....164
Figure 58. SE Values for Observed Scores – 3,000 Examinees 165
Figure 59. RMSE Values for Observed Scores – 3,000 Examinees.....166
Figure 60. Conditional RMSE for SL Observed Scores – 3,000 Examinees 168
Figure 61. Conditional RMSE for HB Observed Scores – 3,000 Examinees.....169
Figure 62. Conditional RMSE for LAV Observed Scores – 3,000 Examinees.....170
Figure 63. Conditional RMSE for CC Observed Scores – 3,000 Examinees.....171

Figure 64. Conditional RMSE for FPC Observed Scores – 3,000 Examinees 172
Figure 65. Bias Values for Classification Accuracy – 1,000 Examinees 177
Figure 66. SE Values for Classification Accuracy – 1,000 Examinees.....178
Figure 67. RMSE Values for Classification Accuracy – 1,000 Examinees.....179
Figure 68. Bias Values for Classification Accuracy – 3,000 Examinees 180
Figure 69. SE Values for Classification Accuracy – 3,000 Examinees.....181
Figure 70. RMSE Values for Classification Accuracy – 3,000 Examinees.....182
Figure 71. Bias Values for Classification Consistency – 1,000 Examinees 187
Figure 72. SE Values for Classification Consistency – 1,000 Examinees.....188
Figure 73. RMSE Values for Classification Consistency – 1,000 Examinees 189
Figure 74. Bias Values for Classification Consistency – 3,000 Examinees 190
Figure 75. SE Values for Classification Consistency – 3,000 Examinees.....191
Figure 76. RMSE Values for Classification Consistency – 3,000 Examinees 192
Figure 77. Linked Item Difficulty Values by Method 201
Figure 78. Empirical Analysis of IRT True Score Equating 204
Figure 79. Empirical Analysis of IRT Observed Score Equating.....205
Figure 80. Conditional Classification Accuracy Rates 207
Figure 81. Conditional Classification Consistency Rates.....207
Figure 82. Recommendations for Addressing Drift.....208
Figure 83. Intended Interpretation of Test Scores for a Licensure Exam 214
Figure 84. Example of Addressing IPD using Kane's Validity Argument.....216

Figure 85. Example of Addressing IPD using the Standards.....220

CHAPTER I

INTRODUCTION

Assessments used for high stakes decisions, such as admission to higher education or qualification for certification, must meet the highest standards of psychometric quality. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014; referred to hereafter as the *Standards*) consider validity to be the most important aspect of developing and evaluating tests. As defined by Messick (1989), validity is an "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). In other words, validity is an argument for the use of test scores for a specific purpose that can be strengthened by support from research and theory.

Alternatively, the strength of a validity argument can also be compromised by sources of construct-irrelevant variance, or variance due to extraneous factors that distort the meaning of test scores (*Standards*). Sources of construct-irrelevant variance take many different forms and can occur during any stage of test construction. For example, administering the same test on paper-and-pencil compared to a computer-based test may lead to differential scores for equally abled examinees due to type of test modality as opposed to examinee ability. Alternatively, for exams requiring scoring by human raters, one rater might assign lower scores than another. These types of construct-irrelevant

variance must be identified and removed to allow for "the comparable and valid interpretation of test scores for all examinees" (*Standards*, p. 63).

The comparability and valid interpretation of test scores is essential for all testing programs because new forms of the same test are routinely developed and administered to increase test security. In the context of item response theory (IRT), in order for scores on different forms to be appropriately compared, the scores must be placed on the same scale. In this dissertation, *linking* is used to describe the process of placing scores, as well as person and item parameter estimates, onto the same scale. There are a variety of linking methods available to psychometricians, although the use of linking is predicated upon several factors. This includes the data collection design (e.g., single group, random group, common-item non-equivalent group) as well as the IRT assumptions that need to be upheld in order to successfully implement linking.

Overview of Item Response Theory

Unidimensional Item Response Theory. Binet and Simon (1905) first laid the foundation for unidimensional IRT, which was further extended in the 1920's (e.g., Thurstone, 1927). Due to the lack of computers and computationally intensive procedures, IRT only started to gain traction when reintroduced by Lord (1980). For dichotomously scored items (i.e., items that are scored either correct or incorrect), IRT uses a mathematical model to express the probability that an examinee answers an item correctly based upon examinee ability and item parameters (i.e., difficulty, discrimination, and pseudo-guessing).

***Assumptions*.** The main advantage of IRT is that of parameter invariance, whereby parameter values remain equal across groups of examinees and measurement conditions (Rupp & Zumbo, 2006). That is, a person with a specific ability, or theta (θ), remains unchanged over different items or tests (i.e., test independent) and item estimates remain invariant over different groups of examinees (i.e., group independent). So, examinee ability can be estimated independently from items, and item parameters can be estimated independently from the ability of examinees (Hambleton & Jones, 1993). When these conditions are met, the IRT property of parameter invariance has been satisfied (Jones, 1960).

Unidimensional IRT requires that a set of items or test measures only one ability. If an item taps into more than one ability, then a multidimensional IRT model is required. For example, a math item that contains a long verbal passage may also be measuring reading ability, which represents a second dimension. When the assumption of unidimensionality is violated, item and person estimates are subjected to bias and may jeopardize the validity of conclusions about an examinee's ability (e.g., Reise et al., 2007).

Similarly related to unidimensionality, local independence assumes that responses to an item are independent from other items given ability. That is, an examinee's response is based upon their level of ability, not on how the examinee responds to another item (De Ayala, 2013). Given a group of items with the same content, or testlet, an examinee might respond to an item based upon their previous response to a similar item. Failure to uphold the assumption of local independence has negative implications for

construct validity and leads to overestimates of reliability (e.g., Sireci et al., 1991; Thissen et al., 1989).

Even though IRT requires strong assumptions that must be upheld (e.g., unidimensionality, local independence, parameter invariance), more valid test score interpretations can be made by more carefully considering content specifications (Linn, 1990). Due to its versatility, unidimensional IRT is widely used today by testing companies for test development, item banking, computer adaptive testing, and equating purposes.

***Data Collection Design*.** Scores can be adequately compared only when the assumptions of IRT are maintained and the appropriate data collection design is implemented. Several data collection designs are available for implementation. The most frequently used data collection design is the common-item non-equivalent groups (CINEG) design (Kolen & Brennan, 2014) because it is most practical for examinees. Under the CINEG design, two test forms are administered to samples from two different populations that differ in their θ distribution. Unlike other approaches (e.g., single group or random group) that require examinees to take two forms of the same test, the CINEG design uses a set of common items shared between forms, also known as anchor items, to separate differences in group ability from differences in form difficulty (Kolen & Brennan, 2014). Although the CINEG is most feasible for examinees, scores from the two forms cannot be directly compared until the scale indeterminacy issue is resolved through linking.

***Scale Indeterminacy*.** In unidimensional IRT, the θ-scale is not fixed to a specific origin or unit of measurement. IRT software programs handle this scale indeterminacy issue by setting the θ-distribution to a standard normal distribution with a mean of 0 and standard deviation of 1. However, estimating item parameters for test forms with two nonequivalent groups in separate calibration runs (i.e., one for each form) will produce item parameter estimates that are on separate scales. While the two θ-scales differ in their origin and unit of measurement, they are linearly related. A linear transformation can be used to place all item parameter estimates onto the same scale through linking.

***Linking Methods*.** A bevy of linking methods are available with the CINEG design, yet all methods can be classified under one of three types: concurrent calibration (CC), fixed parameter calibration (FPC), and separate calibration (SC). For CC, item parameters for multiple test forms are estimated simultaneously with one calibration run. Item parameter estimates are already on the same scale, so no additional linear transformation is required. Although CC benefits from its efficiency, response data from two operational test forms must be available during calibration, which is not often the case because only one form is usually administered at one time.

Under FPC, base form items have been calibrated and unique items from the new form are estimated by fixing the new form common item estimates to those of the base form. Similar to CC, FPC does not require a linear transformation because item parameter estimates are already on the same scale. This design is commonly used in practice, when items are field-tested prior to being used as scored items.

With SC, two test form parameter estimates are independently calibrated and then linked together via linear transformation. Unlike CC and FPC, SC requires a linear transformation to place the item estimates from one form onto the scale of the other form. Although SC requires an extra step, SC can be used to examine item parameter drift among the common items because it produces two sets of item parameter estimates.

Item Parameter Drift

Although the CINEG design is widely used by testing companies, it requires that the IRT property of parameter invariance hold for each of the common items. If the assumption of parameter invariance is violated, items may begin to function differently between subgroups of examinees. When equally abled examinees from different subgroups (e.g., male or female) have different response probabilities to an item, this is referred to as differential item functioning (DIF). Classified by the *Standards* as a threat to fairness and internal structure of the test, DIF studies are carried out to identify items that may be operating differently between subgroups of examinees. Unless there is sufficient justification for why the item is behaving differently, items showing DIF are removed from the scored item set because they represent a source of construct irrelevant variance that jeopardizes the comparability of scores.

When common items function differently over separate testing occasions (Goldstein, 1983) this is referred to as item parameter drift (IPD). IPD is not directly mentioned in the *Standards*, although it is alluded to: “It is important to check that the anchor items function similarly in the forms being equated. Anchor items are often dropped from the anchor if their relative difficulty is substantially different in the forms

being equated" (*Standards*, p. 98). IPD is often considered a special type of DIF (e.g., Babcock & Albano, 2012; Gaertner & Briggs, 2009), operating as a threat to the fairness and validity of examinees and their test scores.

There are a number of reasons that could lead to IPD including: item overexposure, changes in test curriculum or classroom instruction, cheating, a security breach, test-taking strategies, advances in technology, and current news. For example, test-takers that become exposed to common items will have prior knowledge that benefits them, while unfairly penalizing other test-takers without prior knowledge. Because the exposure of an item does not reflect the actual latent ability of a test taker, but instead, an extraneous factor outside of the construct being measured, it is considered a source of construct-irrelevant variance. As a result, the test would be considered unfair, and a detriment to validity, because the test unfairly advantages examinees with prior knowledge and disadvantages examinees without prior knowledge. Another possibility is that the drift may occur because the initial calibration was poor or contained a different population of test-takers (e.g., first-time new graduates versus retest-takers). Items may drift easier or harder, although most of the reasons presented suggest that items would become easier over time because examinees would benefit by receiving information (fairly or unfairly) that would better prepare them for an item.

Messick (1989) referred to two types of construct-irrelevant variance: construct-irrelevant difficulty and construct-irrelevant easiness. Construct-irrelevant difficulty refers to "aspects of the task that are extraneous to the focal construct make the test irrelevantly more difficult for some individuals or groups" (p. 34). An example is

provided where unnecessary reading comprehension requirements are required for subject-matter knowledge. Construct-irrelevant easiness refers to "when extraneous clues in item or test formats permit some individuals to respond correctly in ways irrelevant to the construct being assessed" (p. 34). Messick uses an example where students pick up on clues when the answer to an item is based upon the longest response stem.

While Messick did not specifically refer to IPD within the context of these types of construct-irrelevant variance, his conceptualization can be extended to IPD. In fact, the example of students picking up clues based upon the length of the response options is an example of test-savviness or a test-taking strategy. Each of the examples of IPD (e.g., changes in instruction, security breach, cheating) represent a type of construct-irrelevance. That is, construct-irrelevant easiness and difficulty are contaminating influences on test scores that systematically increase or decrease test scores for an examinee or group (Haladyna & Downing, 2004).

For testing programs where scale stability is a fundamental concern, IPD presents a threat to the stability of the scale because of changes in item parameter estimates (Huggins-Manley, 2017). If the item parameter estimates change over time, forms that are IRT pre-assembled from an item bank are likely to be easier or harder than the actual difficulty level intended. The estimation of ability estimates will also be affected, as groups that perform better on the common items due to IPD are likely to have their ability overestimated, while groups that do not benefit from IPD may have their ability underestimated. Thus, IPD may compromise the comparability of scores between forms, undermine validity, and result in negative consequences for some examinees.

Purpose

Equating is a commonly used statistical process to ensure the comparability of scores by maintaining scale stability over time. Equating is mainly used to correct for minor adjustments in form difficulty, but the presence of IPD may lead to greater differences in form difficulty and produce worse equating outcomes than practitioners are aware of. The inaccuracy of the equating outcomes may be compounded further by the type of linking method used to adjust for group ability differences. Thus, IPD has the potential to effect both the item parameter estimates and the linking constants used to place forms on the same scale.

A considerable amount of research has been conducted on the performance of different unidimensional IRT linking methods (e.g., Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Kang & Petersen, 2011; Kim & Kolen, 2007; Lee & Ban, 2010; Uysal & Kilmen, 2016), but few studies have examined the performance of unidimensional IRT linking methods in the presence of drift. Further research is needed to evaluate the robustness of IRT linking methods with IPD.

Studies have typically found that equating outcomes improve when common items that drift are removed from the linking and equating process (e.g., Hu et al., 2008; Li, 2012; Vukmirovic et al., 2003). However, the removal of common items often leads to construct underrepresentation, which may produce less accurate equating outcomes (e.g., Keller & Keller, 2015; Klein & Jarjoura, 1985; Yang, 2000). IPD detection methods can also report false negatives (e.g., DeMars, 2004b; Donoghue & Isham, 1998), so not all items that exhibit drift may be detected. Furthermore, removing drifted items can be

an iterative, time consuming approach (Gaertner & Briggs, 2009) that could require subject matter experts to determine whether a common item can be removed prior to linking and equating. Hence, it is important to evaluate the impact of IPD when items are not detected or cannot be removed from the common item set.

Drift presents an insidious threat to practitioners who regularly assemble test forms and conduct equating with unidimensional IRT. IPD results in inaccurate equating outcomes that undermine the use and interpretation of test scores and weaken validity evidence. Moreover, IPD may unfairly result in negative consequences for examinees seeking access to greater opportunities in higher education or career advancement. The purpose of this study is to examine the extent to which IPD affects equating outcomes and determine which IRT linking methods are most robust to different conditions of drift.

Current Study and Research Questions

The current study compares the performance of five unidimensional IRT linking methods within the context of IPD: (1) Stocking-Lord, (2) Haebara, (3) concurrent calibration, (4) fixed parameter calibration, and (5) least absolute values. Because drifted items may go undetected, the study aims to determine the impact of drift when common items are not removed prior to linking and equating. Results from the study will contribute to a limited body of research and provide guidelines to help psychometricians confronted with IPD when equating. Implications for validity and recommendations for validation procedures will be identified to provide practitioners best practices for supporting their validity arguments.

In order to explore as many settings and conditions as possible, the present study will explore drift and linking methods through simulated and empirical datasets. Factors that are expected to impact findings include the proportion of drifted items in the common item set, the magnitude of the drifted items, differences in group ability, and sample size. As drift has consequences on both linking and equating outcomes, an inspection of the linking constants, item parameter estimates, equating outcomes, and classification rates will be evaluated. Thus, the research questions are as follows:

1. What is the impact of IPD on linking constants *A* and *B*?
2. What is the impact of IPD on the recovery of linked item parameter estimates?
3. How consequential is the effect of IPD on true and observed equated scores?
4. To what extent does IPD affect classification accuracy rates?
5. To what extent does IPD affect classification consistency rates?

CHAPTER II

LITERATURE REVIEW

This chapter is broken up into six sections. The first section briefly describes IRT true and observed score equating. The second section discusses the common-item non-equivalent groups (CINEG) data collection design. The third section examines seven unidimensional IRT linking methods. The fourth section reviews research conducted on the performance of IRT linking methods. The fifth defines IPD, the reasons for drift, and the implications it has on validity and validation. The last section reviews research conducted on the performance of IRT linking methods in the presence of IPD.

IRT Equating

Although the focus of this dissertation is linking, equating is discussed here because both IRT true and observed score equating results will be used as evaluation criteria to examine the performance of the linking methods being investigated. The equating procedures presented below are not being compared or contrasted but presented as brief introductions as to how equated scores are obtained.

True Score Equating. Equating is a statistical process that adjusts for variations in difficulty among forms (Kolen & Brennan, 2014). To adjust for difficulty, IRT true score equating can be used to find the number-correct true score on Form X that corresponds to the number-correct true score on Form Y. The number-correct true score is computed by summing all of the item characteristic curves at a given θ. Test

characteristic curves are used to find the true score associated with θ on Form X that corresponds to the true score associated with θ on Form Y.

As θ approaches -∞ the probability of correctly answering item *j* approaches c_j instead of 0. As θ approaches ∞ the probability of correctly answering item *j* approaches 1, but never reaches 1. Thus, true scores can only be obtained between the sum of c_j and one point below the total score. However, true scores represent parameters that are unknown, so observed scores are used in practice. Unlike true scores, observed scores can fall between 0 and the highest possible score. To find scores that exist between 0 and the sum of c_j, linear interpolation is used (Kolen, 1981). First, a score of 0 on Form X is set equal to 0 on Form Y. Second, the sum of c_j on Form X is set equal to the sum of c_j on Form Y. Then, linear interpolation is used to find equivalent scores between these two points.

Observed Score Equating. IRT observed score equating consists of three steps. First, a conditional observed score distribution is estimated using the Lord and Wingersky (1984) recursion formula:

$$\begin{aligned} f_r(x|\theta_i) &= f_{r-1}(x|\theta_i)(1-p_{ir}), && x = 0 \\ &= f_{r-1}(x|\theta_i)(1-p_{ir}) + f_{r-1}(x-1|\theta_i)p_{ir}, && 0 < x < r \\ &= f_{r-1}(x-1|\theta_i)p_{ir}, && x = r \end{aligned} \qquad (2.1)$$

where $f_r(x|\theta_i)$ is the distribution of number-correct scores over *r* items for examinees of ability θ_i. The probability of earning a 0 on the first item is defined as $f_1(x = 0|\theta_i) =$

$(1 - p_{i1})$ whereas the probability of earning a 1 on the first item is defined as $f_1(x = 1|\theta_i) = p_{i1}$.

The second step is to obtain the marginal observed score distribution by integrating the conditional distribution over all points of θ:

$$f(x) = \int_{\theta} \psi(\theta) f(x|\theta)\, d\theta, \tag{2.2}$$

where $\psi(\theta)$ represents the synthetic population from the distributions of X and Y; and d is a scaling constant set to 1 or 1.7. In order to perform equating, a single (synthetic) population must be obtained by combining the two populations, X and Y, under the CINEG design.

The last step is to apply the traditional equipercentile method:

$$e_Y(x) = F_Y^{-1}\big(F_X(x)\big), \tag{2.3}$$

where $e_Y(x)$ is the Form Y equivalent of score x on Form X; F_X and F_Y are the cumulative distribution functions for each scale; and $F_Y^{-1}$is the inverse function of F_Y.

Common-Item Nonequivalent Groups (CINEG) Design.

Also referred to as the non-equivalent groups anchor test (NEAT) design, the CINEG design is most widely used in practice. Implemented when only one form per test date can be administered because of security concerns, test forms share a set of common items (anchors) that are used to differentiate group ability from differences in form difficulty (Kolen & Brennan, 2014). Common items that are internal contribute to the total test score, whereas external common items do not count towards the total test score and are mainly used for equating purposes.

While the CINEG design is the most practical design to use for most testing programs, certain conditions need to be met in order to ensure accurate linking. Common items should be proportionally representative of the entire test form from a content and statistical perspective (Kolen & Brennan, 2014), otherwise the common item set will suffer from construct underrepresentation (Messick, 1989). Without the proper proportion of content, linking results may be inaccurate (e.g., Keller & Keller, 2015; Sukin & Keller, 2008) and differences in group ability may not be adequately captured. Additionally, the same set of common items should not be reused for every new test form created. The more frequently a common item is used, the greater the likelihood of that item being exposed to the population of examinees, which may lead to IPD.

Unidimensional IRT Linking Methods

Under the CINEG design, groups are not considered equivalent and the item parameter estimates for each form need to be placed on the same scale. Although groups differ in their θ distributions, software programs constrain each θ distribution to a mean of 0 and standard deviation of 1. A linear transformation can be made to the item parameter estimates so that the IRT model produces the same fitted probabilities of correct responses (Hanson & Beguin, 2002). The two θ-scales are linearly related as follows:

$$\theta_{Y_i} = A\theta_{X_i} + B, \tag{2.4}$$

where A and B are the slope and intercept of the linear equation, respectively, and θ_{X_i} and θ_{Y_i} are the ability values of examinee I on the scale of Form X and Form Y,

respectively. Under the three parameter-logistic (3PL) model, the item parameters are related as follows:

$$a_{Y_j} = \frac{a_{X_j}}{A}, \tag{2.5}$$

$$b_{Y_j} = Ab_{X_j} + B, \tag{2.6}$$

and

$$c_{Y_j} = c_{X_j}, \tag{2.7}$$

such that a_{X_j}, b_{X_j}, and c_{X_j} are the item discrimination, item difficulty, and pseudo-guessing parameters, respectively, for item *j* on Form X; and a_{Y_j}, b_{Y_j}, and c_{Y_j} are the same parameters for item *j* on Form Y. As can be seen below, plugging in the scale transformation equations directly into the 3PL model will produce the same probability of a correct response:

$$P\left(\theta_{Y_i}, a_{Y_j}, b_{Y_j}, c_{Y_j}\right) = c_{Y_j} + \left(1 - c_{Y_j}\right)\frac{\exp\left[Da_{Y_j}\left(\theta_{Y_i} - b_{Y_j}\right)\right]}{1 + \exp\left[Da_{Y_j}\left(\theta_{Y_i} - b_{Y_j}\right)\right]} \tag{2.8}$$

$$= c_{X_j} + (1 - c_{X_j})\frac{\exp\left\{D\frac{a_{X_j}}{A}\left[\left(A\theta_{X_i} + B\right) - \left(Ab_{X_j} + B\right)\right]\right\}}{1 + \exp\left\{D\frac{a_{X_j}}{A}\left[\left(A\theta_{X_i} + B\right) - \left(Ab_{X_j} + B\right)\right]\right\}}$$

$$= c_{X_j} + (1 - c_{X_j})\frac{\exp\left[Da_{X_j}\left(\theta_{X_i} - b_{X_j}\right)\right]}{1 + \exp\left[Da_{X_j}\left(\theta_{X_i} - b_{X_j}\right)\right]}$$

$$= P\left(\theta_{X_i}, a_{X_j}, b_{X_j}, c_{X_j}\right),$$

where $P\left(\theta_{Y_i}, a_{Y_j}, b_{Y_j}, c_{Y_j}\right)$ and $P\left(\theta_{X_i}, a_{X_j}, b_{X_j}, c_{X_j}\right)$ are the probabilities that examinee *i* correctly answers item *j* on scales Y and X. Several different linking methods can be used to obtain linking constants *A* and *B*.

Mean-Sigma. Proposed by Marco (1977), the mean-sigma (MS) method uses the mean and standard deviation of item difficulty estimates from each test form to derive the linking constants:

$$A = \frac{\sigma(a_Y)}{\sigma(a_X)}, \tag{2.9}$$

and

$$B = \mu(b_Y) - A\mu(b_X), \tag{2.10}$$

where $\sigma(a_Y)$ and $\sigma(a_X)$ are standard deviations of *a*-parameter estimates of the common items for Forms Y and X, respectively; and $\mu(b_Y)$ and $\mu(b_X)$ are the means of *b*-parameter estimates of the common items for Forms Y and X.

Mean-Mean. Similar to the mean-sigma method, the mean-mean (MM) method (Loyd & Hoover, 1980) uses the means of the item difficulty and item discrimination estimates from each test form to compute the linking constants. The *B* constant can be calculated using the same equation from the mean-sigma method; however, the *A* constant is computed as follows:

$$A = \frac{\mu(a_X)}{\mu(a_Y)}, \tag{2.11}$$

such that the mean of the discrimination estimates for Form X are divided by the mean of the discrimination estimates for Form Y.

Haebara. One limitation of the MM and MS methods is that they do not consider all of the item parameter estimates simultaneously in the transformation (Kolen & Brennan, 2014). To resolve this issue, Haebara (1980) and Stocking and Lord (1983) developed linking methods using characteristic curves.

The Haebara method takes the difference between each item characteristic curve (ICC) on the base scale and transformed scale, squares the difference, and then sums all the differences over the common items (*j:V*), as such:

$$Hdiff(\theta_i) = \sum_{j:V}\left[p_{ij}(\theta_{Yi}; \hat{a}_{Yj}, \hat{b}_{Yj}, \hat{c}_{Yj}) - p_{ij}\left(\theta_{Yi}; \frac{\hat{a}_{Xj}}{A}, A\hat{b}_{Xj} + B, \hat{c}_{Xj}\right)\right]^2, \quad (2.12)$$

where $p_{ij}(\theta_{Yi}; \hat{a}_{Yj}, \hat{b}_{Yj}, \hat{c}_{Yj})$ represents the item characteristic function on the scale of Form Y, and $p_{ij}(\theta_{Yi}; \frac{\hat{a}_{Xj}}{A}, A\hat{b}_{Xj} + B, \hat{c}_{Xj})$ represents the item characteristic function on the scale of Form X transformed to the scale of Form Y. *Hdiff* is then summed over all examinees, retrieving values of *A* and *B* that minimize the following criterion:

$$Hcrit = \sum_i Hdiff(\theta_i) \quad (2.13)$$

Stocking-Lord. The Stocking and Lord (1983) method uses a similar equation to Haebara, except that they sum the differences of ICCs before squaring:

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}\left(\theta_{Yi}; \hat{a}_{Yj}, \hat{b}_{Yj}, \hat{c}_{Yj}\right) - \sum_{j:V} p_{ij}\left(\theta_{Yi}; \frac{\hat{a}_{Xj}}{A}, A\hat{b}_{Xj} + B, \hat{c}_{Xj}\right)\right]^2. \quad (2.14)$$

The difference here is that the sums of all the differences of common items are taken prior to squaring. That is, the Stocking-Lord (SL) method examines the squared difference between the test characteristic curves for a given θ_i whereas Haebara examines

the squared difference between the ICCs for a given θ_i. *Sldiff* is then summed over all examinees, retrieving values of *A* and *B* that minimize the following criterion:

$$SLcrit = \sum_i SLdiff(\theta_i). \tag{2.15}$$

Least Absolute Values. He et al. (2015) proposed a robust scale transformation method called Least Absolute Values (LAV). The LAV combines ordinary least squares regression and the Haebara method to obtain linking constants. Ordinary least squares regression can be influenced due to outliers, so a weight function is used to reduce the impact of the outliers:

$$\sum_i w_i * r_i^2, \tag{2.16}$$

where w_i is a weight for the *i*th observation, and r_i^2 is the squared residual (i.e., difference between observed and predicted values) of the i^{th} observation. Using equations 2.4 – 2.8, the difference in probability of getting a correct answer based on the base scale and transformed scale is:

$$d_{ij} = p_{ij}\left(\theta_{Yi}; \hat{a}_{Yj}, \hat{b}_{Yj}, \hat{c}_{Yj}\right) - p_{ij}\left(\theta_{Xi}; \hat{a}_{Xj}, \hat{b}_{Xj}, \hat{c}_{Xj}\right) \tag{2.17}$$

$$= p_{ij}\left(\theta_{Yi}; \hat{a}_{Yj}, \hat{b}_{Yj}, \hat{c}_{Yj}\right) - p_{ij}(\theta_{Yi}; \frac{\hat{a}_{Xj}}{A}, A\hat{b}_{Xj} + B, \hat{c}_{Xj})$$

A loss function *L* evaluates the resultant losses, d_{ij}, as such:

$$L\left(d_{ij}\right) = \sum_i \sum_j w_{ij} d_{ij}^2, \tag{2.18}$$

where w_{ij} is the weight assigned to the probability difference for item *j* and examinee *i*. The weight, w_{ij}, can also be defined as $w_{ij} = 1/|\, d_{ij}\,|$, which simplifies to:

$$L_{LAV}(d_{ij}) = \sum_i \sum_j |d_{ij}|. \quad (2.19)$$

Thus, the LAV minimizes the absolute difference between two ICCs. Large values of d_{ij} correspond to smaller weights for the squared difference.

Concurrent Calibration. Concurrent calibration (CC) estimates person and item parameters from two or more forms simultaneously in one computer run. Although separate calibration procedures require linking methods (e.g., Haebara, SL, and LAV) to place the estimates from two forms on the same scale, CC does not require any additional scale transformation procedure. Instead of fixing the θ distribution to a standard normal distribution, CC estimates the distributions simultaneously with the item parameters. Thus, the estimated distributions and item parameter estimates obtained from CC are already on the same scale.

Fixed Parameter Calibration. Fixed parameter calibration (FPC) takes the item parameter estimates from a set of previously calibrated common items (or item bank) and uses these values for the same set of common items on a new form when calibrating the new form field-test items. No scale transformation is required for FPC because the distribution of θ for the new group is estimated using their responses to the common items and the item parameter estimates from the base form. The resulting distribution of θ for the new group will be on the scale of the base form, as well as the unique items from the new form.

Comparison of Unidimensional IRT Linking Methods

Seven methods were presented in the previous section, and although each method has benefits, the moment methods (i.e., MM and MS) have produced less stable results than the Haebara and SL characteristic curve methods (e.g., Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Kim & Lee, 2004; Li et al., 2012; Ogasawara, 2001; Uysal & Kilmen, 2016). For this reason, the moment methods will not be considered further. Unless otherwise specified, the studies listed below all implemented the CINEG under unidimensional IRT.

Investigating the scale stability of the math and verbal sections from the SAT, Petersen et al. (1983) examined the true score equating results using linear (i.e., Tucker and Levine), equipercentile, and IRT equating methods. Three IRT linking methods (i.e., CC, FPC, SL) were used with the 3PL model. Using the LOGIST computer program, linking methods were evaluated according to the weighted mean squared difference between observed and estimated scale scores. For reasonably parallel tests, the linear methods performed similarly to the IRT methods. However, when tests were not reasonably parallel, the IRT methods were more robust than the linear methods. Among the IRT methods for the verbal section, FPC performed the best, followed closely by CC. For the math section, CC was superior to both FPC and SL. Overall, CC was considered to be the most stable.

Using data calibrated from two math forms of the ACT to obtain generating item parameters for a simulation study, Hanson and Béguin (2002) analyzed the performance of MM, MS, SL, Haebara, and CC linking methods under the 3PL model. The following

factors were included: the estimation program (MULTILOG versus BILOG-MG), sample size (3,000 versus 1,000), number of common items (20 versus 10), and equivalent groups sampled from N(0, 1) versus nonequivalent groups with a base group sampled from N(0, 1) and a new group sampled from N(1, 1). IRT true score equating criterion and ICC criterion were used as evaluation criteria. CC performed better than all methods for both evaluation criteria, with the exception of MULTILOG N(1, 1), for which the SL method performed better than the other linking methods under the IRT true score equating criterion.

Similar to Hanson and Béguin (2002), Kang and Petersen (2011) ran a simulation study based on item parameters obtained from two math forms to compare CC, FPC, and SL linking methods. The study varied the sample size (500 versus 2,000), number of common items (10, 20, or 40), and ability distributions with a base group sampled from N(0, 1) and new groups sampled from N(0, 1), N(0.25, 1.1), and N(0.5, 1.2). Additionally, SL and CC were carried out using BILOG-MG, while FPC was calibrated with BILOG-MG and PARSCALE. Using the 3PL model, results were examined using the ICC and TCC evaluation criteria. Most notably, FPC performed significantly worse with BILOG-MG, especially with fewer common items and nonequivalent ability distributions. Otherwise, the SL, CC and FPC with PARSCALE performed comparably.

In assessing academic growth over time with grade-level math data, Jodoin et al. (2003) compared the performance of the MS, FPC, and CC methods. An external anchor CINEG matrix design comprised of 12 field-test blocks was implemented, with PARSCALE used for 2PL, 3PL, and graded-response model (GRM) calibration.

Although truth could not be ascertained from the use of real data, the FCP and CC methods performed similar to each other in terms of the MLE and EAP ability estimates and classification consistency.

Kim and Kolen (2007) conducted a simulation study to analyze factors that could potentially affect the linking process under the 3PL model. Three ability distributions were considered for both the old and new groups (i.e., normal, positively skewed, and negatively skewed) resulting in a total of nine distribution combinations. Haebara, SL, and CC methods were evaluated according to the ICC criterion. All three methods used BILOG-MG for calibration, while POLYST was used for Haebara and SL linking. CC outperformed the Haebara and SL methods in linking accuracy.

Lee and Ban (2010) compared CC, SL, Haebara, and proficiency transformation linking methods. The linkage plan from this study assumed that Form A was administered at two time points. Form B_2 was spiraled with Form A_2, so A_2 and B_2 were considered randomly equivalent, without possessing any common items. Parameter estimates from B_2 were placed onto the scale of A_1 using A_2 as an anchor form. Using a simulation study, two ACT English forms were calibrated using a 3PL model to obtain the generating item parameters. Manipulated factors included the sample size (500 or 3,000), total items (75 or 25), and ability distributions where the base group was sampled from N(0, 1) and new groups were sampled from N(0, 1), N(0.5, 1), and N(1, 1). Expected observed score distribution (ESD) and TCC criteria were used as evaluation criteria. BILOG-MG was used for calibration, and ST was used to carry out linking for

SL and Haebara. Contrary to findings from previous studies (e.g., Petersen et al., 1983; Hanson & Beguin, 2002), Haebara and SL generally performed better than CC.

Studying the accuracy and consistency of IRT true score equating results for a sequence of test forms, Li et al. (2012) used simulated data to compare the performance of the chained equipercentile equating method and IRT true score equating method based on MM, MS, SL, Haebara, and CC linking methods. PARSCALE was used for calibration with the 2PL model and results were evaluated based upon mean squared errors (MSE), bias, and variance. Overall, the SL, Haebara, and CC methods performed better than the moment methods, and were comparable to each other.

A simulation study was conducted by Kim and Cohen (1998) to compare the performance of SL to CC. Using the 2PL model, 500 examinee responses to 50 items were simulated. The number of common items varied (5, 10, 25, and 50) as did the ability distributions, with the base group sampled from a θ distribution of N(0, 1) and the new group sampled from a θ distribution of N(0, 1) and N(1, 1). Evaluation criteria included the root mean squared difference (RMSD) and mean Euclidean distance (MED). SL had smaller RMSD and MED values than CC for most conditions. However, the type of software used may have confounded the results (Hanson & Béguin, 2002), as BILOG was used for SL and MULTILOG was used for CC.

Following their previous study, Kim and Cohen (2002) compared the SL and CC methods using the GRM for a polytomously scored 30-item test. Three different sample size combinations were considered for the two forms: 300 base group examinees/300 target group examinees, 1,000/1,000, and 1,000/300. The ability of the base group was

sampled from a θ distribution of N(1, 1) to match the difficulty of the test, while the target groups were sampled from N(0, 1) and N(1, 1). Common item sets consisted of 5, 10, and 30 items. MULTILOG was used for calibration for both SL and CC and EQUATE was used for the SL scale transformation. Mean distance measure (MDM) and RMSD were the criteria used for evaluation. Results indicated that CC was slightly, but consistently, better than SL for recovery of item and ability parameters.

Keller and Keller (2011) investigated the long-term sustainability of five IRT linking methods (i.e., MM, MS, SL, Haebara, and FPC) over six administrations of a test using the 3PL model. Three different ability distribution shifts were manipulated: none, a mean shift with increments of 0.15 units starting from N(0, 1) and ending at N(0.75, 1), and a skew-shift where the mean increases as in the mean shift condition, and the skewness increased by -0.15 between each administration. PARSCALE was used for calibration and STUIRT was used for all separate calibration methods (all except FPC). Evaluation criteria included root mean square error (RMSE), bias of θ estimates, and classification accuracy. Results indicated SL and Haebara performed similarly with FPC and better than the moment methods. SL and Haebara performed best when there was a mean shift in the data, while FPC was better at handling a skew shift in the data. It was concluded that FPC was the best method to deal with complex changes in examinee performance.

A simulation study was conducted by Li et al. (1997) to compare the performance of FPC and SL methods under the 3PL model. Three different ability distributions were varied according to a standard normal distribution, a positively-skewed chi-squared

distribution with a skewness of 1, and a negatively-skewed chi-squared distribution with a skewness of -1. BILOG was used for calibration and EQUBANK was used for SL linking. Results indicated that FPC produced slightly more stable parameter estimates despite having slightly higher levels of bias.

There are several takeaways from the studies presented. First, CC has typically produced the most stable item parameter estimates and accurate equating results among all methods presented (e.g., Hanson & Beguin, 2002; Kim & Cohen, 2002; Kim & Kolen, 2007; Petersen et al., 1983). If CC did not perform the best, it performed comparably to SL, Haebara, and FPC methods (e.g., Jodoin et al., 2003; Kang & Petersen, 2011; Li et al., 2012). Only in two instances did CC perform worse than SL (i.e., Kim & Cohen, 1998; Lee & Ban, 2010). However, Lee and Ban (2010) linked two forms together with nonequivalent groups (A_1 and B_2) through group A_2, which was considered equivalent to B_2. Because there were no common items between A_2 and B_1, this type of linkage plan differed from other CINEG designs, which may have led CC to perform worse than SL. The results from Kim and Cohen (1998) may have been confounded due to differences in software. Second, FPC performed comparably to CC in most of the studies in which the two methods were used (e.g., Jodoin et al., 2003; Kang & Petersen, 2011; Keller & Keller, 2011). However, FPC has not been studied nearly as extensively as CC. Third, the SL method is the most widely used separate calibration linking method, although the performance between SL and Haebara has been comparable (e.g., Hanson & Beguin, 2002; Keller & Keller, 2011; Kim & Kolen, 2007; Lee & Ban, 2010; Li et al., 2012). More research is needed on the Haebara method.

While the results presented here seem to favor CC, these same linking methods might operate differently within the context of IPD. The final sections will discuss IPD as a threat to measurement as well as validity, and review studies that have examined the performance of linking methods in the presence of IPD.

Item Parameter Drift

One of the greatest attributes of IRT is the property of parameter invariance; item parameters remain the same over different groups of examinees and separate testing occasions. When this assumption is violated, and item parameter estimates deviate over subsequent testing administrations (Goldstein, 1983), item parameter drift (IPD) occurs. IPD is considered a type of DIF (e.g., Babcock & Albano, 2012; Gaertner & Briggs, 2009), but instead of items functioning differently between subgroups (e.g., male versus female), items differ over testing administrations.

IPD can have detrimental effects on linking both directly and indirectly (Han et al., 2012). First, item parameter estimates will be directly impacted. Consequently, procedures that rely on these estimates will also be subjected to IPD. For example, when pre-assembling forms with IRT, statistical specifications should be nearly identical so as to not advantage or disadvantage examinees taking a specific form. Although pre-assembling forms requires that the scored items already be calibrated and fixed to a bank scale, any items that exhibit drift will deviate from the fixed estimate and change the difficulty (easier or harder) of the form without the test developer being aware. Second, the linking constants will be indirectly affected by IPD through the drifted item parameter

estimates (Han et al., 2012). As a result, the linked item parameter estimates will be negatively influenced.

When using the 2PL or 3PL model, three types of IPD can be investigated: *a*-drift, *b*-drift, and *ab*-drift (e.g., DeMars, 2004b; Donoghue & Isham, 1998; Wells et al., 2002). Changes to the discrimination parameter over time are referred to as *a*-drift, changes to the difficulty parameter over time are known as *b*-drift, and changes to both the discrimination and difficulty values over time are referred to as *ab*-drift. Donoghue and Isham (1998) found that detection rates for *a*-drift were significantly lower than detection rates for *b*-drift and *ab*-drift. Of the 13 detection methods investigated, only one method (Lord's χ^2) had an *a*-drift detection rate above 50%, which led the authors to conclude that all methods were insensitive to *a*-drift. Drift in difficulty parameters (*b*-drift) tends to be the most common IPD as the detection of *a*-drift can be challenging.

Reasons for IPD. Any number of causes could result in IPD and the identification of a particular reason could be very difficult in practice. Yet, researchers have proposed and investigated different sources of IPD. First, changes in curriculum could result in items becoming easier or harder (e.g., Bock et al., 1988; DeMars, 2004a; Goldstein, 1983; Sykes & Fitzpatrick, 1992). Using data from the College Board Physics Achievement Test, Bock et al. (1988) found that basic mechanics items became easier over a 10-year span, whereas other specialized topics became harder over time. These findings were supplemented by a curriculum survey indicating that basic topics were more regularly stressed. DeMars (2004a) investigated IPD by comparing items from information literacy and global issues over the course of four years. DeMars (2004a)

found that items from information literacy showed more drift due to the swift rate at which content was likely to change in information literacy, but drift could not always be explained by content alone. Sykes and Fitzpatrick (1992) examined the effect of item position, item type, item content and elapsed time between test administrations on possible changes in item difficulty on a professional licensure exam. Results revealed no significant relationship between item position or item type, but a significant difference for elapsed time and content categories. Sykes and Fitzpatrick (1992) hypothesized that the drift due to the content was attributed to the change in curriculum. In examining reasons for educational attainment over time, Goldstein (1983) suggested that mental arithmetic could be phased out of curriculum due to advances in technology. If an item is presented on an exam that requires mental arithmetic without use of a calculator, examinees may not be as well versed in solving the problem as examinees who were taught mental arithmetic before the regular use of calculators. Thus, the item would become harder. On the other hand, the same item could become easier for subsequent test takers if mental arithmetic is required but the use of a calculator is also permitted.

Depending on the location of the item in the test form, a context effect may occur. Kingston and Dorans (1984) investigated the effect of item location on item-types by spiraling 12 sub-forms of the GRE General Test. Analytical items, which require an extensive set of directions, were susceptible to significant practice effects such that the performance of these items depends on how many items of that type precede it. Therefore, these items could be more difficult if fewer items of the same type are presented, whereas these items could become easier if more items are presented.

One reason that could result in items drifting easier or harder is if the initial calibration is unstable or response behavior is not properly modelled (Glas, 2000). This may be due to small sample sizes, changes in population characteristics between administrations (e.g., first-time testers versus retesters), or due to seasonality effects (Wyse & Babcock, 2016), any of which could cause items to drift in different directions.

Motivation level is another potential reason why an item could drift easier. Glas (2000) suggests that differences in performance can occur between pretest and on-line stages. If examinees are aware that certain items (i.e., field-test, experimental, external) do not count towards their score, they have little incentive to give maximum effort and their motivation may wane. Thus, the item estimate might appear more difficult after pretest but easier once on-line. Alternatively, items at the end of a long test might seem harder due to examinee fatigue or due to lack of time causing examinees to guess.

Common items regularly used on different forms are at risk for overexposure as test-takers may begin to recognize items when taking the exam multiple times (Jurich et al., 2012). Items could also be exposed when test-takers discuss information about the test to one another or post information on "braindump" websites (Smith, 2004). Although items may be thought of as being exposed only after a test has been administered, test security is needed throughout all stages of test development. Security breaches can occur when test materials are hacked online, if booklets are not secured during meetings (e.g., item review, standard setting), or when materials are not properly disposed.

Messick (1989) referred to test savviness as a form of construct-irrelevant easiness whereby students can identify clues in items that lead them to choosing the

correct answer. These items may be easier to test-takers that are more adept at test taking. Additionally, test preparation courses offer examinees the opportunity to take advantage of test-taking strategies to make more efficient use of their time and methods to handle certain types of problems.

Finally, current news and the corresponding media attention given to certain topics may cause certain items to drift. O'Neill et al. (2013) remarked that answering a question about HIV in 1986 represents an esoteric immunology topic, but in 1992 it represents a current events topic due to the outbreak of cases between this time period. The attention given to the topic, and the information available, will be more substantial in 1992 compared to 1986. A list of potential reasons for IPD can be found in Table 1.

Table 1

Potential Reasons and Directionality of IPD

Reason	**Easier**	**Harder**	**Citation**
Changes in curriculum	Yes	Yes	Bock et al. (1988); DeMars (2004a); Goldstein (1983); Sykes & Fitzpatrick (1992)
Technological advances	Yes	Yes	Goldstein (1983)
Item location	Yes	Yes	Kingston & Dorans (1984)
Unstable or poor initial calibration/ improper modeling	Yes	Yes	Glas (2000); Wyse & Babcock (2016)
Motivation	Yes		Glas (2000)
Item overexposure	Yes		Smith (2004)
Cheating	Yes		Jurich et al. (2012)
Security breach	Yes		Jurich et al. (2012)
Test-taking strategies/test savviness	Yes		Messick (1989)
Current news	Yes		O'Neill et al. (2013)

IPD Implications for Validity and Validation. Regardless of the reason or direction for IPD, the presence of drift is a threat to measurement contexts that require a stable scale, such as licensure and certification. All the aforementioned reasons are sources of construct-irrelevant variance that may jeopardize the assumption of parameter invariance, thereby threatening the generalizability of test scores across examinee populations and measurement conditions (Rupp & Zumbo, 2006). IPD also has major implications for the fairness and validity of test scores, as well as the process of validation.

IPD as a Threat to Validity. Although too prescient for the time, the first notions of a theoretical definition of validity emerged from Cronbach and Meehl (1955). They conceptualized the validity triumvirate recognized today—criterion-related validity, content validity, and construct validity. More importantly, they postulated a nomological network to help confirm or disconfirm the interpretation of test scores through a system of laws and relationships that define a theory. As Box (1976) stated, "all models are wrong" (p. 792), including measurement models such as classical test theory (CTT) and IRT. Although we accept these theories (i.e., CTT and IRT) as approximations of someone's true ability, we recognize that some tolerable amount of error is associated with using them. How much error is considered consequential though? Moreover, when IPD is present, how much more error is added when our models and theories break down as a result of the assumptions that define them (e.g., parameter invariance)?

Dorans and Feigenbaum (1994) suggested a raw score difference of 0.5 or greater as a "difference that matters" when rounding is also considered. This would translate to a

difference of one raw score point, which could lead to different interpretations. For example, a difference of one fewer point could result in a student being labeled as "Below Proficient" instead of "Proficient." Another instance is a prospective lawyer or doctor who "Fails" their certification exam by one point. While these labels are the results of interpretations from test scores, Messick (1989) also emphasized the importance of subsequent *actions* based upon interpretations from test scores. In the context of the student, he/she may have to attend a remedial class instead of continuing along the same trajectory of his/her classmates. For the prospective candidate, failing the exam means having to restudy, investing more financial resources in exam preparation materials, or possibly considering a career change.

These hypothetical scenarios may become realities when considering the influence of drift. The two examples above illustrate the consequences that could ensue if items drift harder, whether due to a lack of coverage in curriculum, item location, or an initial calibration suggesting an item is easier than it really is. Alternatively, examinees may also benefit from items drifting easier, which could be a result of item overexposure, cheating, or security breaches. Studies examining IPD on equating results (e.g., Hu et al., 2008; Jurich et al., 2012; Li, 2012; Vukmirovic et al., 2003) have found drift to affect equated scores by one point or more.

If IPD goes undetected, examinees will receive a score that is different from the one they should be correctly awarded if the drift were detected (Rupp & Zumbo, 2003a). Estimated equated scores that differ by even one point are a detriment to the interpretation and use of test scores. Stated by Messick (1989), validity is a matter of

degree—so as scores drift more, the weaker the argument becomes for claiming the test scores are suitable measures for the specified purpose of the test.

Impact on Validation. As part of their conception of construct validity, Cronbach and Meehl (1955) suggested a statement of the proposed interpretation, consideration of alternative interpretations, and the need for extended analysis in validation. Messick (1989) reiterated these same points, which have become widely accepted, as necessary components for a validity argument.

Messick's work largely influenced the *Standards*, which is considered as one of, if not the, premier resource for guidance on testing. Although the *Standards* discusses DIF (p. 16, 51, 82) as a threat to the internal structure of a test, there is no specific reference to IPD (although a number of standards allude to potential reasons for drift). This may be partially attributed to the fact that IPD is considered a special type of DIF (e.g., Babcock & Albano, 2012; Gaertner & Briggs, 2009). Instead, the *Standards* allude to IPD through equating, stating: "It is important to check that the anchor items function similarly in the forms being equated. Anchor items are often dropped from the anchor if their relative difficulty is substantially different in the forms being equated" (*Standards*, p. 98). However, IPD has the potential to not only impact the internal structure of a test, but all five sources of validity evidence: 1) evidence based on test content; 2) evidence based on response processes; 3) evidence based on internal structure; 4) evidence based on relations to other variables; and 5) consequences of testing.

Drift has the potential to affect test content, which speaks to the relationship between the test content and the construct being measured. Content-related validity

evidence should “address issues such as the fidelity of test content to performance in the domain in question and the degree to which test content representatively samples a domain, such as a course curriculum or job.” (*Standards*, p. 218). The *Standards* discuss a number of considerations for test design and development that contribute to providing support for content-related validity evidence. Commentary of Standard 4.8 states that: “When sample size permits, empirical analyses are needed to check the psychometric properties of test items and also to check whether test items function similarly for different groups” (p. 88). Psychometric properties can be reviewed by subject matter experts as in Standard 4.8, or by the test developer in Standard 4.10. Subject matter experts may be able to identify items that are obsolete (e.g., replaced by new findings or laws) or speak to subject areas that have been added or retired. Test developers are likely to evaluate items exhibiting IPD or DIF during scoring (applicable for evidence of internal structure). However, when using IRT, the item bank can be used to assemble domain and test-level difficulty to a specified value. If the bank scale was originally calibrated using items that drifted, then the forms assembled may be easier or harder than they statistically exhibit. This would result in some examinees receiving an easier form than other examinees, despite being assembled to the same statistical specifications. Test developers should ensure that the item bank is calibrated with the right population and an adequate number of examinees (Wyse & Babcock, 2016).

Response processes refer to the cognitive processes (e.g., test-taking strategies, response times, eye movements) that examinees engage in while taking the test (Standard 1.12). These processes may change as a result of drift. For example, if examinees have

knowledge of items as a result of cheating or being exposed to the item, their response time will be very quick. Context and practice effects may also occur (Kingston & Dorans, 1984). Some item formats require more extensive directions than others, which might inhibit examinees that are unfamiliar with their format. However, examinees that retake the test will not be caught off guard by the complexity of the item and can spend more time on other areas. In these examples, the cognitive processes being used by examinees do not tap into the intended processes required to demonstrate competence or mastery; rather, they reflect having access or exposure to items that other test takers do not get to benefit from.

The internal structure of the test refers to whether the obtained scores function as intended. In the context of IRT, it is expected that items will perform similarly across different groups of examinees and conditions. As previously mentioned, the *Standards* recommend removing common items from the anchor set if their difficulty fluctuates over different test forms (Standard 5.15).

When scores from one test correlate with scores from another test measuring the same construct (e.g., SAT and ACT), there is some convergent evidence for relations to other variables (Standard 4.13). The *Standards* also apply relations to other variables in the context of subgroups. That is, test-criterion relationships may differ from one subgroup to another, or the difference may be attributed to different meanings for the groups. These differences result from construct-irrelevant variance such as DIF and IPD. A test form that has been compromised as a result of cheating or a security breach will

yield scores much higher than expected. These scores may be more homogeneous than expected and result in a lower correlation with another form or test.

Finally, evidence should be provided for the consequences of testing. The consequences of test use "follow directly from the interpretation of test scores for uses intended by the test developer" (*Standards*, p. 19). Consequences may be intended or unintended, and in the case of drift, are likely unintended due to its potential to go unnoticed. If drift is undetected, examinees will receive a score with more error that misrepresents their ability. As a result, some unqualified test-takers will pass or advance, and some qualified test-takers will fail. Unqualified test-takers that benefit from drift may go on to hold career positions (e.g., doctors, lawyers) that they are ill-equipped to handle, which could put patients or clients at risk for harm. Those test-takers that are disadvantaged by drift will have to devote more time and financial resources to retaking the exam. Test developers should investigate sources of construct-irrelevant variance that could be contributing to examinees' scores (Standard 1.25).

While the *Standards* discusses the need for composing a validity argument, it does not elucidate how to construct one. Kane (2006, 2013) operationalized a method for practitioners by introducing an argument-based approach to validation that includes an interpretive use argument (IUA) and a validity argument. The IUA provides a framework for the claims being made about test scores through a network of inferences and assumptions as seen in Figure 1.

Figure 1. Kane's Argument Based Approach to Validation.

The IUA requires at least four inferences with the number of inferences depending upon the extent of the claims being made. Each inference can be thought of as a bridge interconnected with the other inferences. If any inference lacks the evidence to support the claim being made, then the bridge collapses. Progression to the subsequent bridge cannot be made until the claims being made for that inference are supported.

The first bridge is the *scoring/evaluation* inference, which assumes that the scoring rule (e.g., answer key, rubric) is appropriate, accurate, and consistent for the purposes of assigning an observed score based on an observed performance. The second is *generalization*, which assumes that the examinee's performance on this occasion would generalize to a universe of other occasions, settings, raters, etc. The third is *extrapolation*, which implies that the performance on the test is indicative of the knowledge, skills, or attributes required for a given job or context. The fourth inference is *utilization*, which suggests that scores can be used to make value-based decisions (e.g., admission decision, pass/fail for licensure). A high-stakes examination based upon research and theory is likely to have more inferences than just a classroom-based assessment. Once the claims, assumptions, and inferences have been laid out in the IUA, they are critically evaluated by a validity argument.

Using Toulmin's (1958) model of inference (Figure 2), the scoring inference takes us from a sample of observations (i.e., grounds or data) to our claim about the observed scores (Kane, 2006, 2013). The inference is supported by our warrant, which states that the appropriate answer key/rubric, testing conditions, and statistical analyses (e.g., item calibration, linking, and equating) are applied accurately, consistently, and are free from any bias. The warrant is based on a number of assumptions that must be empirically or theoretically validated through the backing. Unless there are alternative hypotheses (i.e., rebuttals) that disconfirm our evidence, we can claim that our scoring inference has been supported. Examining the rebuttals in Figure 2, there is evidence of IPD that could be attributed to item overexposure, cheating, or changes in curriculum. Although several items were removed, there are other anchor items that cannot be removed due to a lack of content balance. Therefore, our linking and equating results might be negatively influenced and our claim of observed test scores accurately reflecting examinee performance is not supported. As a result, we are unable to provide enough support for the scoring inference, which means that we cannot claim that our observed test scores are suitable to be used to make decisions about examinees. It also means that we cannot move to any of the next inferences until this is resolved (displayed by the red "X" marks on the arrows).

Although it would not make sense to continue with our IUA after failing to support the scoring inference due to drift, we will evaluate each of the inferences for this hypothetical scenario. Additional reasons for drift will be explored as ways that drift can affect the IUA. Continuing with generalization, observed scores should be

representations of expected scores over parallel versions of tasks, occasions, and raters (Kane, 2006, 2013). In this inference, IPD presents a threat to the task or item over separate testing occasions. For example, an item could be identified by an examinee over repeated administrations and may find the item easier than when it was first presented. Alternatively, the item does not have to be seen by an examinee twice. Instead, an examinee might have knowledge of this item from other test-takers due to cheating or breaches in security. In either instance, the claim that an examinee would receive the same expected score across testing occasions is unsupported. Other examples that could invalidate the generalization inference include presenting an item in different locations of forms (which could change the performance of an item based on context clues from surrounding items), or linking a newly administered form to a bank scale that contains unstable initial calibrations.

Moving to the extrapolation inference, which states that the knowledge, skills, and abilities assessed by the exam for a construct are indicative of the performance relevant to a specified setting. IPD may unduly influence the construct being measured across different examinees. An examinee that takes the test with prior knowledge of the items will receive a score that inflates their true ability, compared to an examinee that takes the test with no prior knowledge of the items. The former test-taker's score is a measure of ability plus familiarity with the exam, whereas the latter examinee's score is a measure of only their ability. Thus, the two observed scores represent different constructs with different meanings and cannot be considered fair to the examinees or in support of the extrapolation inference, compromising the validity argument.

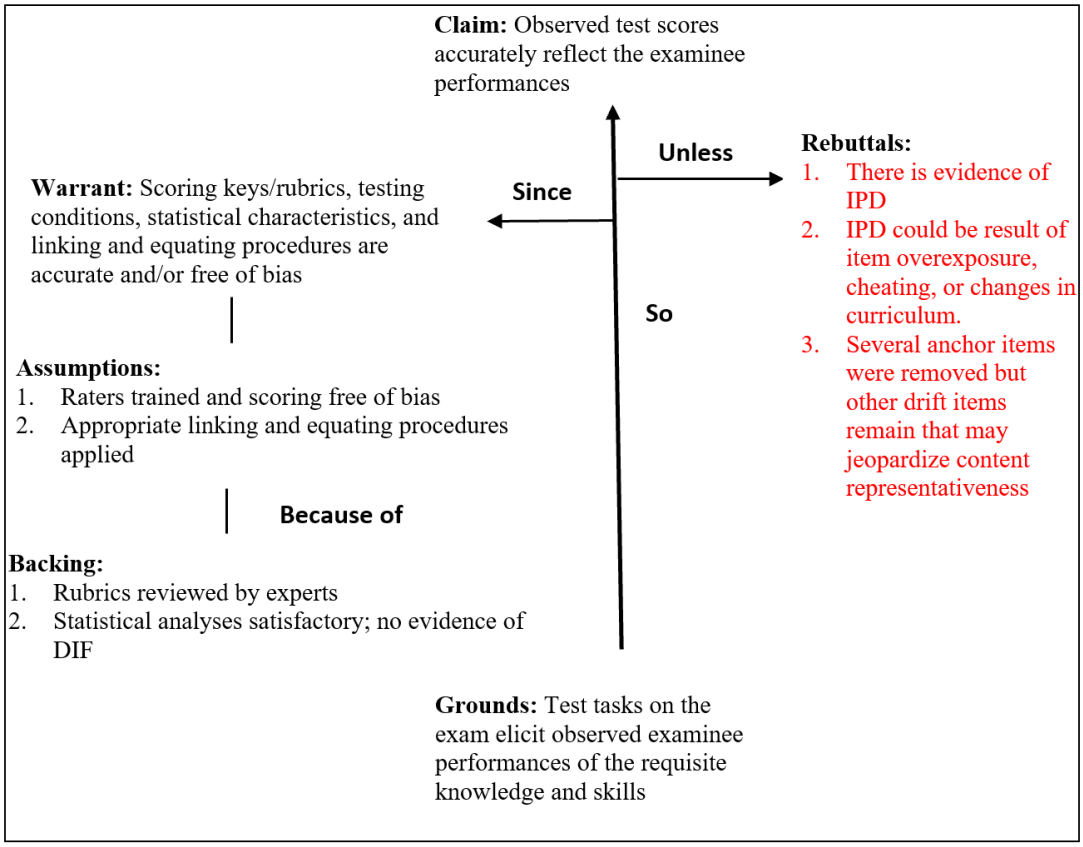

Figure 2. Toulmin's Model of Inference Applied to Kane's Scoring Inference.

Finally, the utilization inference suggests that the scores obtained are useful for making decisions about the competence for a given practice, role, or setting. This inference includes backing that requires longitudinal follow-up (e.g., positive and negative consequences resulting from a decision). If a prospective doctor taking a

certification exam were to pass due to IPD whereby the items shift easier (e.g., due to cheating or item overexposure), he/she would then be allowed to practice medicine. This doctor may treat a patient effectively but could also do more than minimal harm on a patient (e.g., by providing an inaccurate diagnosis or prescribing the wrong treatment) because his/her true ability does not meet the requirements for minimal competence. Instead, his/her exam score is reflective of his/her ability plus familiarity with the test. Therefore, long-term follow-up may suggest that the exam is admitting candidates that are not actually qualified for practice. This may result in resetting the passing standard higher and consequently lowering the pass rate.

While the examples presented above speak mainly to IPD affecting scores through linking and equating, IPD may also operate more insidiously on scores through initial calibrations of item estimates. A bank scale may contain poorly estimated initial values due to calibration with inappropriate sample sizes, unrepresentative populations, seasonality effects, or timing of the calibration (Wyse & Babcock, 2016). As a result, pre-assembled test forms may be easier or harder than thought because the item estimates are not accurate reflections of the difficulty of the item. Therefore, regardless of the linking method used, and the appearance of the results, the obtained equated outcomes may still be inaccurate.

Technically speaking, if the claim(s) from one inference are not supported, then the subsequent inferences will also be invalidated. Thus, one cannot move to the next inference until the issues from the current inference are resolved. However, these examples illustrate how the IUA and validity argument are undermined according to each

inference when IPD goes untreated or undetected and requires the proper validation procedures. This includes appropriate detection measures, proper handling of drifted items, implementing robust linking methods, adherence to test security policies, consistent quality assurance, and following guidelines from empirical research. While an abundance of research is available on DIF (e.g., Haladyna & Downing, 2004), more studies are needed to examine which linking and equating methods handle drift the best.

Comparison of Unidimensional IRT Linking Methods under IPD

The majority of this section discusses research studies comparing linking methods under the context of drift, but an important question must first be asked. Are IRT linking and equating methods robust to IPD? The first part of this section will critique several studies that suggest IRT is robust to drift. The remainder of the section will examine the studies that have investigated one or more linking methods under the influence of drift.

IRT Robustness to IPD? Several studies have found a minimal effect of IPD on linking and suggest that Rasch (1960) and IRT models are robust to IPD (e.g., Rupp & Zumbo, 2003a; 2003b; Stone & Lane, 1991; Wells et al., 2002; Witt et al., 2003). However, this assertion is only partially warranted.

Wells et al. (2002) examined the effect of IPD on θ estimates under the 2PL model using the SL method. Three types of drift were analyzed: 1) the discrimination parameter (*a*-drift) was shifted by +0.5, 2) the difficulty parameter (*b*-drift) was shifted by +0.4, and 3) both parameters (*ab*-drift) were shifted by +0.5 for the discrimination parameter and +0.4 for the difficulty parameter. Four levels for the percentage of drifted items (5%, 10%, 15%, and 20%) in 40-item and 80-item tests with sample sizes of 300

and 1,000 examinees were simulated. Two testing occasions were simulated, both of which randomly sampled θ from a N(0, 1) distribution. BILOG 3 was used for calibration and EQUATE was used to carry out the SL method. RMSE, RMSD, and the mean absolute percentile difference (MAPD) were used to evaluate the recovery of θ estimates. IPD was found to have minimal impact on theta estimates and the authors concluded that IRT remained robust to parameter invariance. However, the authors note that the drifted items were not used in estimating the linking transformation, which may explain why the authors did not find any substantial impact of IPD. As noted by Han et al. (2012), IPD has both an effect on the item estimates as well as the linking constants. Another possible reason for the lack of significant findings is because studies have indicated that *a*-drift is harder to detect and has minimal impact on ability estimates (e.g., Donoghue & Isham, 1998).

Using the Wells et al. (2002) article to supplement their argument of IRT being robust to IPD, Rupp and Zumbo (2003a, 2003b) provided a theoretical and practical perspective on drift. Although they suggested that IRT models do yield relatively stable examinee scores in the presence of IPD, large amounts of drift can exacerbate outcomes.

Witt et al. (2003) examined the effects of drifting item difficulty on ability estimates and classification rates for a 100-item test administered to 187 examinees and a 200-item test administered to 260 examinees. Using the Rasch model, the authors manipulated the ability distribution to be negatively skewed, item difficulties were drifted by negative and positive values of 0.10, 0.25, and 0.50 logits, and the percentage of items drifted were 5%, 10%, and 25%. Winsteps was used to calibrate parameter estimates.

Misclassification rates were found to be no higher than what would be expected by measurement error and negligible differences existed between estimated θ and true θ. The authors concluded that the robustness of the Rasch model is evident even at the most extreme levels of drift and a large number of items (i.e., 25%) is needed to exhibit drift before θ estimates begin to deviate from their true value. Although misclassification rates were relatively low, an inspection of the mean ability and difficulty distributions suggest that classification rates should be low. The mean of the θ values for the 100-item and 200-item tests were 2.05 and 1.45, respectively, while the mean of the difficulty were -0.03 and 0.00. Given the disparity between θ and item difficulty, drift would probably need to be much larger than 0.50 logits to have a significant impact on classification rates.

Tracking the academic growth of preschoolers' math achievement between Fall and Spring instruction, Stone and Lane (1991) examined the stability of item parameter estimates from the Head Start Measures Battery. The assessment includes 19 free-response items, the first six of which are common to all students. Performance on these items determines whether the student receives six additional items for less-able children (Level I) or seven additional items for more-able children (Level II); thus, students do not receive all items in one administration. An unconstrained 2PL model was compared to a constrained 2PL model where discrimination and difficulty parameters were equal across time points. MULTILOG was used for parameter estimation, while G^2 was used to statistically compare the restricted (Model II) and unrestricted (Model II) models for best fit. A third hybrid model (Model III) was used when items from the unrestricted model

indicated drift over time—these items were allowed to vary while the remaining (stable) items were constrained. When comparing the item parameter estimates between seasons with Model III, only eight of the 38 estimated difficulty and discrimination parameters were found to drift. The authors concluded that the parameter estimates were moderately stable between testing occasions. However, two of the six items (33%) were common items flagged for drift in difficulty. With this proportion of drift and magnitudes of drift approaching 1.0, performing linking could produce inaccurate item parameter estimates and may not be as stable as reported.

In summary, these studies illustrate that IRT is a robust model, but they do not accurately capture the full impact of IPD. Although Stone and Lane (1991) found moderate stability of parameter estimates, one-third of the common items were affected by drift and would most likely impact linking outcomes. Both the Wells et al. (2002) and Witt et al. (2003) studies only simulated *b*-drift up to 0.5, but this amount of drift is not considered substantial unless test lengths, anchor set lengths, or sample sizes are small (e.g., Draba, 1977; Kopp & Jones, 2020; Risk, 2016; Wright & Douglas, 1976). While Rupp and Zumbo (2003a, 2003b) reiterate the findings of IRT robustness from Wells et al. (2002), they also acknowledge that IPD can have an extensive impact on ability estimates when drift is substantially large. Moreover, studies have found that it is not the *proportion* of drifted items that impact the accuracy of ability estimates and classification rates, but the *magnitude* of drifted items that is more detrimental to outcomes (e.g., Kopp & Jones, 2020; Li, 2012; Risk, 2016). Thus, only a few drifted items can have a profound impact if the extent of the drift is large enough.

Linking Method Studies Under IPD. Unless otherwise stated, the studies presented here investigate the performance of linking methods within the context of IPD under a unidimensional IRT framework using the CINEG design. Several studies investigating the impact of DIF were included (e.g., Huggins, 2014; Kabasakal & Kelecioglu, 2015; Yurtçu & Guzeller, 2018) because these studies also examined the effect on linking and equating.

Using an externally scored CINEG design, Hu et al. (2008) investigated the issue of whether to remove or ignore outliers (i.e., drifted items) for ten variations of four IRT-based linking methods: CC, FPC, SL, and MS. Forms Y_1, Y_2, and Y_3 were administered in Year 1, while forms X_1, X_2, and X_3 were administered in Year 2. Forms were linked and equated by their respective numbers (e.g., X_1 equated to Y_1). Within each pair of numbered forms, 10 common items were shared, and 72 unique items were presented (36 per form). Each form was comprised of multiple-choice (MC), short answer (SA), and open-ended response (OR) items. The item responses to base forms (Y) were randomly sampled from a N(0, 1) distribution, while the new forms (X) were randomly sampled from N(0, 1) and N(1, 1) distributions. Six combinations of number/score points (i.e., 0, 3, 9) and types of outliers (based on item type and content area) were examined. The six combinations included: 1) no outliers; 2) three MC items with three score points from one content area; 3) three MC items randomly chosen from one of five content areas; 4) three MC items with extreme *b*-parameter estimates (from -1.40 to -3.67); 5) five MC items and one OR item with nine score points from one content area; and 6) five MC items and one OR item with nine score points randomly chosen from one of five content areas. An

outlier was defined as any common item that exceeded two score points from the intersection point (two perpendicular straight lines drawn from each item's x-axis and y-axis position) of two plotted *b*-parameters from nonequivalent groups. However, only outliers located on the left side of the straight line (i.e., only items drifting easier) were investigated. Each form (e.g., Y_1) had 2,000 responses whose item parameters were estimated with PARSCALE under the 2PL, 3PL and GR models. Evaluation criteria included the MSE for *b*-parameters and MSE for number-correct true scores.

There are four takeaway points from Hu et al. (2008). First, all methods performed equally well (and better) without outliers and with equivalent groups. Second, the SL and MS performed better than CC and FPC without outliers under non-equivalent groups. Third, CC and FPC performed better than SL and MS when groups were equivalent, with 3 and 9 score point outliers included. Finally, no systematic pattern could be determined when groups were non-equivalent, with 3 and 9 score point outliers included or excluded. The authors suggested removing outliers as opposed to keeping them in the common item set and recommended the SL and MS methods for linking. However, if the groups to be linked are homogeneous, the use of CC and FPC is also acceptable. One interesting question is how much more the estimated equated scores would have been affected for each of the linking methods (and if the same pattern of results would hold) if an internal anchor was used, as Jurich et al. (2012) found the accuracy of equated scores was better under an external anchor design.

Examining the impact of cheating on the recovery of equated scores obtained with IRT true score equating and linking constants, Jurich et al. (2012) compared the

performance of the MM, MS, SL, Haebara, and FPC methods in a simulation study with 100 items and 3,000 responses per form. Factors that were considered included the proportion of cheating examinees (5%, 10%, 25%), the proportion of compromised items (25% and 100%), anchor item methods (external versus internal), and new form ability distributions of N(0,1), N(-0.5,1), N(0,1.25), and N(-0.5, 1.25). BILOG-MG was used for calibration under the 3PL model and results were evaluated in terms of bias and RMSE of the linking constants and equated scores. Results indicated that linking methods had little impact on the linking constants and equated scores.

Similar to Hu et al. (2008), a dissertation by Chen (2013) investigated whether drifted items should be included or excluded from linking in a simulation study comparing the CC, FPC, and SL methods. Item parameter estimates from a 60 multiple-choice item (30 unique and 30 common) real math assessment administered in consecutive years contained were treated as generating item parameters. Using a modified 3PL model where the pseudo-guessing parameter was fixed to 0.2 for all items, Chen manipulated the percentage of drifted items (10% or 25%), type and magnitude of drift (a shifted by ± 0.4, b shifted by ± 0.2 or ± 0.4), and group ability distributions. For the first year, θ was randomly sampled from a N(0, 1) distribution. Three different normal distributions were used in the second year: N(0, 1), N(0.2, 1), and N(-0.2, 1) distributions. PARSCALE 4 was used for calibration and the accuracy of θ estimates was assessed by examining bias, RMSE, and classification rates.

Chen (2013) made several conclusions which can be briefly summarized. Drifted items had little impact on the performance of FPC and SL methods, but CC only

performed as well as the FPC and SL when drifting items were removed from linking. Interestingly, when drifting items were removed from linking, CC estimated θ more accurately as drift increased or if the drift was positive (item became harder). Furthermore, CC did a better job when the two groups are of equal ability or when the mean ability of the year two group was higher than that of the year one group.

While Chen's (2013) study provides some insight into a limited field of research, there are a couple limitations that should be addressed. First, the conditions of the study did not allow for an evaluation of IPD at high magnitudes of drift. Similar to previous studies (e.g., Donoghue & Isham, 1998; Wells et al., 2002; Witt et al., 2003) Chen used *b*-drift values no greater than 0.4. As mentioned earlier, magnitude of drift has been found to be more important than the proportion of items exhibiting drift, and studies have manipulated drift up to 1.0 units (e.g., DeMars, 2004b; Kopp & Jones, 2020; Risk, 2016), with guidelines suggesting drift of 0.5 units or less to be acceptable or commonly used in practice (e.g., Draba, 1977; Han & Guo, 2011; O'Neill et al., 2013; Wright & Douglas, 1976). Second, the finding that CC performed better as drift increased seems unlikely because greater drift typically leads to worse linking outcomes (e.g., Kopp & Jones, 2020; Risk, 2016); however it is not implausible because item parameters were only shifted by 0.2 and 0.4 to introduce drift. Furthermore, the findings that FPC and SL performed similarly with and without drifted items contradicts a large body of research suggesting that outliers/drifted items should be mitigated, unweighted, or removed from the common item set prior to linking and equating (Bejar & Wingersky, 1981; DeMars, 2004b; Donoghue & Isham, 1998; He & Cui, 2020; He et al., 2015; Huynh, & Meyer,

2010; Li, 2012; Stocking & Lord, 1983; Wollack et al., 2006; Veerkamp & Glas, 2000; Vukmirovic et al., 2003).

Keller and Keller (2015) investigated the performance of the SL, FPC, and CC methods when test form content changes from year to year. Test form composition, anchor test composition, and examinee ability distributions were considered. A total of four forms were administered with two test form composition scenarios: 1) where all content was represented and the anchor was a miniature version of the entire test, and 2) where the content on various forms changed across administrations. Each form shared three out of five content areas with other forms, so anchor test composition was based upon the content areas shared between each of the four forms. Examinee ability had three different levels. In the first, ability didn't change from year to year (null condition). In the second, there was a multidimensional mean shift in the ability distribution whereby examinee ability improved by 0.10 and an additional shift of 0.05 was implemented for specific domains that would hypothetically receive more instruction time due to the extra content allotted to a given form (mean shift condition). Finally, there was a skewed condition where not all examinees exhibited the same amount of growth across administrations whereby skewness changed by -0.25 between each administration to reflect a situation where lesser-abled examinees became more able over time. PARSCALE was used to fit the 3PL model for FPC and STUIRT used to carry out the SL scale transformation, while BILOG-MG was used for CC. RMSE and bias were evaluated for the accuracy of parameter estimation as well as classification consistency. Results indicated that CC and FPC typically performed better than SL. When groups

were equivalent or there was a skewed shift, FPC was more robust to changes in content representation. However, CC performed better in the mean shift condition. In terms of classification, CC had the most accurate rates in all conditions, while SL and FPC performed comparably. Overall, it was determined that CC produced more stable results than the SL method.

Wollack et al. (2005) examined the effect of naturally occurring drift on a German placement test over a seven-year period. Ten different linking designs varying in method (i.e., FPC, CC, SL), direct or indirect linking, and with or without drift testing were considered. MULTILOG was used for calibration under a modified 3PL model with the pseudo-guessing parameter constrained to 0.2 and EQUATE was used for the SL method. RMSD was evaluated for ability estimates and equated scores with IRT true score equating The authors concluded that choice of linking method and IPD model could have a large effect on ability estimates and passing rates, although they could not distinguish which linking method was robust to IPD. Closer inspection revealed that items did not drift to the extent that would allow for them to detect differences in outcomes.

In a follow-up to their investigation of naturally occurring drift (Wollack et al., 2005), Wollack et al. (2006) examined the impact of compounding IPD in both a simulated and real data set. Using a 3PL model, the authors crossed the magnitude of IPD (drift of 0.25 and 0.40 units) with the ability distribution shifting by 0 or 0.15 every year for 5 years. The performance of the SL and FPC methods were compared using MULTILOG 7.0 for calibration. RMSE and bias were examined for the recovery of ability and item parameters. The authors found that the linking method was unaffected by

the magnitude of IPD but affected by increased ability. FPC was more influenced by changes in ability than SL. Due to the uncertainty of how much items may drift and whether the ability distributions may change over time, the authors recommended using the SL method with IPD testing. An application of the simulation findings to an empirical example produced consistent findings, albeit less pronounced.

Using data from a large-scale math state assessment administered to grades 3, 6, and 7, Arce Ferrer & Bulut (2017) compared the performance of SL and CC methods on IPD detection rates and magnitude of linking constants and equated cut scores. MULTILOG was used to calibrate responses under the 3PL model, while STUIRT and POLYEQUATE were implemented to carry out the SL transformation and IRT true score equating, respectively. When anchor sets were stable (i.e., no IPD detection method was used and anchor items were assumed to be stable), the SL and CC approaches lead to similar linking constants. However, the equated cut scores had more precision when using the SL method.

Investigating the effect of DIF on the stability of parameter estimation, Kabasakal and Kelecioglu (2015) compared traditional item response models (IRMs) (e.g., Rasch, 3PL model) to multilevel item response models (MIRMs). MIRMs combine hierarchical linear models with item response models, allowing for the examination of the effects of covariates (e.g., sex, race). The following factors were considered: sample size (500 or 2,000 per form), total items (20 or 40), and magnitude of DIF (0.6 and 1.0). Under the 1PL model, parameter estimates for CC were calibrated using BILOG-MG and PARSCALE was used for SL. IRTEQ was used for the SL scale transformation. Bias and

RMSE were calculated to examine the stability of item and ability parameter estimates. The MIRMs produced less error in smaller sample sizes and shorter test lengths than SL and CC, but because it's insensitive to increases in sample size and test length, it was concluded that MIRMs are best useful for small sample equating. Between CC and SL, CC was less affected by the presence of DIF items and errors decreased more as sample size and test length increased.

A simulation study was conducted by Sukin and Keller (2008) to examine the effect of retaining or removing a single drifted common item on classification rates. The performance of MM, MS, SL, and Haebara methods were compared under the 3PL model. The common item was drifted by 0.5 or 0.8 units. Ability estimates for the base form were randomly sampled from a N(0, 1) distribution, while the new forms were randomly sampled from N(0, 1) and N(0.2, 1) distributions. PARSCALE was used for calibration and STUIRT for linking. Classification rates were not affected whether the aberrant item was retained or removed and no differences between linking methods were observed.

Evaluating the effect of DIF on equating error, Yurtçu and Guzeller (2018) compared the performance of the MM, MS, SL, and Haebara methods. A total of 1,000 responses per form were simulated from a N(0, 1) distribution. Among 55 items, 15 were common items with either five or ten items exhibiting DIF. PARSCALE was used for calibration under the 3PL model and IRTEQ for linking and equating. Evaluation of RMSD values indicated that the characteristic curve methods (SL and Haebara)

performed better than the moment methods (MM and MS), with SL performing slightly better than Haebara.

The LAV and Area-Weighted (AW) methods were proposed by He et al. (2015) as new robust scale transformation methods to be used with the IRT CINEG. Similar to the LAV, the AW assigns a weight using a Huber function, instead of using the absolute difference between two ICCs of the equated test forms (equation 2.19). Compared to the SL method, outcomes were evaluated in terms of bias and RMSE for the recovery of item parameters, while the weighted absolute bias and weighted RMSE were used to evaluate equated scores obtained with the IRT true score equating method. One dichotomous item was drifted in both *a* and *b* parameters. A random number from a uniform distribution U(0.1, 0.5) was used to drift *a*, while *b* varied under four mild to moderate conditions: U(0.1, 0.5), U(-0.5, 0.1), U(0.5, 1.0), and U(-1.0, -0.5). Base form responses from 1,000 examinees were randomly sampled from a N(0, 1) distribution, while three new form responses were randomly sampled from N(0, 1), N(0.25, 1.1), and N(0.5, 1.2) distributions. BILOG-MG3 was used for calibration under the 3PL model. Results indicated that the AW and LAV were slightly less accurate than SL without outliers but were more accurate under the presence of outliers. The AW and LAV methods produced similar amounts of bias but the LAV had less RMSE than the AW method.

In a follow-up to their 2015 study, He & Cui (2020) examined the performance of the LAV, AW, SL with outlier elimination (if absolute difference of *a* or *b* parameters > 0.5), and SL with Raju's differential functioning of items and tests (DFIT). The following factors were considered: total items (45 or 120), common items (15 or 40), drifted items

(0, 1, or 3), drift magnitude (*a* randomly varied between from a uniform distribution of 0.1 and 0.5, *b* varied from a uniform distribution of -0.5 and -0.1, or -1.0 and -0.5), and ability distributions. 3,000 responses to the base form were randomly sampled from a N(0, 1) distribution and 3,000 new form responses were randomly sampled from N(0.25, 1.1) and N(0.5, 1.2) distributions. BILOG-MG3 was used for calibration under the 3PL model. RMSE and bias was used to evaluate the recovery of parameters and the weighted absolute bias and weighted RMSE was used to evaluate equated scores obtained with IRT true score equating. Although the LAV occasionally produced larger bias values than the elimination and DFIT methods, it yielded lower RMSE values under almost all conditions. The LAV also performed better than the AW and was concluded to be the best linking method overall.

Based on a statewide assessment test administered to seventh graders in subsequent years, Han et al. (2012) examined the effect of different multidirectional drift patterns on linking and equating outcomes. The performance of the MM, MS, and SL methods were compared based on classification errors, and the RMSE of the linking constants, linked item parameter estimates, and proficiency estimates. An external linking design was used on ten test forms administered each year. Among the ten forms, there was a total of 40 unique items and 20-30 external linking items. For the first year, 50,000 examinees were drawn from a N(0, 1) distribution. For the second year, 50,000 examinees were drawn from a N(0.1, 1) distribution. Four patterns of multidirectional drift were used that varied in IPD direction (unidirectional, bidirectional), to or from the mean item difficulty, and changes in standard deviation. The magnitude of drift varied by

0, ±0.25, and ±0.50. PARSCALE was used for calibration and IRTEQ was used for linking and equating. Han et al. (2012) found that multidirectional drift doesn't necessarily cancel or "wash-out" itself out. The MM method was found to be consistently robust against multidirectional drift, regardless of the IPD pattern.

The following studies have not examined the performance of different linking methods under IPD. However, they are reported here because they speak to the impact of drift on linking and equating outcomes, as well as the various factors (e.g., ability distributions, proportion of drifted items, magnitude of drift) that influence linking and equating outcomes.

In investigating the longitudinal scale stability of small sample licensure programs, Kopp and Jones (2020) examined the performance of FPC within the context of IPD. Sample sizes were manipulated to consist of 10, 25, and 50 examinees with an ability shift ($\Delta\theta$) of -0.1, 0, and 0.1 each year over the course of seven years. Of the 200 items on the test, 80 (40%) items were randomly chosen to serve as common items. The proportion of drifted common items included 0%, 10%, and 20% of items, while the magnitude of the drift shifted items by ±0.2, ±0.5, or ±1.0. Under the Rasch model, parameters were estimated using WINSTEPS. Classification accuracy and ability precision were evaluated according to bias and RMSE. Positively biased ability estimates were found when items became easier, whereas ability estimates were negatively biased when items drifted harder. RMSE was inflated at the smallest sample size (N = 10) or when the magnitude of drift was at 0.5 or higher. Classification accuracy was also

problematic at 0.5 drift or higher. Therefore, findings indicated that the magnitude of drift was more influential than the proportion of drifted items.

Examining whether to remove or keep polytomous items exhibiting drift, Li (2012) evaluated the impact on linking and true score equating results. Among a 60-item mixed format test, there were two sets of 20 items or four sets of 40 common items, with each set consisting of one, two, or four polytomous items. Drift was simulated according to weighted root mean squared differences (WRMSD) of 0.10, 0.15, and 0.20. The ability distribution of the group taking the base form was a N(0,1) distribution, while the ability distribution of the groups taking the new form were N(0,1), N(0.25, 1), and N(0.5, 1). The 2PL and GPC models were used to estimate parameters using PARSCALE. Linking constants were evaluated for each of the conditions by inspecting RMSE values and equating results were examined using weighted root mean squared errors (WRMSE). Results showed that as IPD increased, linking and equating errors also increased. It was also found that longer anchor lengths and fewer drifted items were associated with better linking and equating results. The one factor that did not have an effect on results was the difference of ability distributions. It was concluded that items exhibiting drift should be removed as a notable improvement in results was evident.

Using the Rasch model, Risk (2016) evaluated the impact of IPD on computer-adaptive testing (CAT). The following factors were manipulated: total bank items (300, 500, and 1,000), items drifted in the bank (50, 75, and 100), and the drift magnitude (0.5, 0.75, and 1.0). Drift was also simulated to be multidirectional, with 75% of the items becoming easier and 25% becoming harder. θ estimates for 500 examinees were sampled

from an ability distribution coming from a high-stakes certification exam with N(0.93, 0.73). Bias, RMSE, and absolute average difference (AAD) were used to measure θ estimates. Classification accuracy was evaluated by misclassification rates, as well as the number of false positives and false negatives. It was found that the magnitude of drift has a greater impact on the precision of scores than the number of items with IPD in the item bank. No systematic pattern appeared for total misclassifications, but more false-positives occurred at higher magnitudes of drift (1.0 logits), whereas more false-negatives occurred at lower magnitudes of drift (0.5 logits).

Under the FPC method, Vukmirovic et al. (2003) investigated whether to retain or remove outliers for linking and equating under the presence of IPD. Dichotomous and polytomous items were drifted to account for 10%, 20%, and 30% of the total points on the test, drift was unidirectional or bidirectional (50% easier and 50% harder), and ability distributions consisted of the means of distributions to 0, 0.25, and 0.50. Data was fitted to the 2PL, 3PL, and GR models using PARSCALE for calibration. The RMSD was used to evaluate the differences between TCCs and θ estimates. Findings suggested that outliers had a significant impact on FPC, especially with unidirectional drift, and should be removed prior to linking and equating.

Examining the effects of DIF on anchor items in subpopulations of examinees, Huggins (2014) used the MM, MS, SL, and Haebara methods to evaluate the differences. A total of 50 dichotomously scored items with 10 anchor items were simulated. DIF was manipulated across populations and forms with three levels of DIF magnitude (0.30, 0.60, and 0.90), three levels of proportion of DIF items (20%, 40%, and 60%), directionality of

DIF (unidirectional or bidirectional), mean differences in subpopulation ability levels (none or mean differences), and differential anchor form DIF (DIF in both anchors or DIF in one anchor form). The 3PL model was used and calibration was carried out with BILOG-MG. R was used for the simulation and analyses. Results were evaluated in terms of RMSD, root expected mean square difference (REMSD), root expected squared difference (RESD), and root squared difference (RSD) on true equated scores. Findings indicated that the MM, SL, and Haebara methods were more robust to DIF while the MS method was negatively influenced by the DIF introduced into the *b*-parameters. Furthermore, when DIF varied across forms, score equity between subpopulations was compromised.

Evaluating the consequences of IPD on linking and equating, Han (2008) carried out three simulation studies. In examining the effect of unidirectional drift (Study 1), generating item parameters from a K-12 statewide math assessment were used to simulate two dichotomously scored 40-item test forms with 10 common items. There were five levels of the percentage of drifted items (10%, 20%, 30%, 40%, and 50%) and the magnitude of drift was simulated from 0.05 to 1.00 in increments of 0.05. 5,000 responses per test form were randomly drawn from a N(0, 1) distribution. The MS method was used for scale transformation with the item estimates calibrated by PARSCALE under the 3PL model. Classification rates were examined, as were the RMSE and bias between estimated and true linking constants and item parameter estimates. Increasing the magnitude and proportion of drifted items resulted in heavily

affected linking constants and parameter estimates. Furthermore, misclassification rates climbed as drift increased.

Although the use of the MS method may have exacerbated results, Study 2 by Han (2008) compared the performance of the MM, MS, and SL methods with bidirectional drift. The MM method yielded the most unbiased estimation for *a*-parameters, while the SL method produced the most unbiased estimation of *b*-parameters. Finally, in Study 3, the MS and SL methods were considered when *c*-parameters were manipulated under four different calibration strategies. Findings indicated that SL outperformed MS with internal anchors, whereas the performance between the two methods was relatively similar when using external anchors.

Stahl & Muckle (2007) examined multidirectional drift using the displacement statistic in Winsteps for the Rasch model. The following factors were manipulated: total items (30, 100, and 200), the percentage of drifted items (10%, 20%, and 50%), and the type of drift (symmetrical with all items drifting one direction and asymmetrical with 70% of items drifting easier and 30% drifting harder). They found that artificial positive displacement (i.e., artificial drift) was more pronounced when drift was unidirectional. When drift was manipulated so that 70% of items became easier and 30% became harder, the effects of artificial positive displacement were ameliorated to a lesser extent. When drift was symmetrical (50% easier and 50% harder), there were no problems with displacement.

Babcock & Albano (2012) also used the Rasch model to investigate multidirectional drift on longitudinal scale stability in a high-volume certification testing

program. The proportion of drifted items, direction of the drift, and amount of latent trait change ($\Delta\theta$) were manipulated. Five levels of proportion of newly calibrated items were chosen for drifting (.00, .05, .10, .15, and .20) every year for five years, while the direction of drift could be easier, harder, or a combination of both every year for five years. The $\Delta\theta$ included a 1%, 5%, or 10% change over 20 years depending upon job analysis (JA) updates (every 6 years for full JA and every 3 years for interim JA). θ was randomly sampled from a distribution with a mean of 1.75 and a variance of 0.51, a common distribution in credentialing programs. WINSTEPS was used for calibration under the FPC method. RMSE and bias were evaluated between true and estimated item and person parameters, as well as the pass rate and classification accuracy. Findings indicated that a Rasch scale can maintain stability for about 15 years under little item drift and small to moderate changes in ability. However, large amounts of drift or substantial changes in ability greatly reduced the longevity of the scale.

While results from studies looking at linking methods *without* drift were very conclusive (e.g., FPC and in particular, CC, performed the best), findings from studies examining linking methods *with* drift were rather ambiguous. However, several conclusions can be drawn.

First, there is no single linking method that has performed the best. CC only performed the best in several studies (i.e., Kabasakal & Kelecioglu, 2015; Keller & Keller, 2015) and when groups were equivalent (i.e., Hu et al., 2008); however, linking is unnecessary when groups are equivalent. Several studies found SL to perform the best (i.e., Arce-Ferrer & Bulut, 2017; Chen, 2013; Wollack et al., 2006; Yurtçu & Guzeller,

2018) and Hu et al. (2008) found SL was best with nonequivalent groups. FPC has received support from Chen (2013), who found FPC and SL to be comparable, and similar to CC when used with equivalent groups (Hu et al., 2008). Furthermore, a couple studies have found no difference between linking methods (e.g., Jurich et al., 2012; Sukin & Keller, 2008).

Second, the Haebara method remains understudied, as most researchers have opted to use the SL method. Results from studies not examining drift have found the Haebara method to be comparable to SL (e.g., Hanson & Beguin, 2002; Keller & Keller, 2011; Kim & Kolen, 2007; Lee & Ban, 2010; Li et al., 2012). When considering drift, Yurtçu and Guzeller (2018) found Haebara and SL to perform better than the moment methods, with a slight edge to SL. On the other hand, both Jurich et al. (2012) and Sukin and Keller (2008) found no difference between Haebara and SL.

Third, the LAV method appears to be a promising new robust scale transformation method (i.e., He & Cui, 2020; He et al., 2015). However, to date, only the two studies have investigated the LAV's performance in comparison to the SL method. More research is needed to evaluate the LAV's performance under different conditions and against other linking methods.

Fourth, the magnitude of drift has more of a profound impact on linking and equating results than the proportion of drifted items (e.g., Kopp & Jones, 2020; Li, 2012; Risk, 2016). Studies that have not reported an effect of drift on linking and equating (e.g., Wells et al., 2002; Witt et al., 2003) only simulated a small amount of drift (e.g., a magnitude of drift less than <0.5), which is not considered substantial unless test lengths,

anchor set lengths, or sample sizes are small (e.g., Draba, 1977; Kopp & Jones, 2020; Risk, 2016; Wright & Douglas, 1976).

Fifth, it is unclear how much of an affect different ability distributions, in conjunction with IPD, have on linking and equating results. Ability differences have been found to differentially affect linking methods (e.g., Chen, 2013; Hu et al., 2008). Although Babcock and Albano (2012) found ability differences to have a significant effect on longitudinal scale stability, findings have generally indicated ability to have no effect on linking and equating (e.g., He et al., 2015; Li, 2012; Witt et al., 2003).

Finally, few studies elaborate on the effect of IPD in relation to its impact on validity and validation (e.g., Kabasakal & Kelecioglu, 2015; Risk, 2016). The *Standards* are referenced by several studies (e.g., Arce-Ferrer & Bulut, 2017; Babcock & Albano, 2012; Huggins, 2014) pertaining to proper procedures for linking and equating, the handling drifted items, maintaining scale stability, or fairness. He and Cui (2020) discuss the importance of maintaining content representativeness when linking and equating, which the LAV method seeks to preserve. Guidelines from the International Test Commission are referred to by Arce-Ferrer and Bulut (2017) in the context of handling drifted items and by Huggins (2014), who discusses IPD's impact on fairness. Han (2008) and Han et al. (2012) elucidate the influence of drift as a threat to construct validity, fairness, and the need for IPD analyses as part of validity evidence for test validation. Aside from these studies, validity is not considered to the extent that it should be.

Taken altogether, further research is needed to determine which linking method(s) perform best under different conditions of drift and the extent to which IPD affects parameter estimates, linking constants, equated scores, and classification rates. More attention should focus on the Haebara and LAV methods, as both appear to be comparable to the SL method. While the magnitude of drift has been reported to have more of an effect on equating than the proportion of drifted items, the role of ability remains equivocal. Furthermore, the impact of IPD on validity and validation warrants more detailed analysis, as do the consequences that may ensue when improperly monitored or handled.

CHAPTER III

METHODS

This chapter is broken up into two sections. The first section details the procedures, conditions, and evaluation criteria used for the simulation study. The second section provides an overview of the procedures and evaluation criteria used for data from a real certification examination.

Overview

A simulation study was conducted to examine the effect of IPD on five linking methods (i.e., SL, Haebara, CC, FPC, and LAV) used to link two forms administered in separate years. The items on the forms, and examinee responses to the items, imitated those found from a large-scale certification exam. The simulation included variations of the following conditions: (1) the proportion of drifted items, (2) the magnitude of the drifted items, (3) examinee ability differences, and (4) sample size. The study evaluated each linking method's performance based upon recovery of linking constants, recovery of item parameters, equating accuracy, and classification rates.

Simulation Design

Data Generation. Two dichotomously scored 100-item test forms were created for the simulation study. Although this number of items is longer than a typical educational achievement test of approximately 60 items (e.g., Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Kang & Petersen, 2011), licensure and certification programs

require a larger number of items to ensure that a candidate displays minimal competence over a range of skills and abilities specified in the job analysis (e.g., Kane, 1982). A total of 20 common items (20%) were shared between forms and generated to be statistically similar in difficulty to their respective forms. This number of common items is consistent with research suggesting anchor sets between 20-40% are efficient and practical (e.g., Budescu, 1985; Kolen & Brennan, 2014). Item parameters were generated using the catR package (Magis & Raîche, 2012) in R (R Core Team, 2017) with the 3PL model:

$$P(Y_j = 1|\theta) = c_j + (1 - c_j)\frac{\exp[Da_j(\theta - b_j)]}{1 + \exp[Da_j(\theta - b_j)]} \quad (2.23)$$

where the probability of a correct response to an item, $Y_j = 1$, is based upon examinee ability (θ), a_j is the item discrimination parameter, b_j is the item difficulty parameter, c_j is the item pseudo-guessing parameter, D is a scaling constant set to 1.0 for this study, and exp is an exponential constant with a value of 2.718. These values constitute the generating item parameters upon which parameter estimates are compared to.

Item discrimination was randomly sampled from a normal distribution with a mean of 1 and standard deviation of 0.3, bounded between 0.5 and 1.5. Item difficulty was randomly sampled from the standard normal distribution, bounded between -3.0 and 3.0. The pseudo-guessing parameter was randomly sampled from a uniform distribution between 0.05 and 0.35.

Response data was generated in R by computing the probability of correctly answering an item given a certain θ. The probability of correctly answering each item was then compared to a random number from a uniform distribution bounded between 0

and 1. If the probability of a correct response was greater than the uniform number, then the response was scored as correct.

Linking Methods. The SL, Haebara, CC, FPC, and LAV methods were used in this study. The flexMIRT software (Cai, 2017) was used to calibrate item parameters with SC, CC, and FPC. The R package equateIRT (Battauz, 2015) was used to perform linking for the SL and Haebara methods. R code was provided by He et al. (2015) to implement the LAV method. No linear transformation was required for the CC and FPC methods as the item estimates were already on the same scale after calibration. The equateIRT package provided equated scores for IRT true score and observed score equating for all linking methods.

Conditions. The following conditions have been identified by research as important factors within the context of IPD. Varying levels of each of the conditions helped to identify which linking methods performed most robustly to drift.

Proportion of Drifted Items. The proportion of items exhibiting drift within high-stakes certification exams is likely to vary based upon a number of factors like how often the items have been used on other forms (i.e., risk of overexposure). The longer the test blueprint has gone without a new job analysis could also contribute to items changing over time, particularly when recent news or findings draws attention to more obscure topics (e.g., O'Neill et al., 2013). Items that may have once been harder could become easier as a result of improved candidate training (e.g., Kopp & Jones, 2020).

A testing program that maintains tight security protocols and continuously updates its testing cycle and forms administered may have few items that drift, whereas a

program that is subjected to a security breach may result in a substantial number of drifted items. Thus, the proportion of drifted items was set to 0%, 25%, and 50% of the anchor items. Although these proportions might seem extreme, they have been observed in a couple of studies (e.g., Jurich et al., 2012; Stahl & Muckle, 2007). In their study examining the effects of compromised items and cheaters, Jurich et al. (2012) simulated the proportion of compromised anchor items to be 100%, representing a scenario where items were exposed after first administration. Furthermore, the purpose was to examine how robust each linking method is to drift, even in extreme circumstances.

Magnitude of Drifted Items. Research has found magnitude of drift to have more impact on linking outcomes than the proportion of drifted items (e.g., Kopp & Jones, 2020; Li, 2012; Risk, 2016). Only *b*-drift was considered for this study because most certification programs only consider the difficulty parameter and because *a*-drift is difficult to detect (e.g., Donoghue & Isham, 1998). This study only focused on unidirectional drift where the items become easier over time because most reasons (e.g., cheating, item overexposure) suggest there are more potential scenarios for this type of drift. Furthermore, studies have found multidirectional drift to have little effect on linking outcomes and less influential than unidirectional drift (e.g., Babcock & Albano, 2012; Stahl & Muckle, 2007).

The magnitude of drifted items included a change of -0.25, -0.50, and -1.00 in the difficulty parameters. These reflected similar magnitudes that have been used in other studies (e.g., Kabasakal & Kelecioglu, 2015; Kopp & Jones, 2020; Risk, 2016; Sukin & Keller, 2008). Studies that have not found an effect of drift on linking outcomes have

only simulated *b*-drift up to 0.5 (e.g., Wells et al., 2002; Witt et al., 2003), but this magnitude is not considered substantial in practice (e.g., Draba, 1977; Kopp & Jones, 2020; Risk, 2016; Wright & Douglas, 1976).

Ability Distributions. One of the more contentious factors is the role of different ability distributions, combined with drift, on the effect of linking outcomes. Some studies have reported no effect on outcomes (e.g., He et al., 2015; Li, 2012; Witt et al., 2003), whereas others have found an effect on scale stability and linking methods (e.g., Babcock & Albano, 2012; Chen, 2013; Hu et al., 2008).

Negatively skewed ability distributions are commonly observed in practice (e.g., Kim & Lee, 2017), particularly in licensure and certification (e.g., Witt et al., 2003). This study considered five different ability distributions. Simulated examinees to the base form were randomly sampled from the standard normal distribution N(0, 1). The new form had normal distributions of N(0, 1), N(0.5, 1), and N(1, 1). A negatively skewed (S) distribution with a mean of 0.5 and standard deviation of 1, S(0.5, 1), was used for the fourth ability distribution. A negatively skewed distribution S(1, 1) was used for the final ability distribution. For both skewed distributions, a skewness of -0.75 was implemented (e.g., Kim, 2019; Pearson & Please, 1975). The R package sn (Azzalini, 2020) was used to generate the skewed distributions.

Sample Sizes. Sample size was an important consideration because it affects the stability of item calibration (e.g., Linacre, 1994; Lord & Wingersky, 1984). Although larger sample sizes are better for improving stability, the minimum sample size requirements for the 3PL model has varied. Some researchers have recommended using

1,500 examinees per form (e.g., Harris & Crouse, 1993; Kolen & Brennan, 2014) while others have suggested 1,000 examinees per form (e.g., Hulin et al., 1982; Swaminathan & Gifford, 1983).

The most frequently observed studies investigating linking with the 3PL model have used 1,000 examinees (e.g., Hanson & Beguin, 2002; He et al., 2015; Wollack et al., 2006; Yurtçu & Guzeller, 2018) and several others have used 3,000 examinees (e.g., Hanson & Beguin, 2002; He & Cui, 2020; Jurich et al., 2012; Lee & Ban, 2010). For this study, sample sizes of 1,000 and 3,000 were used as conditions.

Thus, this study had a total of 70 different conditions: 60 conditions for the 25% and 50% drifted item conditions (5 different ability distributions x 2 sample sizes x 2 levels of proportion of drifted items x 3 levels of magnitude of drift) and 10 conditions for the 0% drifted item conditions[1]. For each condition, a total of 100 replications were performed. Table 2 summarizes the conditions used in this study.

Table 2

Simulation Study Conditions

Condition	Levels
Proportion of Drifted Items	0%, 25%, 50%
Magnitude of Drifted Item Difficulties	-0.25, -0.50, -1.00
Ability Distributions	N(0, 1), N(0.5, 1), N(1, 1), S(0.5, 1), S(1, 1)
Sample Sizes	1,000 and 3,000

[1] Drift of 0% precludes the possibility of the magnitude of drifted items. The 10 levels reflect the 5 levels of ability * 2 sample sizes.

Evaluation Criteria. The first research question ascertained the effect of IPD on linking constants *A* and *B*. Although linking constants are not provided by the CC and FPC methods, the mean and standard deviation of the estimated theta distribution for the new form should be similar to *A* and *B*, respectively. Estimated linking constants were compared to the true linking constants. The "true" linking constant values were based upon the mean and standard deviation of each of the five ability distributions. For example, randomly sampling examinees from the standard normal distribution had true linking constants of *A*=1 and *B*=0. Likewise, a negatively skewed distribution with a mean of 0.5 and standard deviation of 1 had true linking constants of *A*=1 and *B*=0.5. The recovery of the linking constants were assessed by three criteria: bias, standard error (SE), and root mean squared error (RMSE). Each of the criteria are defined as follows:

$$Bias = \frac{1}{R}\sum_{r=1}^{R}\hat{l}_r - l\,, \tag{2.24}$$

$$SE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{l}_r - \bar{\hat{l}})^2}, \tag{2.25}$$

$$RMSE = \sqrt{Bias^2 + SE^2}, \tag{2.26}$$

whereby *R* is the number of replications (i.e., 100); $\hat{l}_r$ is an estimate of linking constant *A* or *B* for a given replication *r*; *l* is the true linking constant (*A* or *B*); and $\bar{\hat{l}}$ takes the average standard deviation from all replications of the linking constant.

The second research question examined how well each of the item parameters are successfully recovered. The estimated item parameters for the new form were compared

to the generating item parameters for the new form. Bias, SE, and RMSE were evaluated for the discrimination, difficulty, and pseudo-guessing parameters using formulas similar to the first research question:

$$Bias_j = \frac{1}{R}\sum_{r=1}^{R} \hat{v}_{jr} - v_j\,, \tag{2.27}$$

$$SE_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R} (\hat{v}_{jr} - \bar{\hat{v}}_j)^2}, \tag{2.28}$$

$$RMSE_j = \sqrt{Bias_j^2 + SE_j^2}, \tag{2.29}$$

whereby R is the number of replications (i.e., 100); $\hat{v}_{jr}$ is an estimate of item j for a given replication r (v refers to the item difficulty, discrimination, or pseudo-guessing parameter estimates); v_j is the same parameter for the same item; and $\bar{\hat{v}}_j$ takes the average standard deviation from all replications of the same item. For bias, the absolute values were averaged to prevent the cancellation of positive and negative values across items.

The third research question examined the extent to which IPD influenced true and observed equated scores. Equated scores obtained with IRT true and observed score equating was compared to the criterion equating relationship. The criterion equating relationship was defined as the equated scores obtained from the generating item parameters for the baseline condition. There were two criterion equating relationships – one for equated observed scores and one for equated true scores. For observed score equating, synthetic weights were set to 0.5 to reflect the equal examinee sample sizes of

the base and new forms. Similar to the first two research questions, bias, SE, and RMSE served as the evaluation criteria for both true and observed scores.

$$Bias(x) = \frac{1}{R}\sum_{r=1}^{R} \hat{e}_Y^{(r)}(x) - e_Y(x)\,, \tag{2.30}$$

$$SE(x) = \sqrt{\frac{1}{R}\sum_{r=1}^{R} [\hat{e}_Y^{(r)}(x) - \bar{\hat{e}}_Y^{(r)}(x)]^2}, \tag{2.31}$$

$$RMSE(x) = \sqrt{Bias(x)^2 + SE(x)^2}, \tag{2.32}$$

whereby *R* is the number of replications (i.e., 100); *x* is a particular score point; $\hat{e}_Y^{(r)}(x)$ is the estimated old form equivalent of score *x* obtained from the r^{th} replication; $e_Y(x)$ is the old form equivalent of *x* obtained using the generating item parameters; and $\bar{\hat{e}}_Y^{(r)}(x)$= $\frac{1}{R}\sum_{r=1}^{R} \hat{e}_Y^{(r)}(x)$. For bias, the absolute values were averaged to prevent the cancellation of positive and negative values across items.

The fourth research question examined the extent of IPD on classification accuracy rates. Classification accuracy was defined as the extent to which actual classifications using observed cut scores agreed with "true" classifications based on known true cut scores (Lee, 2010; Lee et al., 2002). If an examinee passes based on his/her true score, the examinee should also pass based upon his/her observed score. Although true scores cannot be observed, a set of quadrature points are used in its place (e.g., 49 quadrature points spanning from -6 to 6). For each quadrature point, it can be determined whether an examinee passes or fails – this reflects the "true" status. If the cut score is on the number correct score metric, the quadrature points (θs) can be converted

to number correct scores using the test characteristic curve. For each θ, it can be determined whether the number correct score is a pass or a fail. This procedure, which uses a distribution of θ, is referred to as the D method (Lee, 2010). The observed classification is obtained using the Lord and Wingersky (1984) recursion formula specified in equation 2.1. Then, the probability of passing or failing was computed based on the true status of the quadrature point. For a specific θ, the probability of failing is the sum of the conditional probabilities up to the cut score, whereas the probability of passing is the sum of the conditional probabilities from the cut score to the highest attainable score. Equation 2.2 is applied to integrate the conditional probabilities over all quadrature points.

The performance of each linking method's classification accuracy was compared to the true classification criterion, which is computed as the proportion of examinees that have been classified as pass-pass or fail-fail for both the true and observed θ status'. The D method was used to get the true classification criterion. There were five true classification criteria, one for each distribution condition. Code was manually written in R to compute classification accuracy for each linking method and results were evaluated using bias, SE, and RMSE.

Licensure and certification programs often have cut scores where the pass rate for first-time test takers is typically between 70 and 90% (e.g., Accreditation Council for Graduate Medical Education, 2018; Breitbach et al., 2013; Okrainec et al., 2011; Shea et al., 1991). Therefore, the cut score was set at the raw score equivalent associated with θ = -0.49 (57 out of 100) to align with a typical pass rate of 75% for a certification exam.

The last research question analyzed the impact of IPD on the consistency of classification rates. Classification consistency is the degree to which classifications agree over two independent administrations of a test (Lee, 2010; Lee et al., 2002). Most methods are developed to estimate consistency based upon a single form, but Lee's IRT method (2010; Lee et al., 2002) was used here. Similar to classification accuracy, the Lord and Wingersky (1984) recursion formula (equation 2.1) was used to come up with the conditional observed score distribution for each given θ. For a given θ, a classification is consistent when the two observed score statuses are either pass-pass or fail-fail. Assuming the two testing occasions are independent, the probability of passing from the first testing occasion is expected to be the same as the probability of passing for the second testing occasion. Thus, the probability of pass-pass is the probability of passing squared (this is also done for fail-fail). Therefore, the consistent classification rate is the squared probability of passing plus the squared probability of failing. As explained above, the D method was used, which assumes a distribution of θ to integrate the individual passing rates over all quadrature points using equation 2.2.

The performance of each linking method's classification consistency was compared to the true classification consistency criterion, which was obtained using the generating item parameters. Code was manually written in R to compute classification accuracy for each linking method and results were evaluated using bias, SE, and RMSE.

To examine items that may exhibit drift with the 3PL model, the DIF function in the R package mirt (Chalmers, 2012) was used. The likelihood ratio test was used to compare the likelihood values between two models in nested conditions: the baseline

model and a less constrained model (i.e., backward procedure). For the baseline model, all item parameters are constrained to be equal. In the less constrained model, one common item is freely estimated while all other items are constrained. If the likelihood value differs between the two models, then the item shows DIF. This is repeated for all common items.

Empirical Data Analysis

Data from two forms of a high-stakes certification program were analyzed. Each form had four different field-test blocks of 10 items with the same set of 110 scored items for a total of 120 items per form. There was a total of 66 internal common items (60%) and 44 unique items (40%) per form. All items were dichotomously scored multiple-choice items. Some items that were previously administered as field-test items on the base form were administered as scored items on the new form. Both forms were built to the same content specifications and approximate item difficulty parameters under the Rasch model. However, the current study utilized the 3PL model, so the items were re-estimated using flexMIRT (Cai, 2017) to include the difficulty, discrimination, and pseudo-guessing parameters.

A total of 1,990 candidates were administered the base form, while 1,979 candidates were administered the new form. A cut score of 83 out of 110 scored items was established for the base form, based on ratings from subject matter experts during a standard setting meeting. Rasch pre-equating was used to find the θ cut score on the new form that was equivalent to the θ cut score on the base form. Using the 3PL model, a raw cut score of 85 out of 110 scored items was set for the new form. The R package

equateIRT (Battauz, 2015) was used to link the SL and HB methods, while R code from He et al. (2015) was used for the LAV method. The PIE software (Hanson & Zeng, 1995) was used for IRT observed score and true score equating. PIE obtained scores below the sum of the pseudo-guessing parameters through linear interpolation. This was done to compare the results of observed and true scores. To examine items that may exhibit drift with the 3PL model, the DIF function in the R package mirt (Chalmers, 2012) was used.

Evaluation Criteria. For each research question, observed estimates for linking constants, item parameters, equated scores, classification accuracy, and classification consistency were computed for all linking methods. However, unlike the simulation, where observed estimates can be compared to true θ values and item parameters, the empirical data analysis does not have known θ values or item parameters. Therefore, it could not be determined which of the linking methods performed the best. Instead, the observed estimates were compared across linking methods to evaluate the similarity of their performance. The observed estimates were also used to validate the results from the simulation. In line with the simulation, the D method (Lee, 2012) was used to examine classification accuracy and consistency, but no statements could be made as to which linking method performed the best, only how similarly the methods performed between each other.

CHAPTER IV

RESULTS

This chapter is organized into three sections. The first section presents results of the simulation study. The second section presents results from the empirical data analysis. The third section relates the findings to implications for validation, based upon the frameworks of the *Standards* five sources of evidence, and Kane's argument-based approach to validation.

Simulation Study

The base form and the new form were both built to the same statistical specifications summarized in Table 3. Both forms were constructed to have a mean IRT difficulty of 0 and a standard deviation of 1, a mean discrimination of 1 and standard deviation of 0.30, and a pseudo-guessing parameter between 0.05 and 0.35. The common items were generated to have the same statistical specifications as each form. A list of all generating item parameters can be found in Appendix A.

Drift Detection. Prior to evaluating the impact of IPD on each research question, it was important to determine whether any common items, particularly those selected to drift, exhibited IPD. Tables 4 and 5 illustrate the percentage of common items that were accurately detected for drift under the 1,000 and 3,000 sample-size conditions, respectively.

Table 3

Descriptive Statistics for Generating Item Parameters

	Base Form		New Form	
	Mean	Standard Deviation	Mean	Standard Deviation
Common Items				
Discrimination	1.012	0.281	1.012	0.281
Difficulty	0.019	0.891	0.019	0.891
Pseudo-guessing	0.233	0.098	0.233	0.098
Unique Items				
Discrimination	0.987	0.248	0.965	0.261
Difficulty	0.019	0.995	0.020	1.075
Pseudo-guessing	0.195	0.078	0.201	0.087
All Items				
Discrimination	0.992	0.254	0.975	0.264
Difficulty	0.019	0.970	0.019	1.036
Pseudo-guessing	0.202	0.083	0.208	0.090

Results are reported using the likelihood ratio in the dif function of the mirt package (Chalmers, 2012; Kim & Yoon, 2011). In order not to inflate Type I error, the *p*-value for the likelihood-ratio tests was set to 0.0025 (.05/ 20 common items) for all conditions using the Bonferroni correction. This *p*-value was chosen because the 20 items are dependent upon each other, but the replications are independent from one another since they contain different examinee responses (e.g., Bland & Altman, 1995; Cabin & Mitchell, 2000). The no drifted item condition represents the Type I error rate because no common items were manipulated to drift. Thus, any detection is considered a false positive. The remaining rows represent the power of detecting drift. For the 25% drifted item conditions, the correct detection of the first five drifted common items was taken for all 100 replications and averaged for the reported percentage. For the 50% drifted item

conditions, the correct detection of the first ten drifted common items was taken for all 100 replications and averaged for the reported percentage.

Drift detection rates were higher for the 3,000 sample-size conditions, but both sample sizes followed the same general patterns. When no items were manipulated to drift, the Type I error rate was near the nominal alpha of .05 for all ability distributions, which is consistent with other drift detection studies (e.g., DeMars, 2004b; Donoghue & Isham, 1998). Detection rates were higher for the 25% drifted item conditions compared to the 50% drifted item conditions, but within each proportion of drifted item conditions (25%, 50%), the detection rates increased as the magnitude of drift increased. For the 1,000-sample size and the highest magnitude of drift (-1.00), detection rates ranged from 74% to 91% under the 25% drifted item condition. However, rates were much lower for the 50% drifted item condition, ranging from 26% to 49%. This lack of power might have been due to not having enough examinees to detect the difference with the 3PL model. When sample size increased to 3,000, these percentages increased for the highest magnitude of drift – near 100% for all ability distributions of the 25% drifted item conditions and from 69% to 90% for the 50% drifted item conditions. Interestingly, the correct detection rates decreased as the ability distributions (normal and skewed) moved further away from a mean of 0. This decrease could be attributed to linking results being affected greater by increases in the proportion and magnitude of drifted items, which results in fewer correctly detected items.

Table 4

Drift Detection Results – 1,000 Examinees

Drifted Items	Drift Magnitude	Ability Distribution				
		N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
None	None	6%	4%	4%	5%	4%
	-0.25	3%	4%	3%	3%	2%
25%	-0.50	31%	27%	19%	24%	15%
	-1.00	91%	86%	75%	82%	74%
	-0.25	1%	1%	1%	1%	0%
50%	-0.50	10%	7%	5%	7%	4%
	-1.00	49%	38%	27%	33%	26%

Table 5

Drift Detection Results – 3,000 Examinees

Drifted Items	Drift Magnitude	Ability Distribution				
		N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
None	None	5%	5%	5%	7%	7%
	-0.25	22%	16%	12%	18%	10%
25%	-0.50	82%	76%	62%	72%	62%
	-1.00	100%	100%	99%	100%	98%
	-0.25	7%	5%	4%	5%	4%
50%	-0.50	44%	38%	28%	35%	24%
	-1.00	90%	83%	72%	80%	69%

Linking Constants. The first research question examined the impact of IPD on linking constants *A* and *B*. Bias, SE, and RMSE were calculated to determine the performance of each linking method.

Linking Constant A. For all linking methods and conditions, the expected value of *A* should approximate 1 because the standard deviations for each of the focal group populations was set to 1 and the slope was not manipulated to drift. The estimates for linking constant *A* with 1,000 and 3,000 sample sizes are summarized in Tables 6 and 7, respectively. Values of bias, SE, and RMSE can be found in Appendix B. Figures 3 – 5 illustrate the bias, SE, and RMSE values for the 1,000 sample-size condition. Figures 6 – 8 illustrate the bias, SE, and RMSE values for the 3,000 sample-size condition.

It should be noted that the linking constants reported here represent values that have slightly different meanings but are considered "linking constants" strictly for the purposes of comparison. The linking constants for the separate calibration methods were derived only from the common items. However, the "linking constants" for CC and FPC were the estimated mean and standard deviation from the new group ability distribution, which considers all items from the exam, not just the common items. Therefore, the comparison between the linking methods is not completely impartial.

For the 1,000 sample-size conditions, the separate calibration methods tended to underestimate *A* (i.e., values were less than 1.00), whereas CC and FPC tended to overestimate *A* (i.e., values were greater than 1.00). When no drift was present, the separate calibration methods (i.e., SL, HB, and LAV) typically recovered *A* better than CC and FPC. That is, the separate calibration methods produced smaller amounts of

RMSE than CC and all ability distributions except N(1,1) for FPC. This is because CC and FPC had larger values of bias than the other linking methods. RMSE increased as the mean of the normal and skewed ability distributions moved further away from 0.

When 25% of the common items were drifted, RMSE increased for the separate calibration methods as the mean of the ability distributions moved further away from 0 for both normal and skewed distributions. No systematic pattern was evident for CC and FPC as the ability distributions moved further away from 0. As the magnitude of the difficulty of drifted items became easier (items were drifted by -1.00 difficulty), the separate calibration methods produced greater values of RMSE and bias, whereas FPC and CC produced smaller values of RMSE and bias. FPC tended to recover *A* the best in terms of RMSE and bias, particularly when the ability distributions deviated from N(0,1).

When 50% of the common items were drifted, estimates of *A* were less accurate for the separate calibration methods as the ability distributions increased. No systematic patterns were evident for CC or FPC. As drift magnitude increased (i.e., became easier), both FPC and CC recovered *A* more accurately, whereas the separate calibration methods recovered *A* less accurately. FPC produced the smallest RMSE and bias for nearly all conditions, including the most extreme drift condition (50% items drifted, -1.0 magnitude). The LAV method produced the highest RMSE values, possibly because drift was only manipulated in the difficulty parameter. As a result, the LAV may have attempted to minimize the effect of drift in difficulty and linking constant *B* at the expense of linking constant *A*. Furthermore, the LAV method had consistently higher values of SE than any other linking method for nearly all conditions of drift.

Table 6

Estimated Linking Constant A – 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	1.002	0.966	0.930	0.989	0.975
	25%	-0.25	0.988	0.952	0.916	0.985	0.960
		-0.50	0.970	0.938	0.903	0.964	0.951
		-1.00	0.939	0.897	0.862	0.931	0.912
	50%	-0.25	0.976	0.948	0.905	0.971	0.961
		-0.50	0.961	0.923	0.873	0.956	0.927
		-1.00	0.898	0.856	0.808	0.895	0.864
HB	None	None	1.006	0.973	0.939	0.996	0.985
	25%	-0.25	0.984	0.951	0.923	0.985	0.965
		-0.50	0.953	0.926	0.897	0.953	0.942
		-1.00	0.885	0.854	0.824	0.882	0.870
	50%	-0.25	0.968	0.943	0.906	0.966	0.960
		-0.50	0.932	0.901	0.859	0.933	0.907
		-1.00	0.813	0.780	0.742	0.819	0.794
LAV	None	None	1.006	0.976	0.944	0.994	0.987
	25%	-0.25	0.987	0.954	0.929	0.987	0.967
		-0.50	0.965	0.939	0.910	0.964	0.949
		-1.00	0.968	0.922	0.889	0.939	0.919
	50%	-0.25	0.968	0.948	0.915	0.963	0.964
		-0.50	0.925	0.897	0.858	0.925	0.902
		-1.00	0.772	0.748	0.739	0.773	0.760
CC	None	None	1.128	1.106	1.097	1.096	1.110
	25%	-0.25	1.114	1.094	1.099	1.095	1.102
		-0.50	1.094	1.087	1.092	1.081	1.103
		-1.00	1.061	1.060	1.080	1.059	1.089
	50%	-0.25	1.100	1.094	1.090	1.084	1.108
		-0.50	1.085	1.083	1.085	1.078	1.093
		-1.00	1.032	1.056	1.084	1.052	1.088
FPC	None	None	1.077	1.054	1.044	1.047	1.052
	25%	-0.25	1.062	1.042	1.042	1.045	1.042
		-0.50	1.042	1.034	1.033	1.026	1.040
		-1.00	1.014	1.006	1.016	1.002	1.019
	50%	-0.25	1.046	1.040	1.033	1.032	1.047
		-0.50	1.033	1.026	1.020	1.023	1.027
		-1.00	0.981	0.988	0.999	0.979	1.001

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

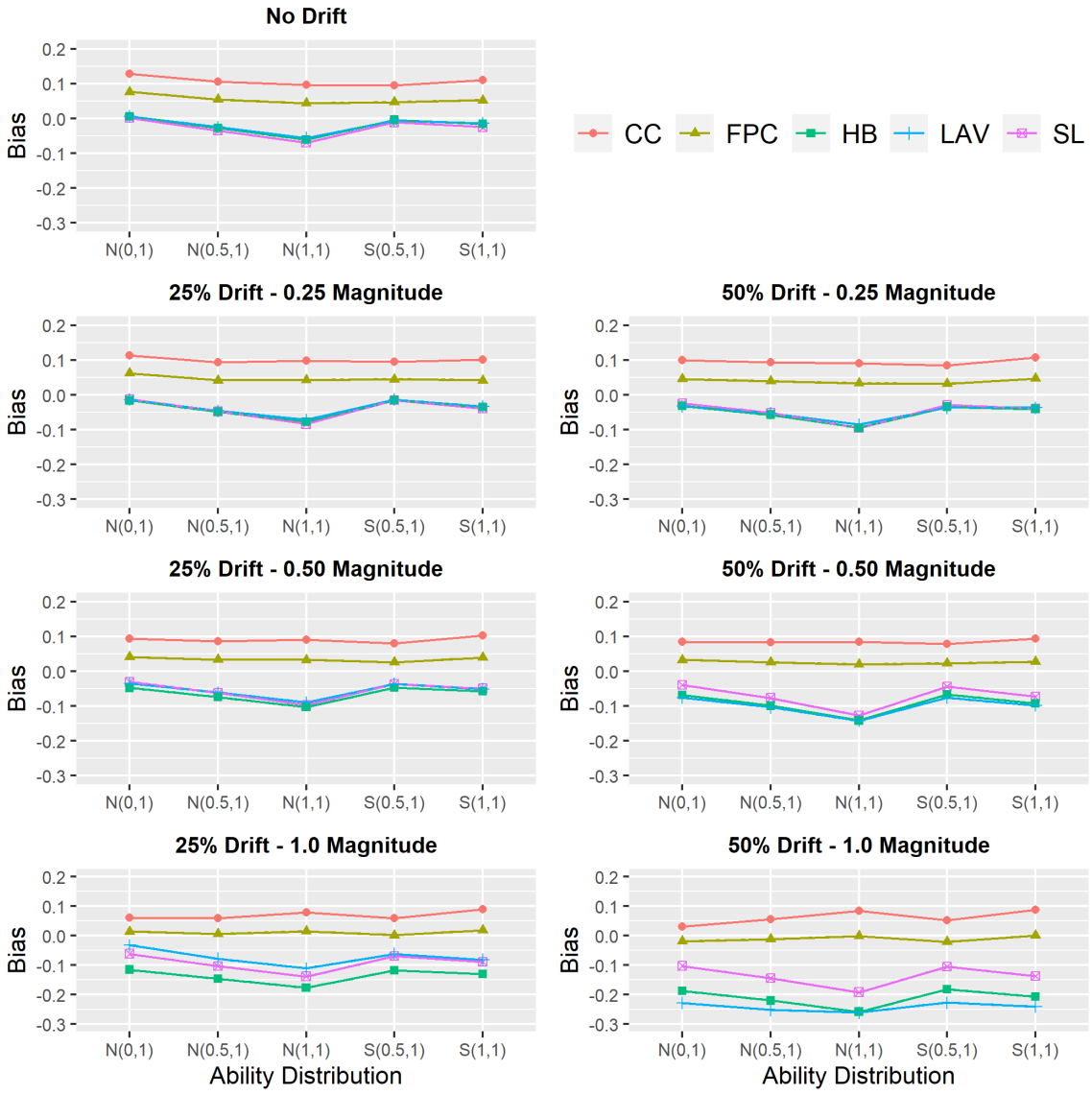

Figure 3. Bias Values for Linking Constant A – 1,000 Examinees.

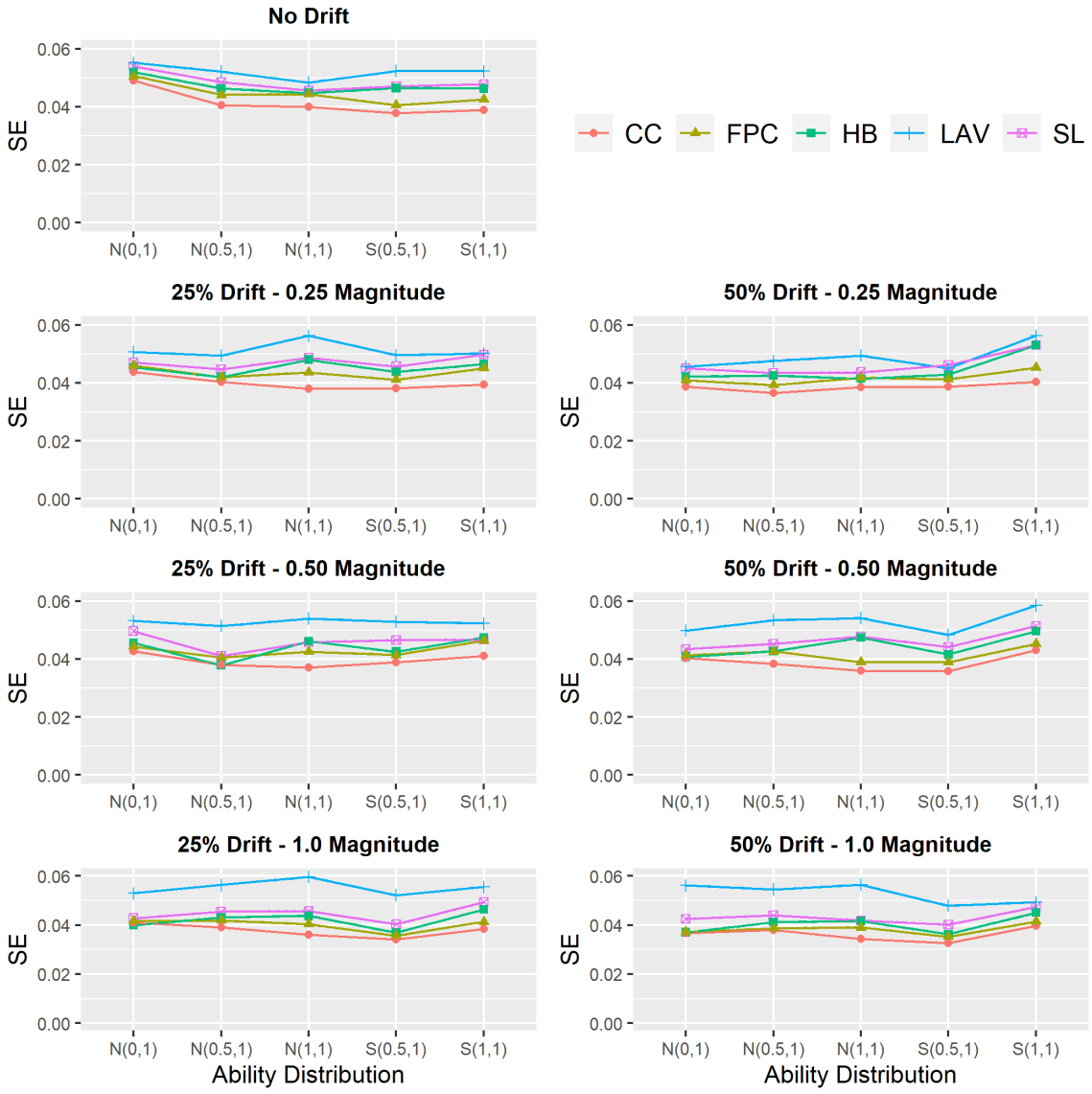

Figure 4. SE Values for Linking Constant A – 1,000 Examinees.

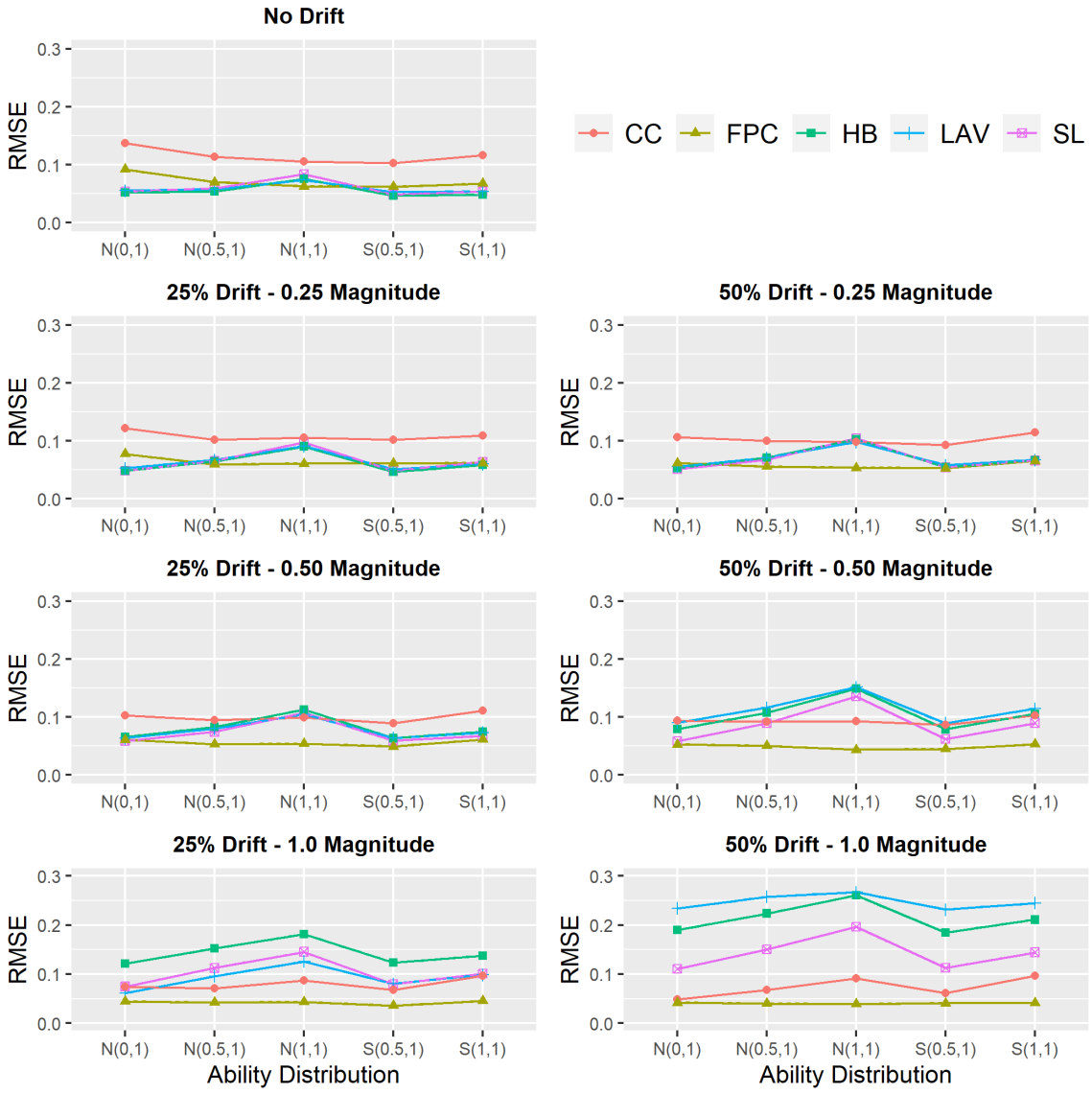

Figure 5. RMSE Values for Linking Constant A – 1,000 Examinees.

When the sample size increased to 3,000, lower values of bias, SE, and RMSE were observed for all methods and most of the conditions. When no drift was present, RMSE increased for the separate calibration methods as ability increased for the normal distributions but remained the same under the skewed distributions. For CC and FPC, RMSE decreased as the mean of the normal ability distributions increased. As the mean of the skewed distributions increased, RMSE increased for CC, but remained stagnant for FPC. The separate calibration methods recovered *A* better than CC and FPC for N(0,1) and the skewed distributions, but CC and FPC recovered *A* better for N(1,1).

For the 25% drifted item conditions, RMSE increased for the separate calibration methods but remained unchanged for FPC and CC as the mean for the normal ability distributions moved further away from 0. This could be attributed to increases in both bias and SE. Values of RMSE hardly changed under the skewed distributions. As drift magnitude increased, RMSE increased for the separate calibration methods, but no pattern was evident for CC and FPC. FPC recovered *A* best at moderate magnitudes of drift (-0.25 and -0.50) and CC recovered *A* the best when the drift magnitude was -1.00.

When drift was manipulated for 50% of the common items, the separate calibration methods were less accurate in recovering *A* as ability increased. No pattern was evident for CC and FPC as the ability distributions changed. As drift magnitude increased, the accuracy of the separate calibration methods and FPC decreased, while CC remained unchanged. CC produced the smallest values of RMSE, followed by FPC and SL. HB and LAV were most impacted by the highest magnitudes of drift. Similar to the 1,000 sample-size, the LAV exhibited the largest SE for nearly all conditions of drift.

Table 7

Estimated Linking Constant A – 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.997	0.974	0.941	0.998	0.997
	25%	-0.25	0.989	0.964	0.926	0.989	0.988
		-0.50	0.974	0.950	0.918	0.981	0.980
		-1.00	0.939	0.909	0.872	0.952	0.950
	50%	-0.25	0.984	0.956	0.921	0.987	0.986
		-0.50	0.961	0.932	0.893	0.968	0.970
		-1.00	0.903	0.872	0.831	0.924	0.912
HB	None	None	1.000	0.981	0.951	1.002	1.004
	25%	-0.25	0.985	0.963	0.930	0.985	0.988
		-0.50	0.956	0.935	0.909	0.965	0.967
		-1.00	0.885	0.860	0.833	0.899	0.903
	50%	-0.25	0.975	0.950	0.922	0.980	0.982
		-0.50	0.932	0.908	0.876	0.941	0.945
		-1.00	0.819	0.795	0.761	0.843	0.837
LAV	None	None	1.000	0.980	0.956	0.996	0.999
	25%	-0.25	0.990	0.967	0.935	0.982	0.984
		-0.50	0.980	0.960	0.927	0.973	0.976
		-1.00	0.977	0.948	0.915	0.958	0.967
	50%	-0.25	0.971	0.948	0.924	0.975	0.978
		-0.50	0.906	0.885	0.859	0.908	0.914
		-1.00	0.755	0.752	0.729	0.783	0.780
CC	None	None	1.052	1.040	1.033	1.036	1.049
	25%	-0.25	1.042	1.032	1.026	1.027	1.045
		-0.50	1.024	1.020	1.024	1.019	1.043
		-1.00	0.983	0.985	1.003	0.996	1.031
	50%	-0.25	1.036	1.026	1.026	1.026	1.047
		-0.50	1.011	1.011	1.018	1.010	1.042
		-1.00	0.954	0.980	1.006	0.988	1.026
FPC	None	None	1.024	1.012	1.001	1.005	1.008
	25%	-0.25	1.014	1.003	0.993	0.993	1.001
		-0.50	0.998	0.992	0.987	0.984	0.997
		-1.00	0.964	0.958	0.962	0.957	0.978
	50%	-0.25	1.008	0.996	0.990	0.991	1.002
		-0.50	0.984	0.979	0.976	0.971	0.992
		-1.00	0.932	0.941	0.944	0.936	0.955

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

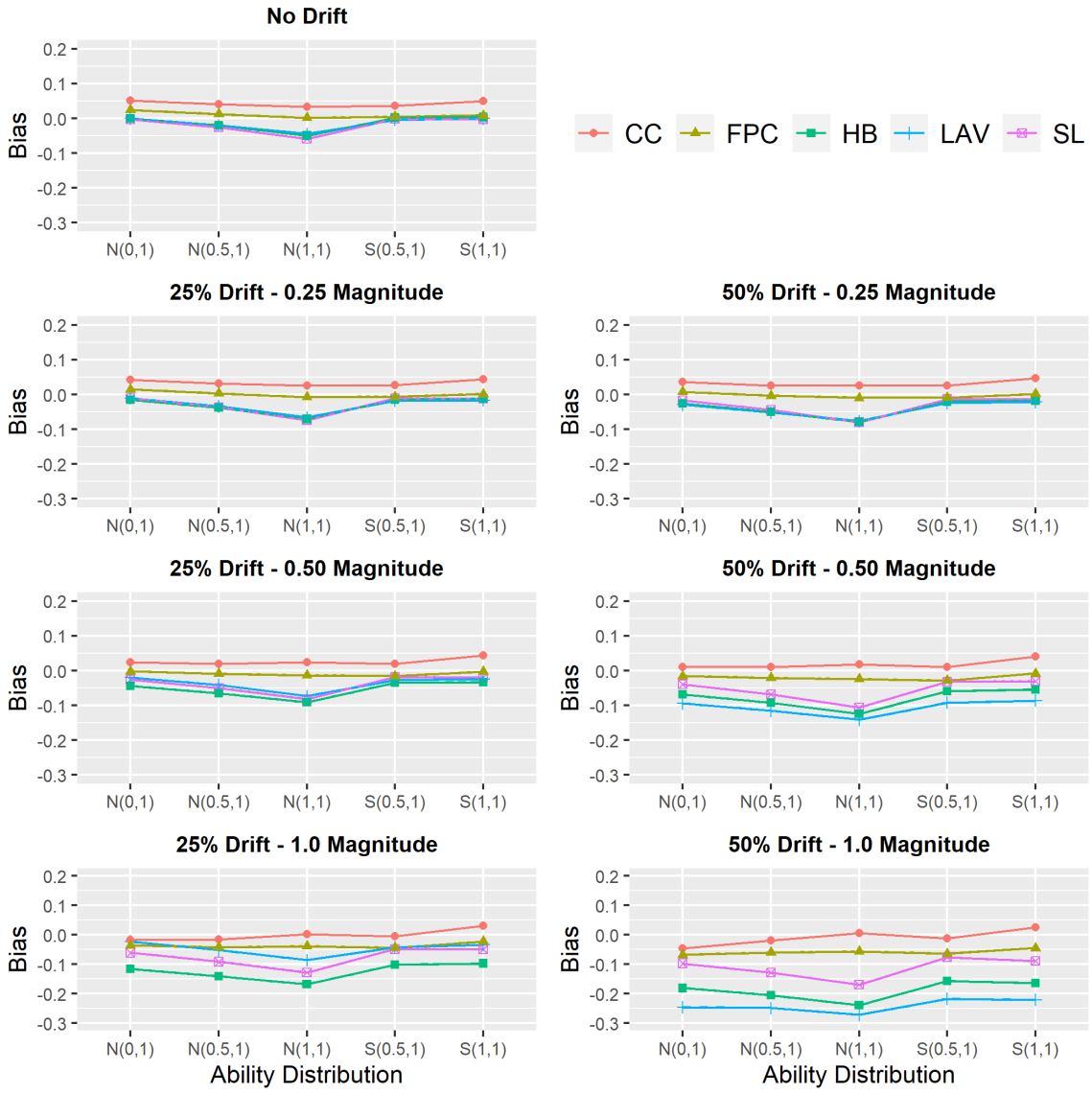

Figure 6. Bias Values for Linking Constant A – 3,000 Examinees.

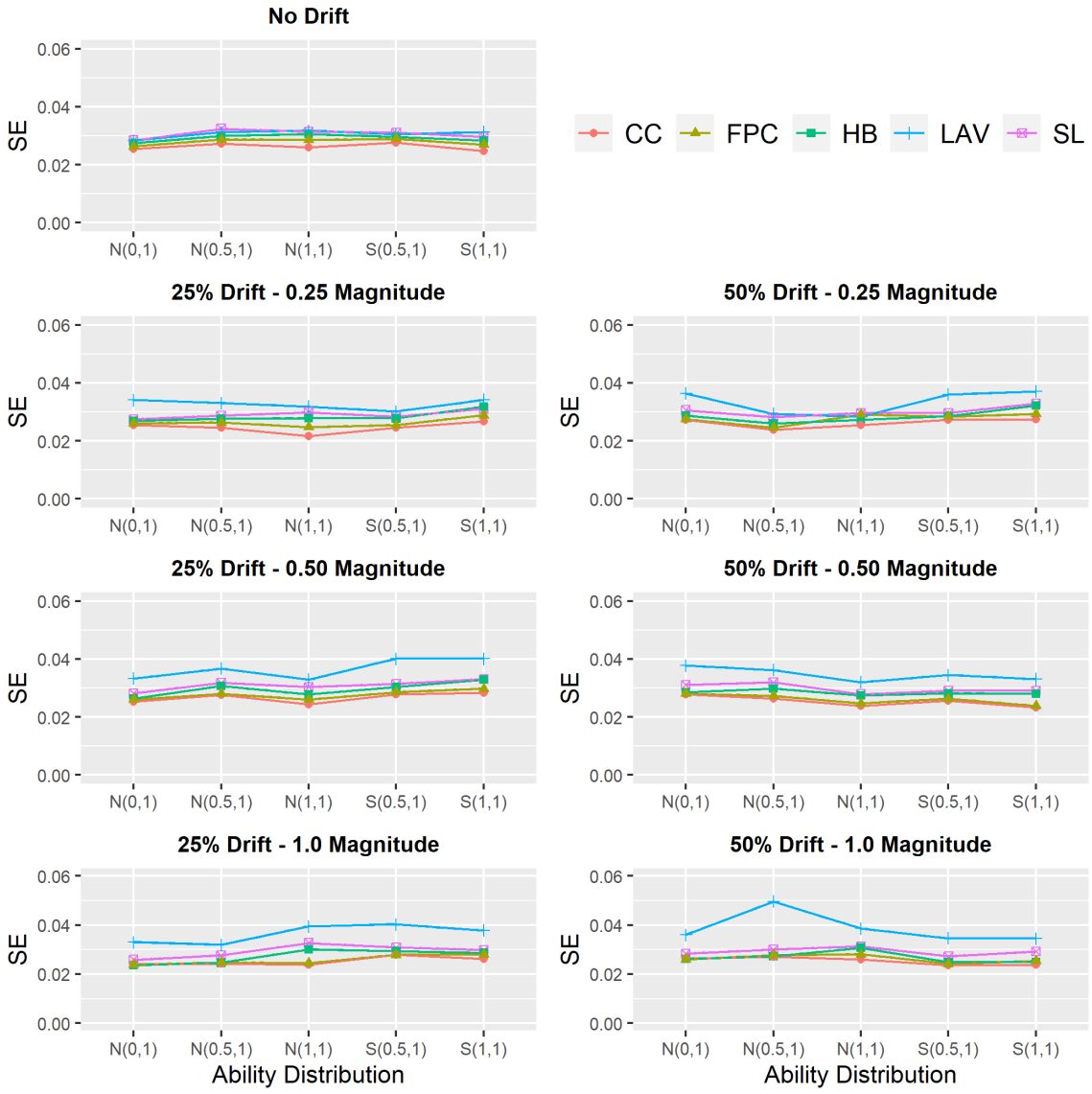

Figure 7. SE Values for Linking Constant A – 3,000 Examinees.

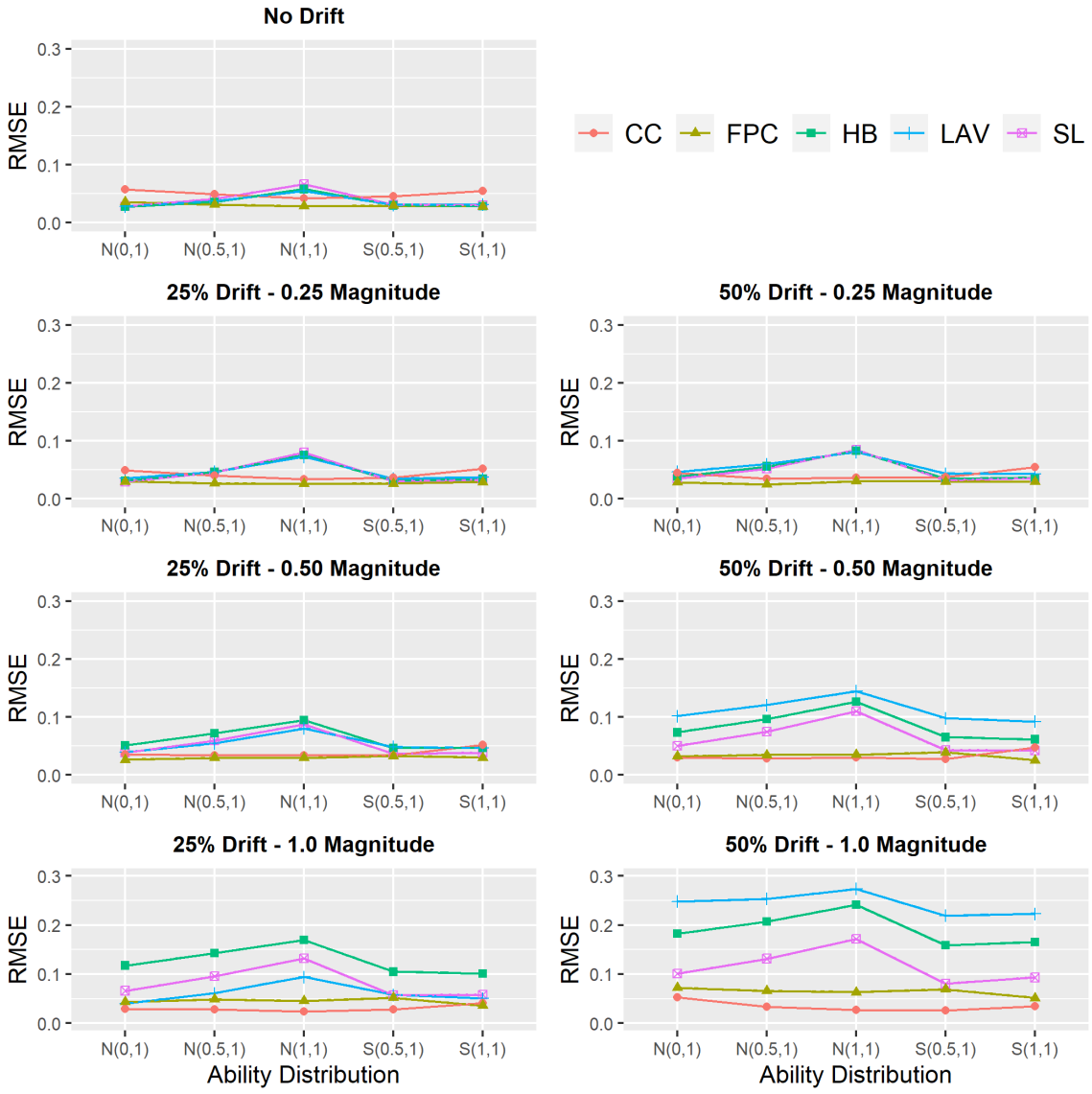

Figure 8. RMSE Values for Linking Constant A – 3,000 Examinees.

Linking Constant B. The expected value of *B* is dependent upon the ability distribution the examinees originated from. For N(0,1), the expected value of *B* should be 0 while the expected value should be 0.5 when the population ability is N(0.5,1). Estimates for linking constant *B* are found in Tables 8 and 9 for the 1,000 and 3,000 sample-size conditions, respectively. Bias, SE and RMSE for the 1,000 sample-size condition are illustrated in Figures 9 – 11. Bias, SE, and RMSE for the 3,000 sample-size condition are illustrated in Figures 12 – 14. Actual values for bias, SE, and RMSE can be found in Appendix B.

For the 1,000 sample-size conditions, all linking methods tended to overestimate *B* for most of the conditions (this was also true for the 3,000 sample-size condition). Since drift was introduced to make items easier on the new forms, the overestimation of linking constant *B* implies that the items were easier for the new form examinees, which was expected. When no drift was present, all linking methods had similar values of RMSE. For all linking methods, RMSE tended to increase as the ability distributions moved further away from a mean of 0. Similarly, the more drift is introduced, the more RMSE increased. However, the combined effect of increasing drift (proportion and magnitude) and increasing ability led to smaller RMSE. This might occur because there are fewer examinees that will benefit from drift when the ability distribution is already high (i.e., most examinees are already likely to answer an item correctly). Jurich et al. (2012) noted that this ceiling effect arises when higher ability examinees cannot benefit from the effects of drift. This pattern was also evident when the sample size increased to 3,000.

When 25% of common items were drifted, there was no systematic pattern of RMSE on the recovery of *B* for the separate calibration methods and for FPC. However, bias systematically decreased for the separate calibration methods as group differences increased. RMSE and bias increased for CC as ability distributions deviated greater from a mean of 0. As the magnitude of drift increased, RMSE and bias also increased for all the linking methods under almost all conditions. The LAV tended to recover *B* better than the other linking methods. Values of RMSE and bias were much larger for CC and FPC than the separate calibration methods. Hu et al. (2008) found that group equivalence was the most important factor for CC and FPC in the recovery of difficulty parameters and equated true scores. CC and FPC produced RMSE values similar to SL and HB under N(0,1), but produced RMSE greater than SL and HB when groups were not equivalent.

When 50% of common items were drifted, the recovery of *B* improved for all linking methods except for CC as the ability distributions moved further away from 0. As drift magnitude increased, the recovery of *B* became less accurate for all linking methods. The LAV performed better than the other linking methods for almost all conditions despite having slightly higher SE values. Since drift occurred exclusively in the difficulty parameter, the LAV may have recovered *B* the best because its weight function was able to minimize the effect of the drifted items by assigning them smaller weights in the linking process.

Table 8

Estimated Linking Constant B – 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.006	0.468	0.911	0.482	0.926
	25%	-0.25	0.067	0.523	0.955	0.534	0.986
		-0.50	0.125	0.578	1.002	0.597	1.031
		-1.00	0.229	0.669	1.072	0.680	1.100
	50%	-0.25	0.126	0.572	1.001	0.586	1.031
		-0.50	0.234	0.673	1.091	0.692	1.125
		-1.00	0.442	0.848	1.231	0.873	1.262
HB	None	None	0.005	0.466	0.912	0.479	0.925
	25%	-0.25	0.064	0.517	0.949	0.526	0.976
		-0.50	0.117	0.560	0.981	0.579	1.006
		-1.00	0.196	0.614	0.997	0.621	1.024
	50%	-0.25	0.121	0.563	0.991	0.577	1.017
		-0.50	0.222	0.649	1.058	0.665	1.089
		-1.00	0.395	0.766	1.122	0.792	1.148
LAV	None	None	0.007	0.465	0.916	0.481	0.925
	25%	-0.25	0.051	0.507	0.946	0.520	0.971
		-0.50	0.066	0.518	0.947	0.537	0.971
		-1.00	0.055	0.509	0.929	0.517	0.946
	50%	-0.25	0.120	0.561	0.995	0.573	1.021
		-0.50	0.197	0.625	1.045	0.641	1.065
		-1.00	0.313	0.667	1.025	0.667	1.028
CC	None	None	-0.038	0.504	1.047	0.533	1.065
	25%	-0.25	0.023	0.562	1.099	0.587	1.127
		-0.50	0.085	0.625	1.158	0.658	1.188
		-1.00	0.200	0.746	1.276	0.769	1.303
	50%	-0.25	0.088	0.617	1.152	0.644	1.183
		-0.50	0.204	0.735	1.272	0.765	1.301
		-1.00	0.450	0.977	1.517	1.015	1.535
FPC	None	None	-0.018	0.496	1.010	0.513	1.014
	25%	-0.25	0.044	0.553	1.056	0.566	1.073
		-0.50	0.103	0.609	1.106	0.628	1.123
		-1.00	0.206	0.703	1.186	0.711	1.201
	50%	-0.25	0.108	0.605	1.107	0.622	1.127
		-0.50	0.220	0.712	1.206	0.730	1.226
		-1.00	0.430	0.897	1.374	0.916	1.383

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

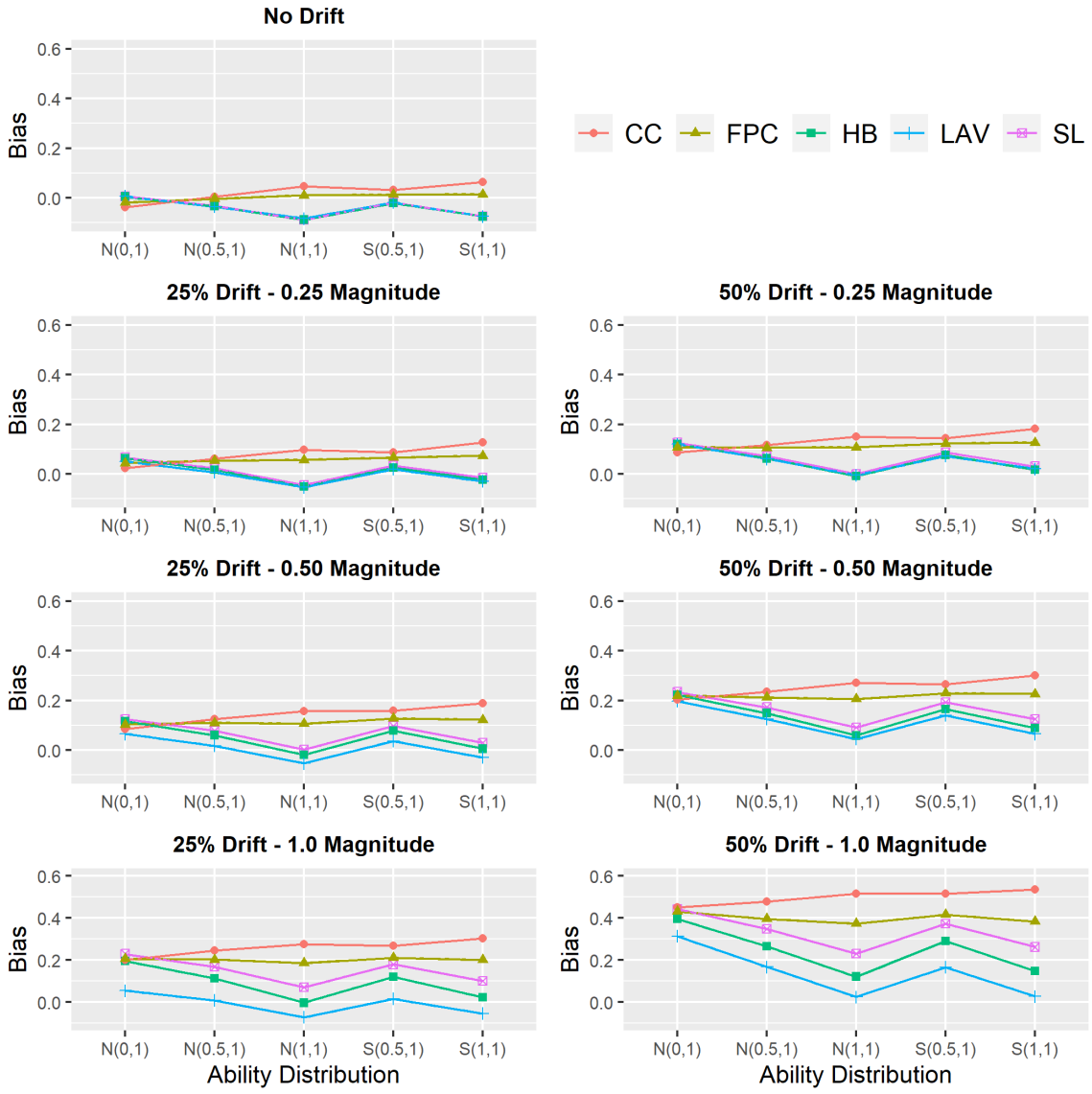

Figure 9. Bias Values for Linking Constant B – 1,000 Examinees.

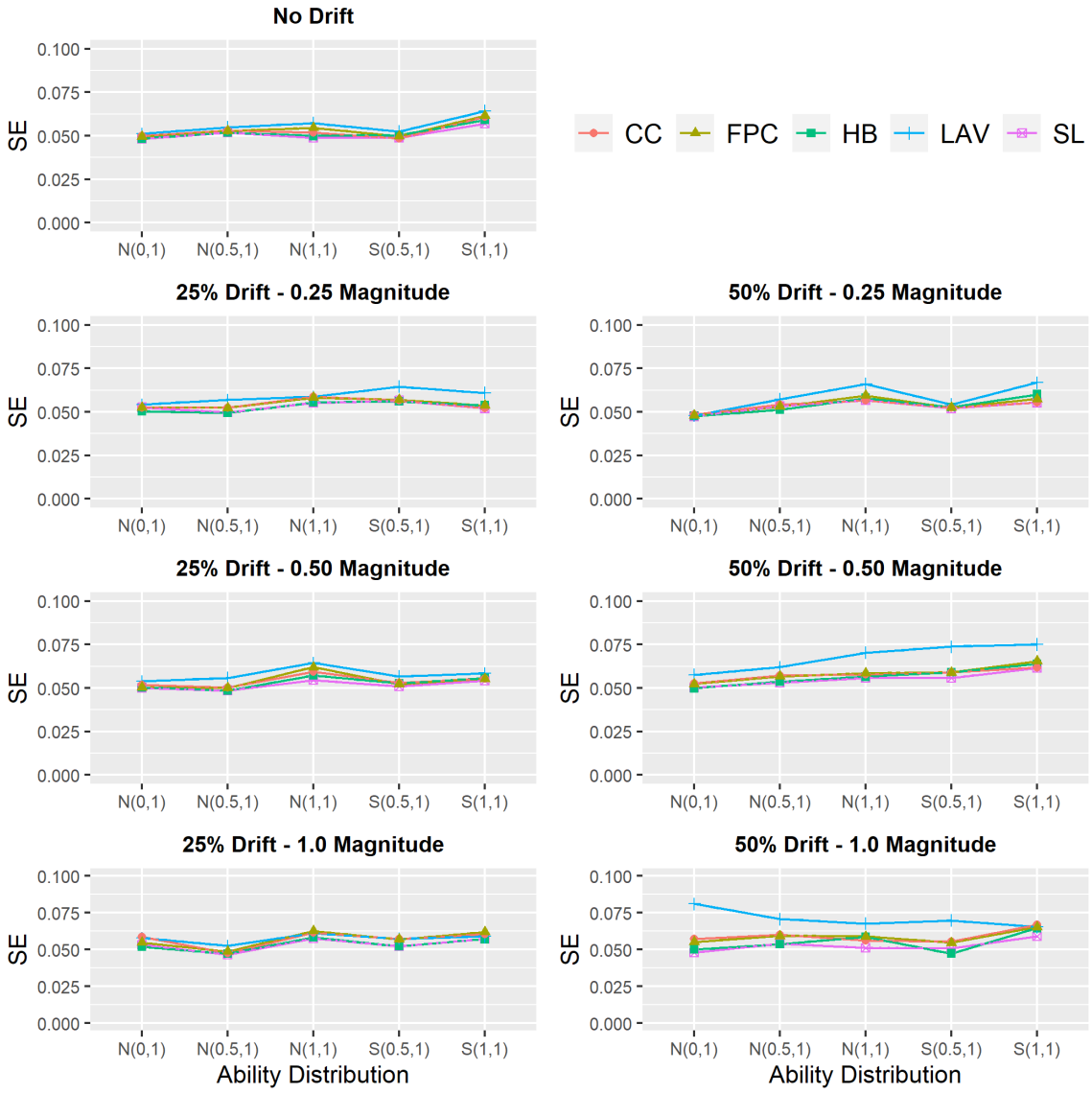

Figure 10. SE Values for Linking Constant B – 1,000 Examinees.

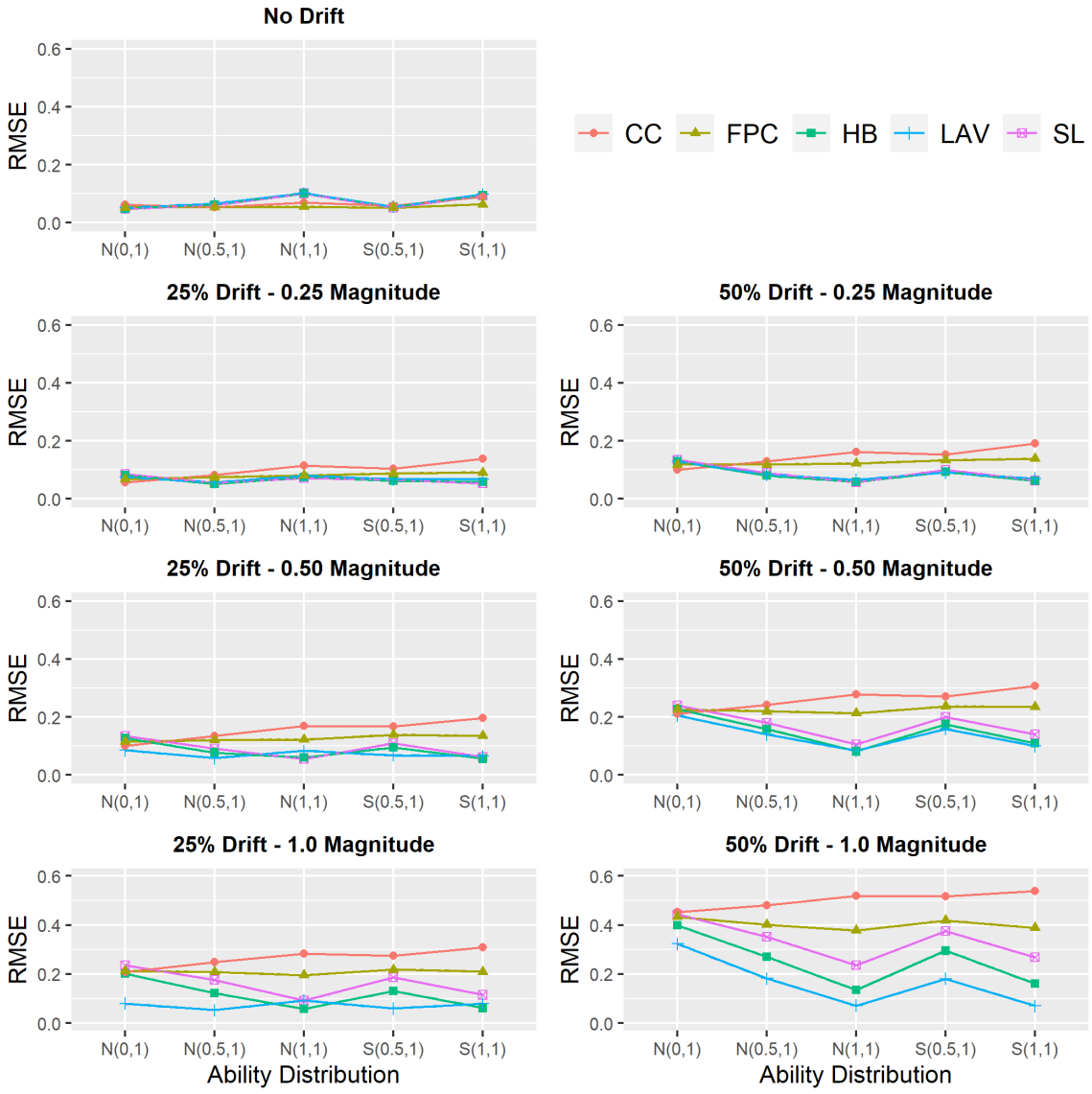

Figure 11. RMSE Values for Linking Constant B – 1,000 Examinees.

Under the 3,000 sample-size conditions, most of the values of *B* were also overestimated for all linking methods and conditions. CC and FPC had smaller RMSE values with 3,000 examinees than 1,000 examinees. SL and HB had smaller RMSE values when drift was null or small (25% drifted items), but larger RMSE values with the highest proportion and magnitude of drifted items. In most conditions, the LAV was more accurate in recovering *B* with a larger sample size.

When no drift was present, the RMSE for *B* increased as the ability distributions moved further away from N(0,1). Although all linking methods recovered *B* well, FPC performed the best, followed by CC.

Among the 25% and 50% of common items drifted, the separate calibration methods and FPC recovered *B* better for most conditions as drift and ability increased. The addition of drift does not greatly impact the probability of correctly responding to an item when that probability is already very high, thus, less error is produced. On the other hand, the RMSE for CC slightly increased as the ability distributions increased from N(0,1) due to an increase in bias. The LAV method had smaller RMSE values than the other linking methods for nearly all conditions despite having larger SE values.

Table 9

Estimated Linking Constant B – 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.005	0.481	0.941	0.504	0.988
	25%	-0.25	0.061	0.539	0.993	0.563	1.039
		-0.50	0.120	0.585	1.041	0.613	1.083
		-1.00	0.223	0.676	1.105	0.707	1.154
	50%	-0.25	0.115	0.585	1.039	0.617	1.091
		-0.50	0.234	0.683	1.126	0.724	1.178
		-1.00	0.434	0.861	1.265	0.901	1.332
HB	None	None	0.003	0.479	0.944	0.501	0.987
	25%	-0.25	0.057	0.533	0.988	0.555	1.029
		-0.50	0.111	0.568	1.020	0.593	1.055
		-1.00	0.192	0.619	1.031	0.646	1.070
	50%	-0.25	0.111	0.577	1.032	0.606	1.078
		-0.50	0.222	0.660	1.096	0.696	1.140
		-1.00	0.388	0.780	1.153	0.812	1.212
LAV	None	None	0.004	0.480	0.946	0.503	0.985
	25%	-0.25	0.035	0.516	0.978	0.536	1.012
		-0.50	0.037	0.509	0.974	0.532	1.005
		-1.00	0.030	0.504	0.960	0.528	0.986
	50%	-0.25	0.108	0.567	1.029	0.594	1.068
		-0.50	0.177	0.606	1.042	0.622	1.067
		-1.00	0.309	0.668	1.027	0.644	1.039
CC	None	None	-0.013	0.499	1.011	0.523	1.042
	25%	-0.25	0.045	0.562	1.072	0.586	1.100
		-0.50	0.108	0.614	1.133	0.643	1.156
		-1.00	0.224	0.733	1.246	0.765	1.269
	50%	-0.25	0.105	0.613	1.127	0.644	1.159
		-0.50	0.234	0.730	1.245	0.766	1.271
		-1.00	0.472	0.976	1.485	1.009	1.517
FPC	None	None	-0.004	0.495	0.990	0.510	1.009
	25%	-0.25	0.053	0.555	1.047	0.570	1.062
		-0.50	0.114	0.602	1.096	0.619	1.107
		-1.00	0.216	0.695	1.171	0.711	1.181
	50%	-0.25	0.112	0.605	1.098	0.627	1.118
		-0.50	0.234	0.709	1.197	0.734	1.211
		-1.00	0.437	0.898	1.357	0.913	1.378

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

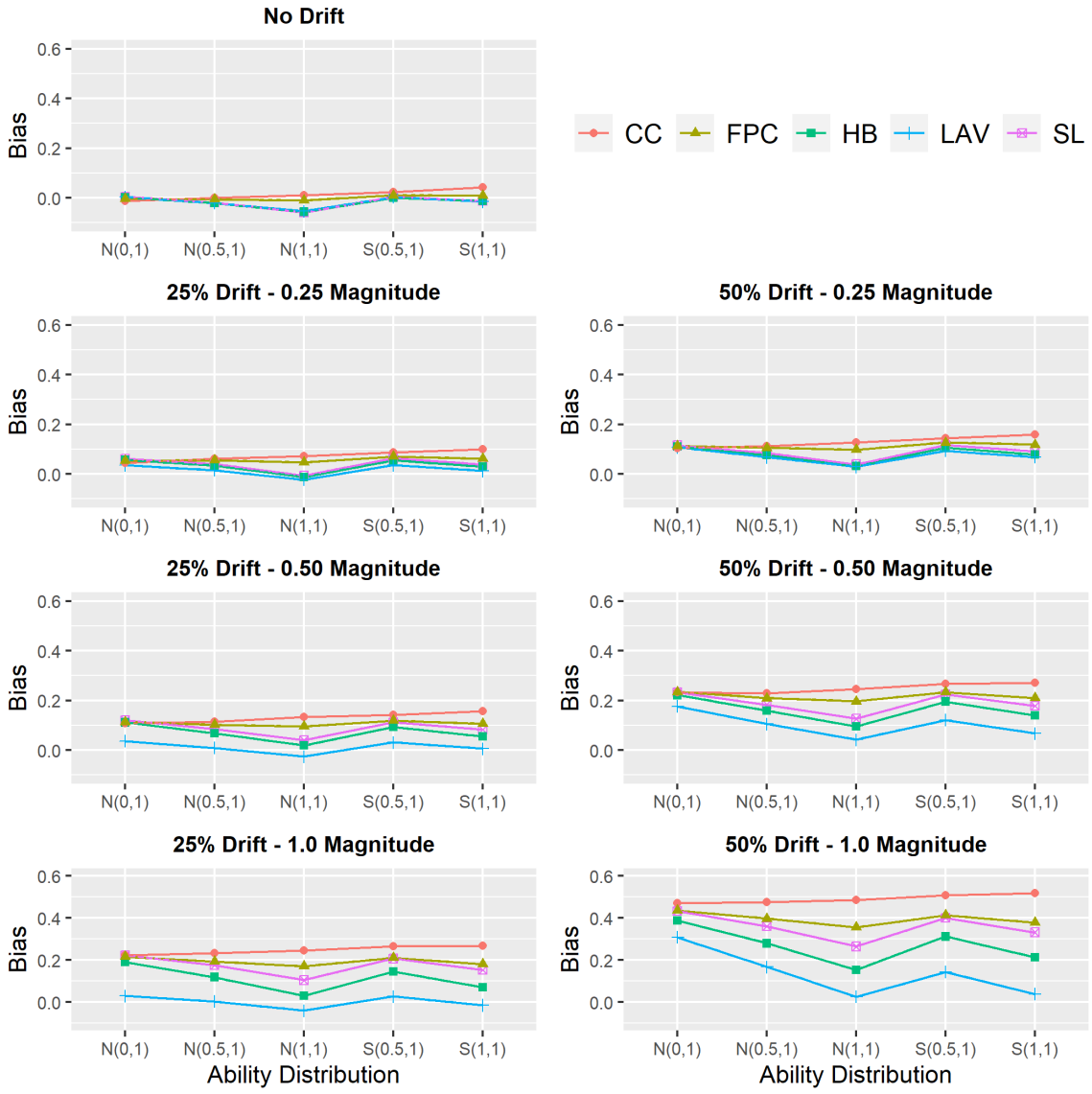

Figure 12. Bias Values for Linking Constant B – 3,000 Examinees.

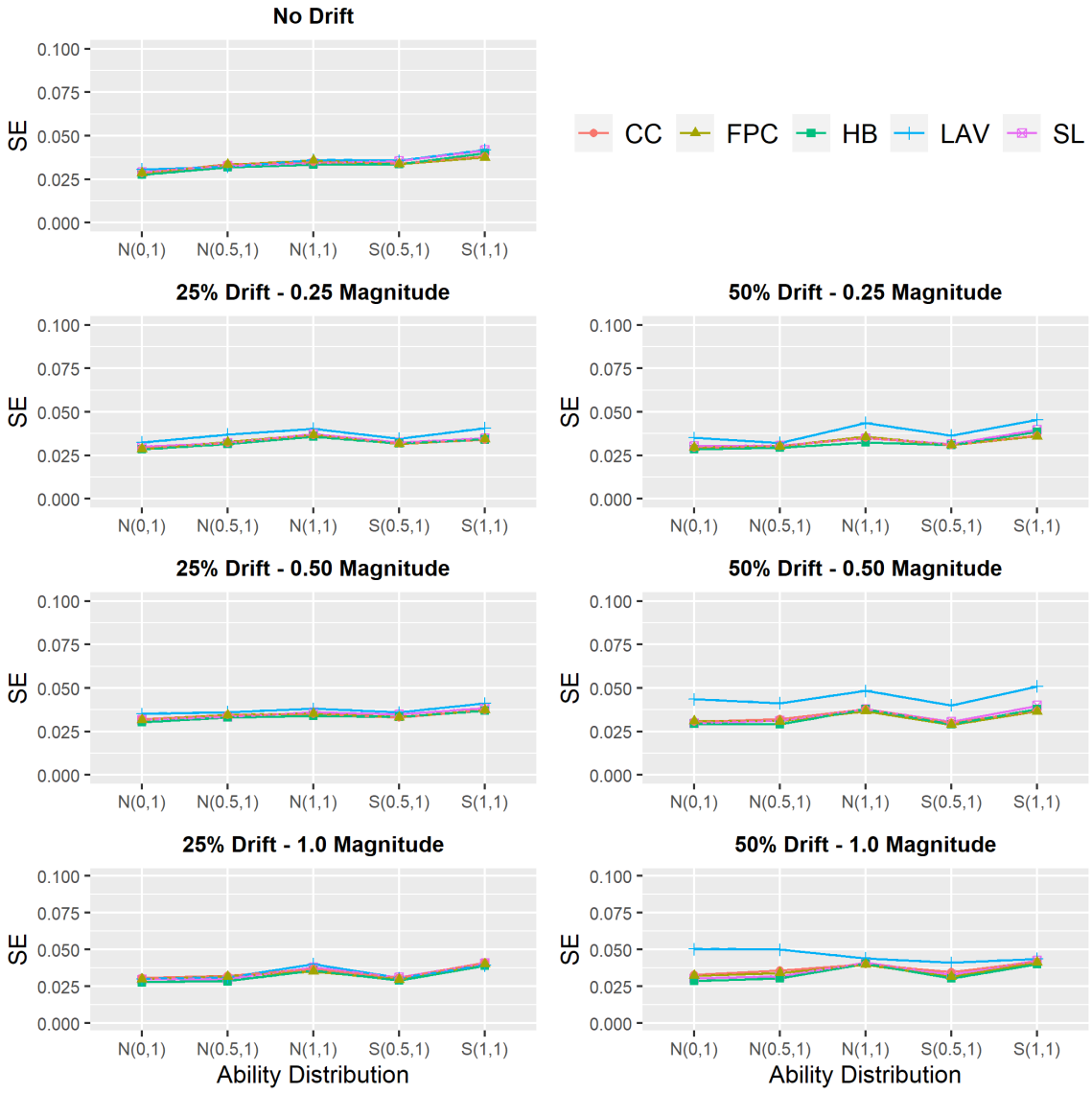

Figure 13. SE Values for Linking Constant B – 3,000 Examinees.

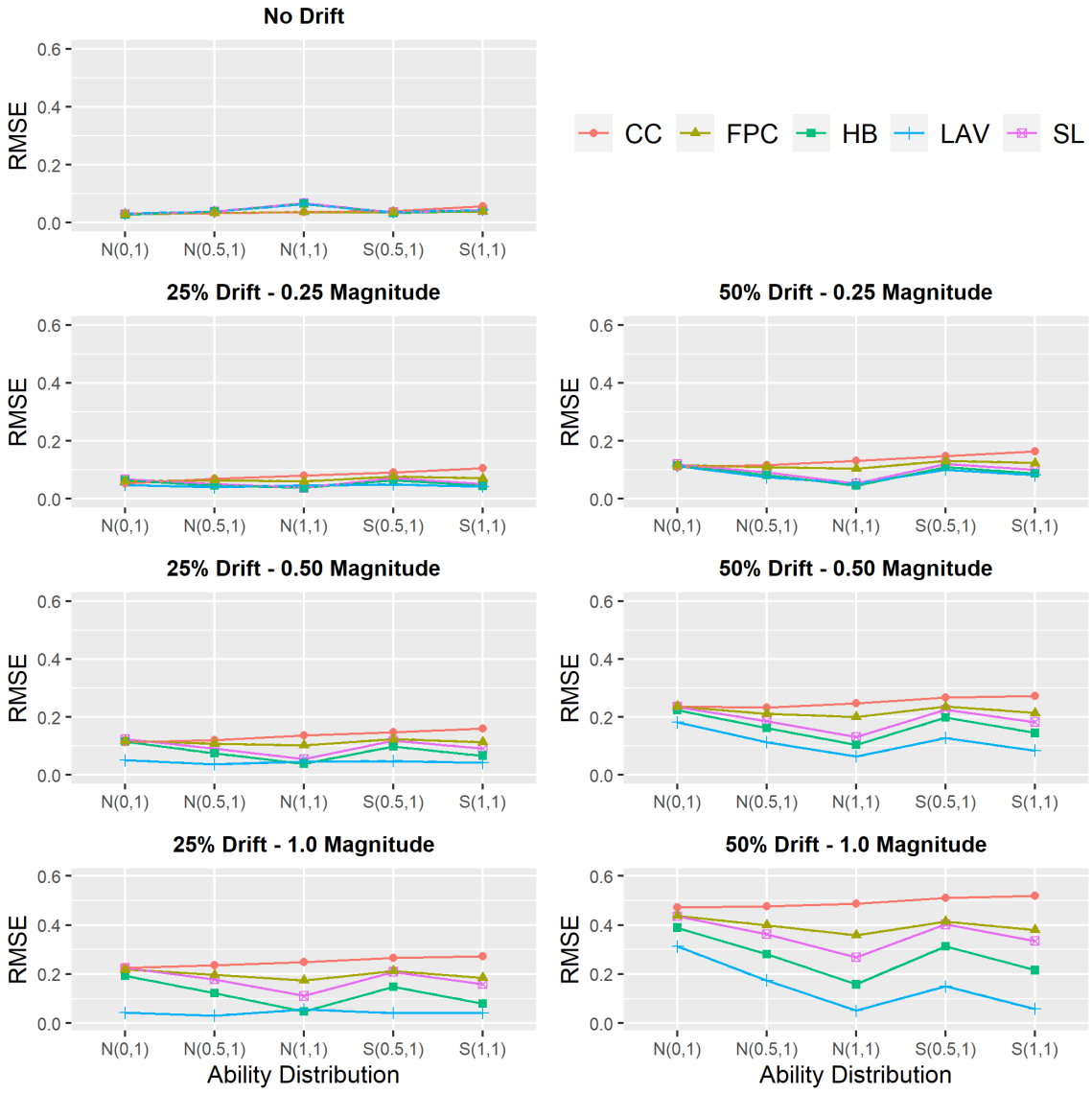

Figure 14. RMSE Values for Linking Constant B – 3,000 Examinees.

Linked Item Parameter Estimates. The second research question examined the impact of IPD on the recovery of the linked item parameter estimates: discrimination (*a*), difficulty (*b*), and pseudo-guessing (*c*).

Linked Item Parameter Estimate a. Bias, SE, and RMSE values were calculated by comparing the linked item parameter estimates for the new form unique items to the generating item parameters. Bias, SE, and RMSE is illustrated in Figures 15 – 17 for the 1,000 sample-size and in Figures 18 – 20 for the 3,000 sample-size. Values of bias, SE, and RMSE for the 80 unique items can be found in Appendix C. Bias, SE, and RMSE was also evaluated for all 100 items and can be found in Appendix D.

Under the 1,000 sample-size conditions, when no drift was present, the RMSE for *a* increased for the separate calibration methods under the normal distributions as the mean deviated further away from 0. However, RMSE slightly decreased for the separate calibration methods under the skewed distributions as the mean deviated further away from 0. The RMSE for CC and FPC decreased for both normal and skewed ability distributions as the mean ability increased, which was consistent with the RMSE values from linking constant *A*. CC and FPC recovered *a* best for all ability distributions.

For the 25% and 50% drift conditions, RMSE tended to increase for the separate calibration methods as the mean of the ability distributions increased. For CC and FPC, RMSE typically decreased as the mean of the ability distributions increased. As the magnitude of drift increased, RMSE tended to increase for the separate calibration methods, but remained relatively unchanged for CC and FPC. CC and FPC produced the lowest levels of RMSE among all of the linking methods. The LAV was influenced the

greatest under the most extreme conditions of drift (50% drifted items, -1.0 magnitude). These findings are not surprising since the LAV yielded the greatest RMSE values for linking constant *A* at the most extreme conditions of drift, which has direct influence on item estimate *a* (equation 2.5).

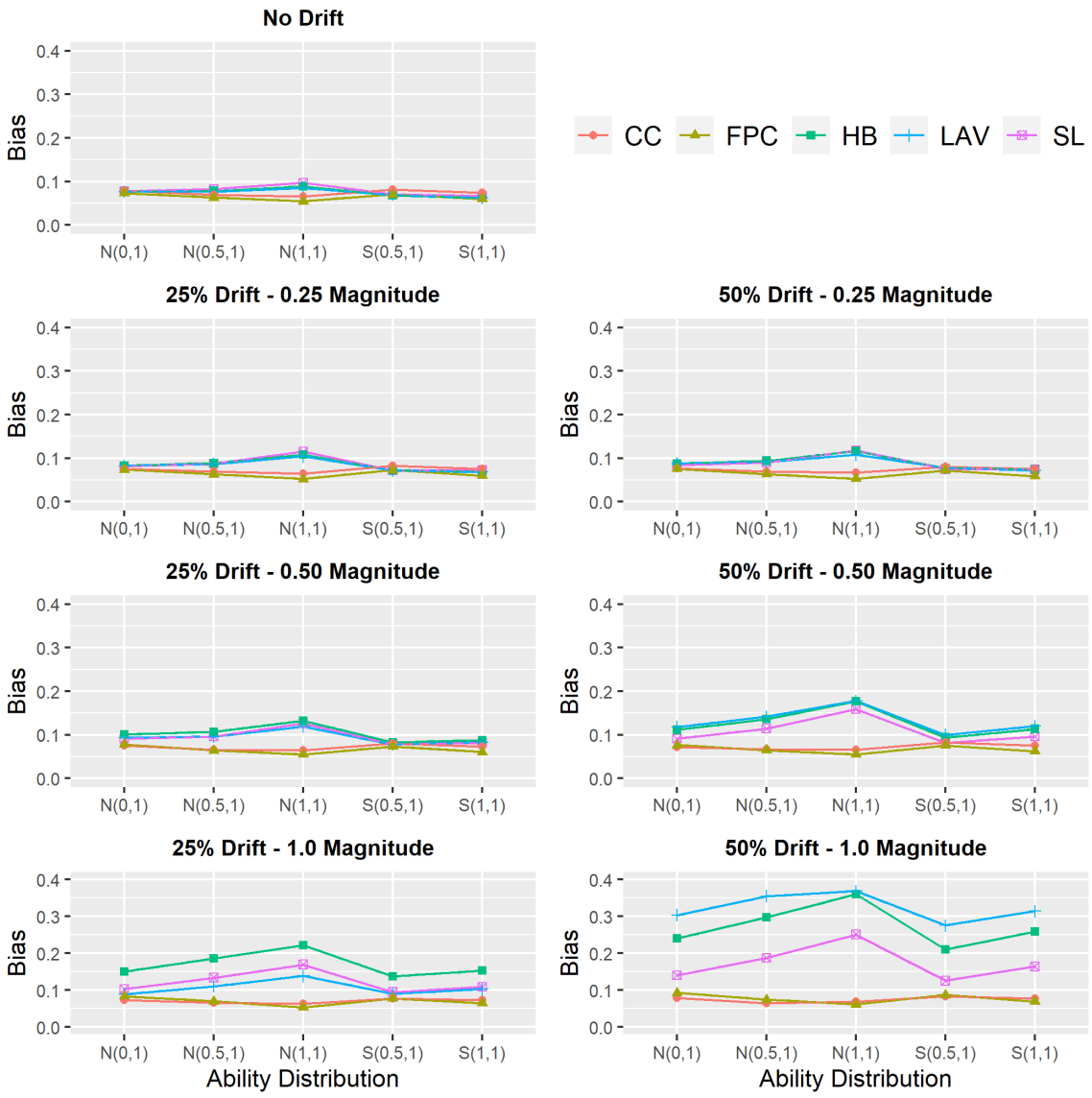

Figure 15. Bias Values for Item Estimate a – 1,000 Examinees.

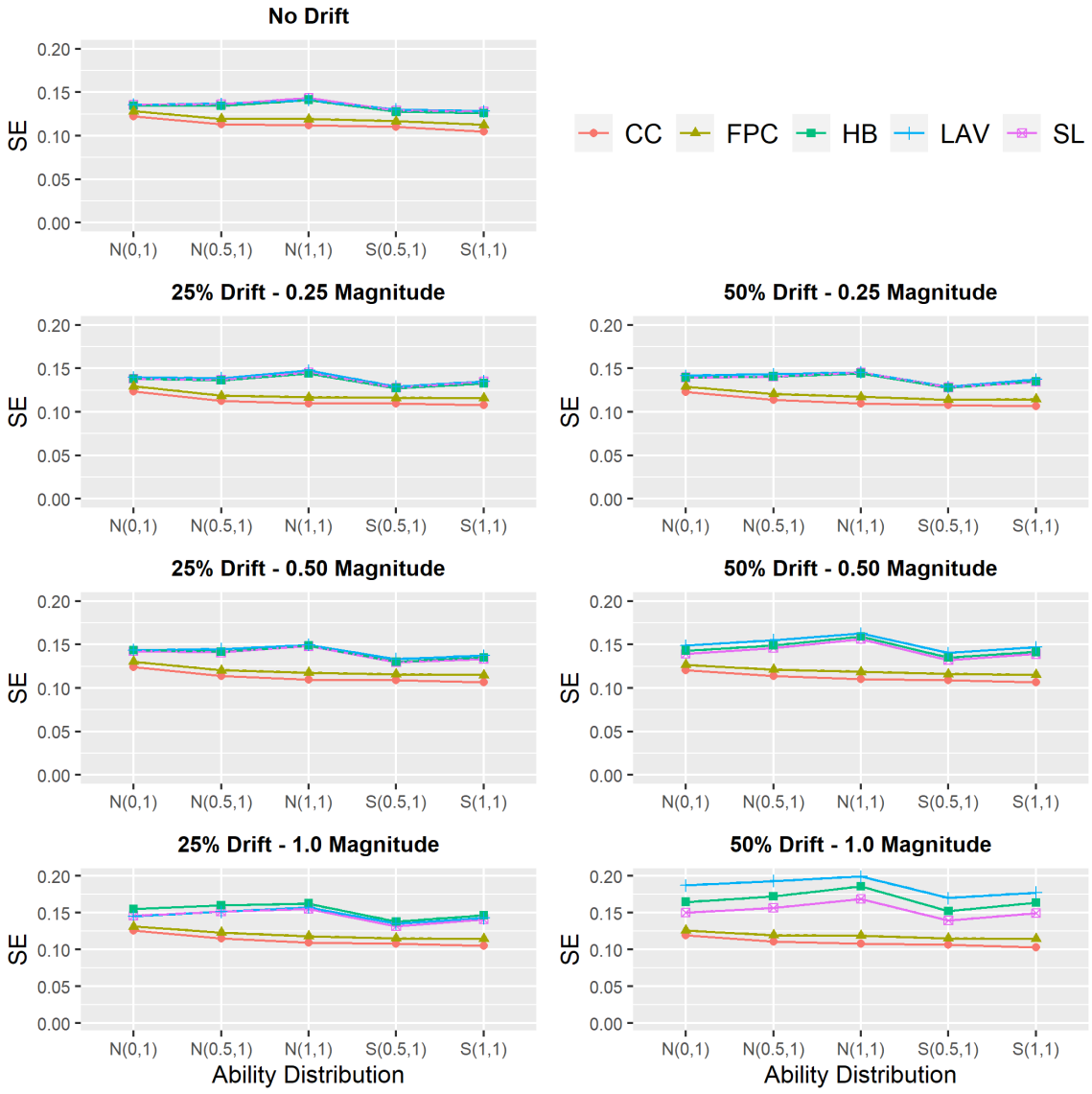

Figure 16. SE Values for Item Estimate a – 1,000 Examinees.

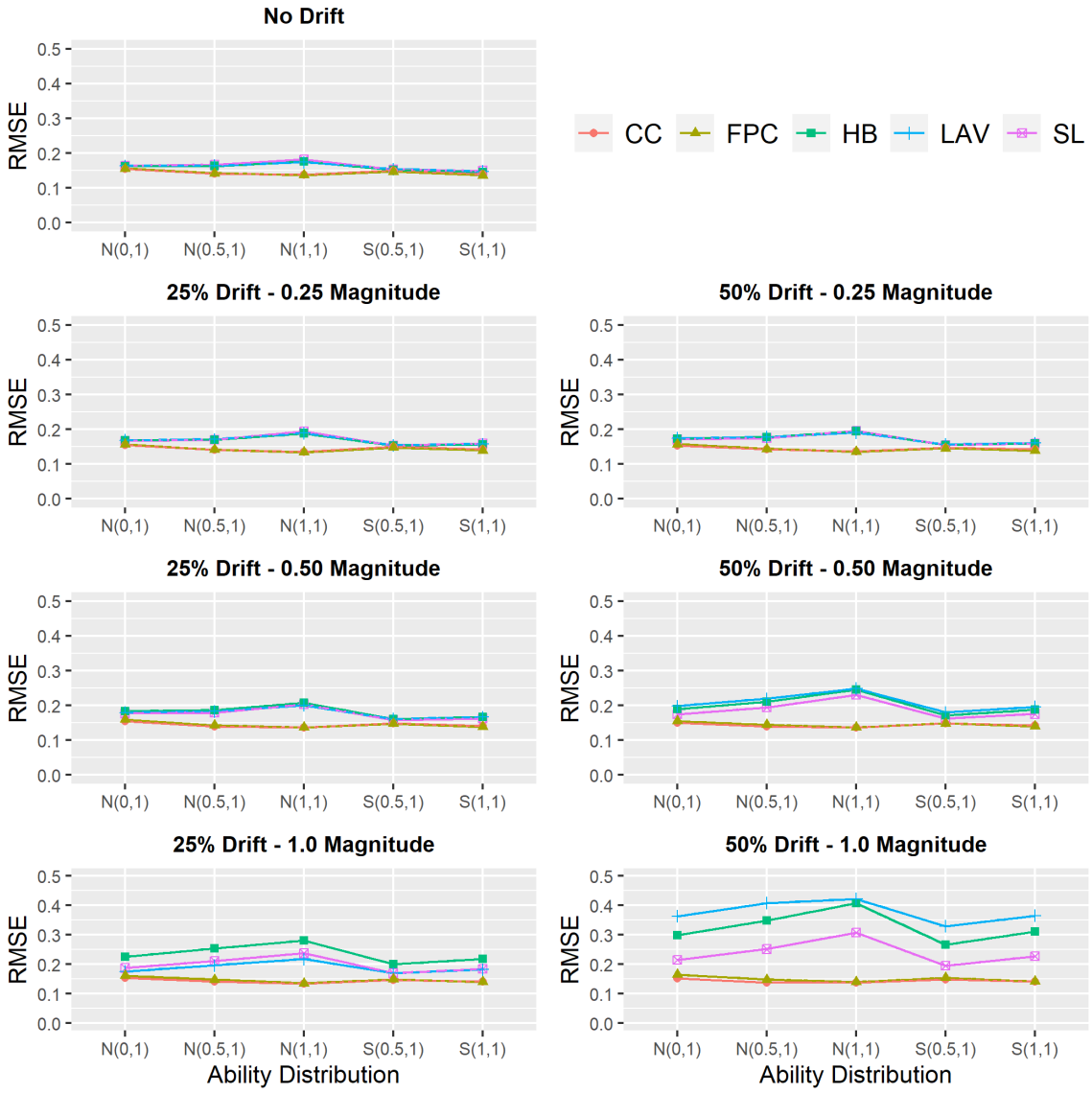

Figure 17. RMSE Values for Item Estimate a – 1,000 Examinees.

Compared to the 1,000 sample-size conditions, all linking methods provided smaller values of bias, SE, and RMSE than the 3,000 sample-size conditions. SE decreased the most, which is to be expected, since larger sample sizes yield smaller standard errors. When no drift was present, the RMSE for the separate calibration methods increased as the mean ability increased for the normal distributions; however, RMSE remained unchanged for the separate calibration methods as the mean ability increased for the skewed ability distributions. The RMSE for CC and FPC slightly decreased as the mean ability increased for both normal and skewed distributions. CC and FPC produced smaller RMSE values for the normal ability distributions, but all linking methods had similar RMSE values for the skewed distributions.

For the 25% drift conditions, the RMSE for *a* increased for the separate calibration methods as the mean of the normal ability distributions increased, but remained unchanged for the skewed ability distributions. CC and FPC yielded slightly smaller values of RMSE as the mean ability increased for normal distributions. As the magnitude of drift increased, all linking methods produced greater values of RMSE. Overall, CC produced the smallest values of RMSE, followed closely by FPC. The LAV produced the smallest values of RMSE among the separate calibration methods.

For the 50% drift conditions, RMSE typically increased for the separate calibration methods as the mean ability increased for both normal and skewed distributions. However, RMSE decreased for CC and FPC as the mean ability increased for both normal and skewed distributions. As drift magnitude increased, all linking methods produced larger values of RMSE. CC produced the smallest values of RMSE,

followed by FPC. SL yielded the smallest RMSE values among the separate calibration methods. On the other hand, the LAV method produced the largest RMSE, which is directly attributable to the large RMSE values the LAV exhibited for linking constant *A*.

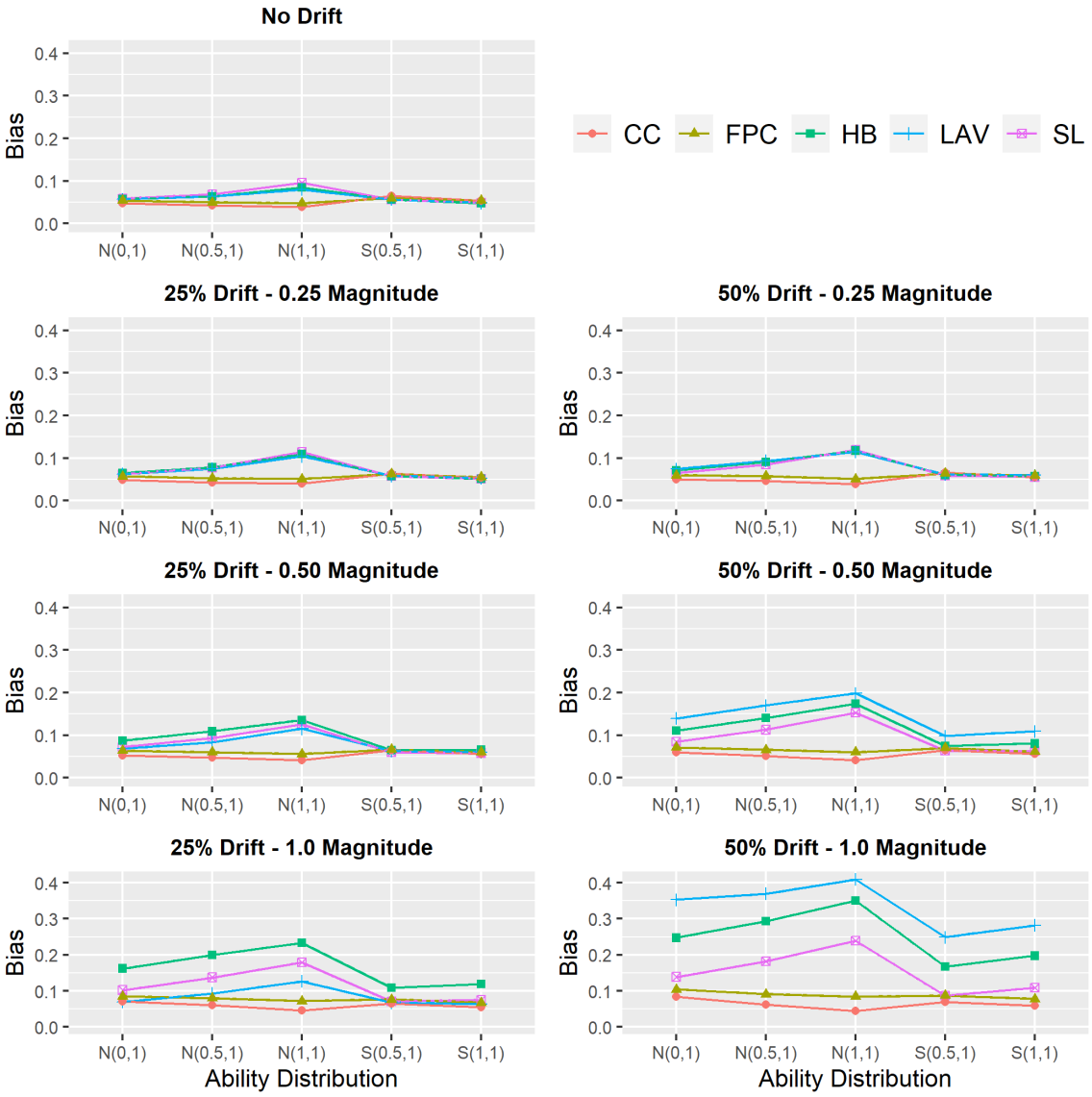

Figure 18. Bias Values for Item Estimate a – 3,000 Examinees.

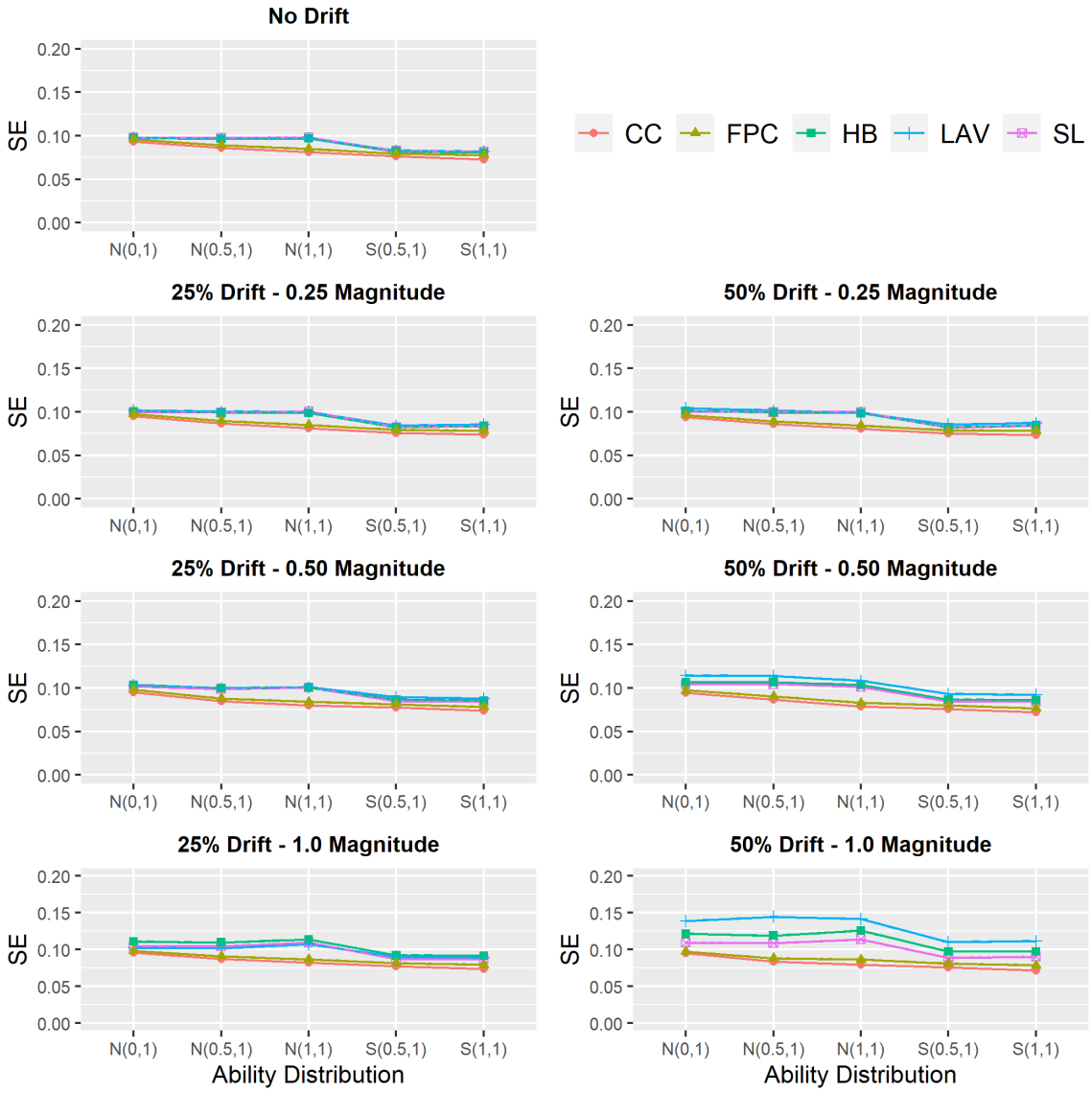

Figure 19. SE Values for Item Estimate a – 3,000 Examinees.

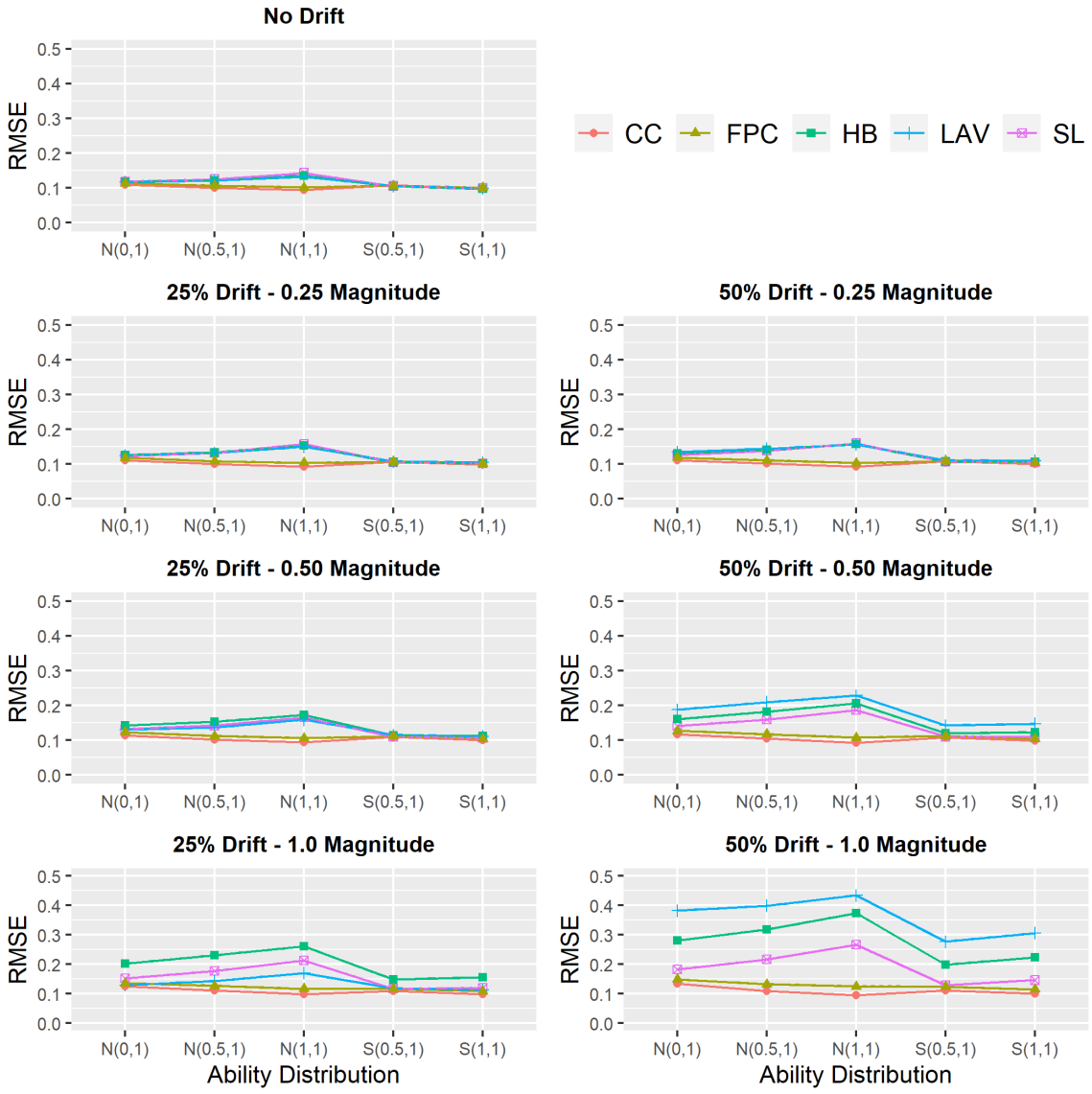

Figure 20. RMSE Values for Item Estimate a – 3,000 Examinees.

Linked Item Parameter Estimate b. Figures 21 – 23 illustrate the bias, SE, and RMSE values for item difficulty under the 1,000 sample-size conditions. Figures 24 – 26 illustrate the bias, SE, and RMSE values for item difficulty under the 3,000 sample-size conditions. Specific values for each of these three outcomes can be found in Appendix C. These outcomes were also evaluated for all items, common and unique, as found in Appendix D.

Overall, the findings here mimic those found for linking constant *B*. That is, the LAV most accurately recovered item estimate *b* for the conditions that it most accurately recovered linking constant *B*. This is because both linking constants effect the recovery of item estimate *b*. Linking constant *A* is multiplied to the linked estimate and then linking constant *B* is added to the linked estimate (as in equation 2.6), so better recovered linking constants will result in better recovered item parameter estimates. Although the LAV did not recover *A* as well as the other methods, LAV RMSE values for linking constant *B* were substantially smaller, often two to three times smaller for all ability distributions except N(0,1), than all other linking methods. The FPC also recovered item estimate *b* well, particularly at N(1,1). This may have been due to the direct effect that linking constant *A* has on the linked difficulty parameter (equation 2.6).

Under both sample sizes, when no drift was present, RMSE for *b* increased as the mean ability of the normal and skewed distributions increased. The performance of each linking method was fairly similar for most conditions, although CC and FPC returned smaller values of RMSE for N(0.5, 1) and N(1,1).

When 25% of the common items were drifted, RMSE for all linking methods typically increased as the mean ability increased for normal and skewed distributions. As drift magnitude increased, RMSE increased for all linking methods except for the LAV, which remained unchanged. The LAV recovered *b* the best for most conditions of drift, particularly -1.00 magnitude, followed by FPC and CC.

For the 50% drift conditions, RMSE typically increased for all linking methods as the normal and skewed distributions deviated further away from a mean of 0. As the magnitude of drift increased, RMSE increased for all linking methods. Between the linking methods, LAV and FPC performed the best, but under certain conditions. The LAV produced the smallest RMSE, which can be attributed to bias, for N(0,1) and when the magnitude of drift was the greatest (-1.00). However, at this level of drift, estimates of *b* are rather inaccurate. FPC had the smallest RMSE due to smaller bias values for N(0.5,1) and N(1,1) for the magnitudes of drift at -0.25 and -0.50. CC also performed similarly to FPC for the magnitudes of drift at -0.25 and -0.50.

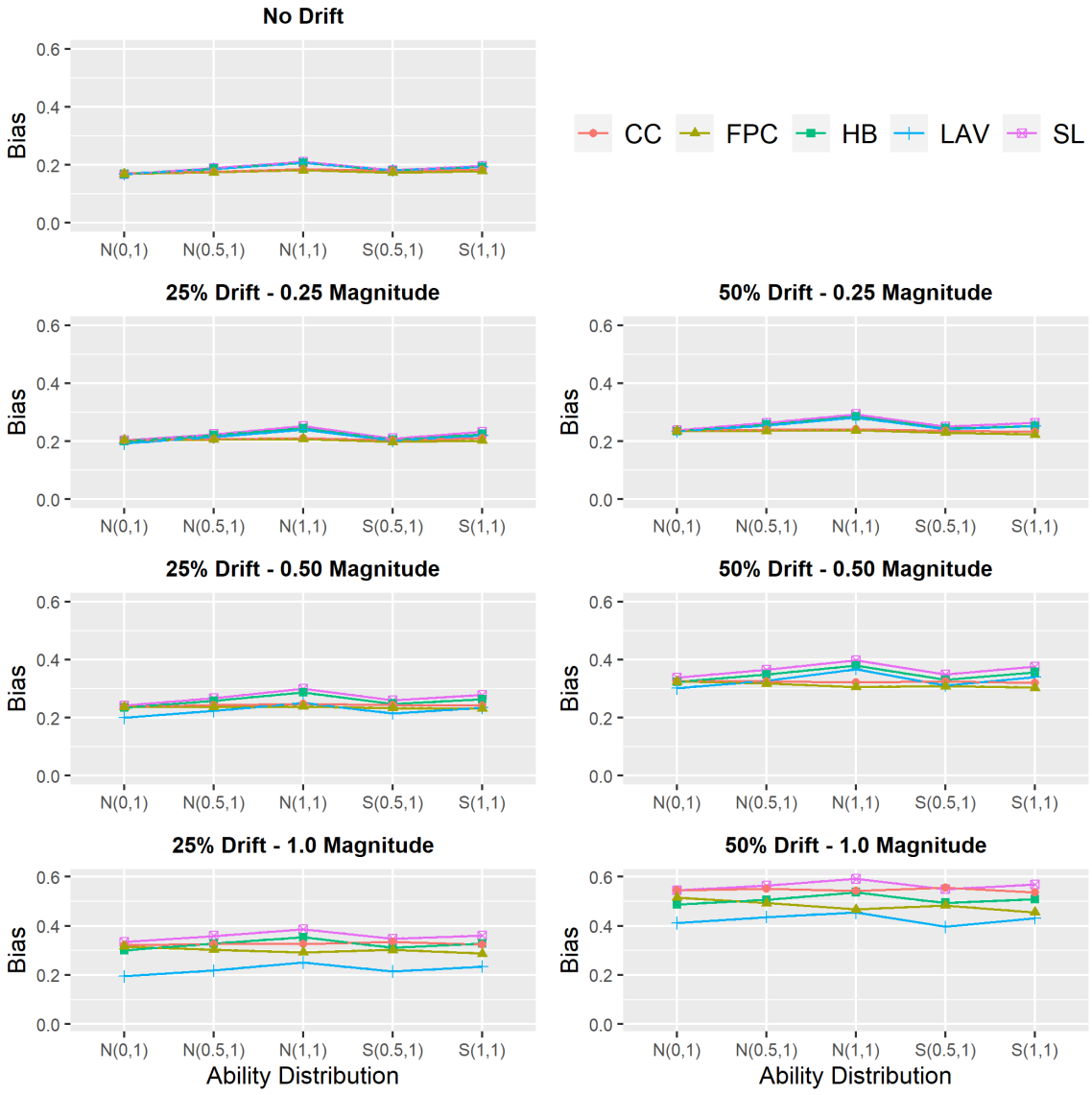

Figure 21. Bias Values for Item Estimate b – 1,000 Examinees.

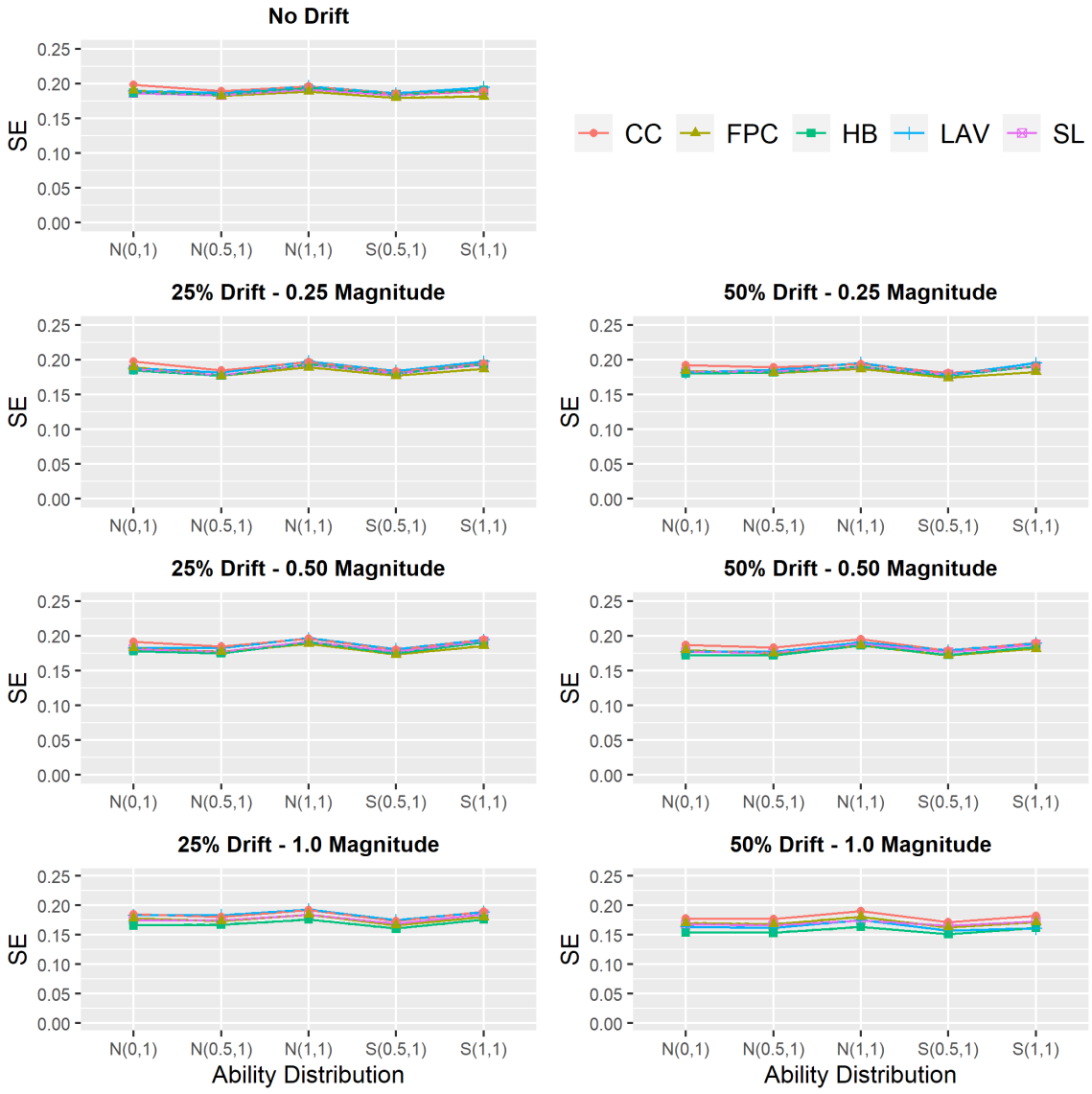

Figure 22. SE Values for Item Estimate b – 1,000 Examinees.

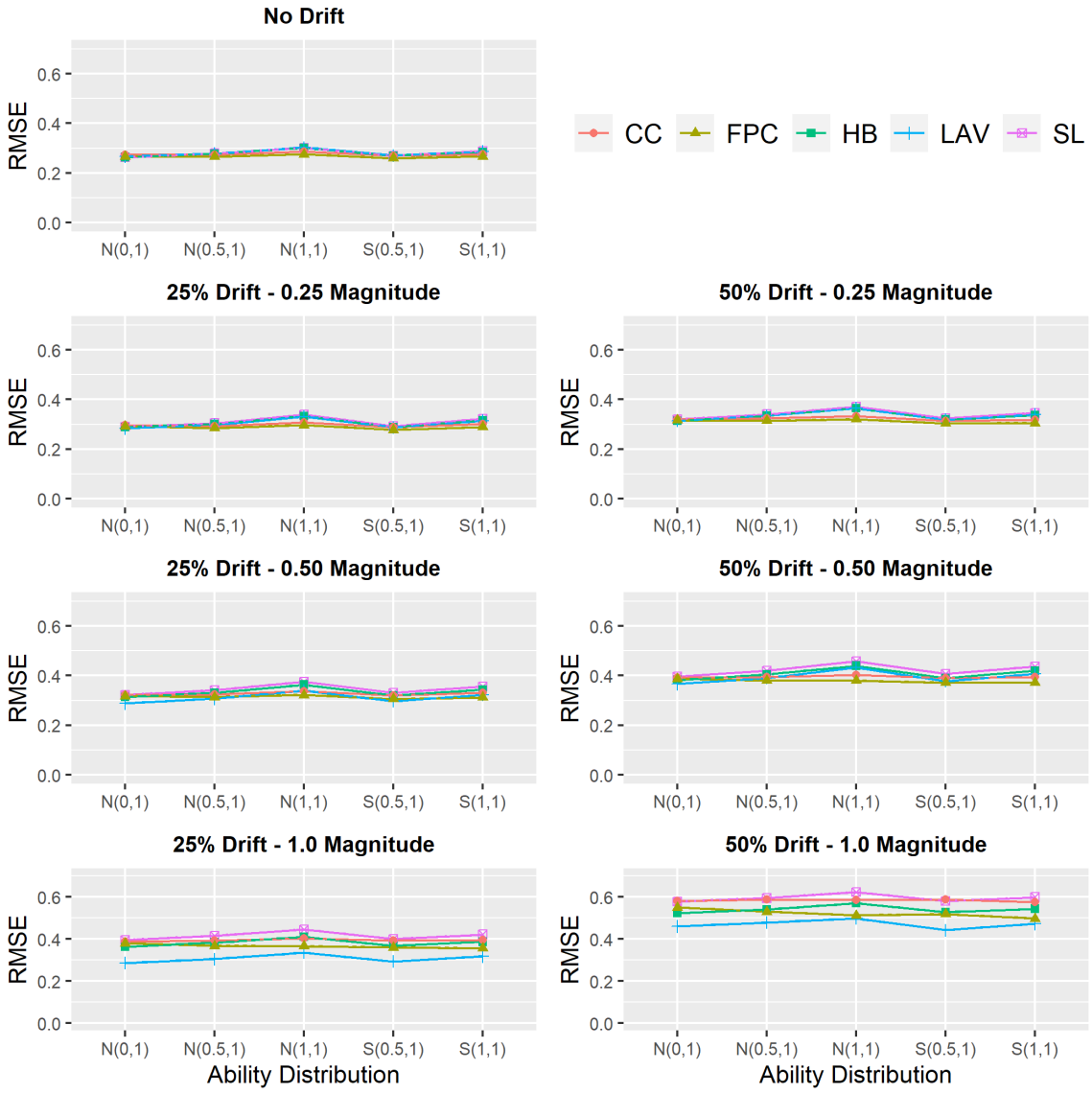

Figure 23. RMSE Values for Item Estimate b – 1,000 Examinees.

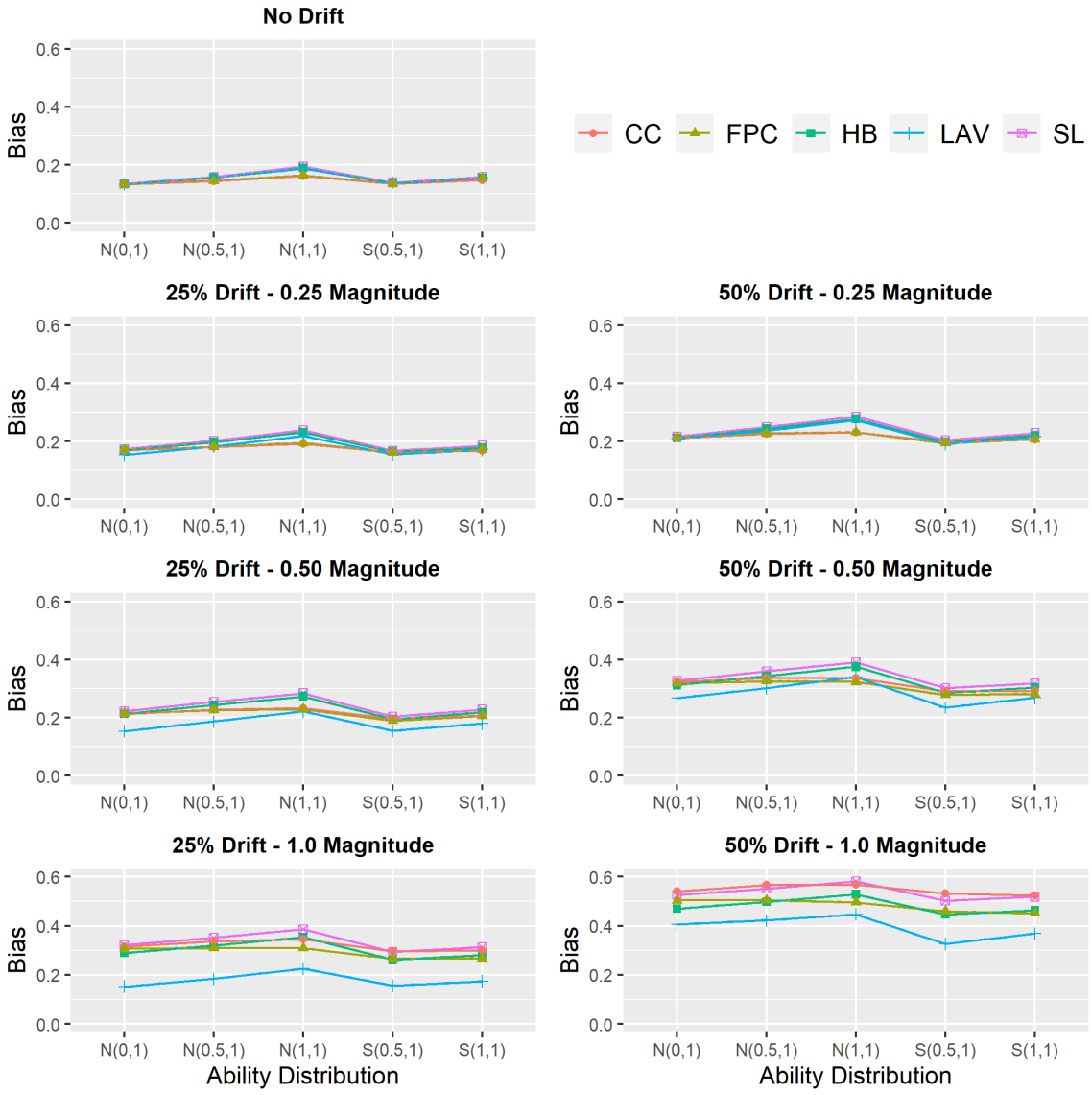

Figure 24. Bias Values for Item Estimate b – 3,000 Examinees.

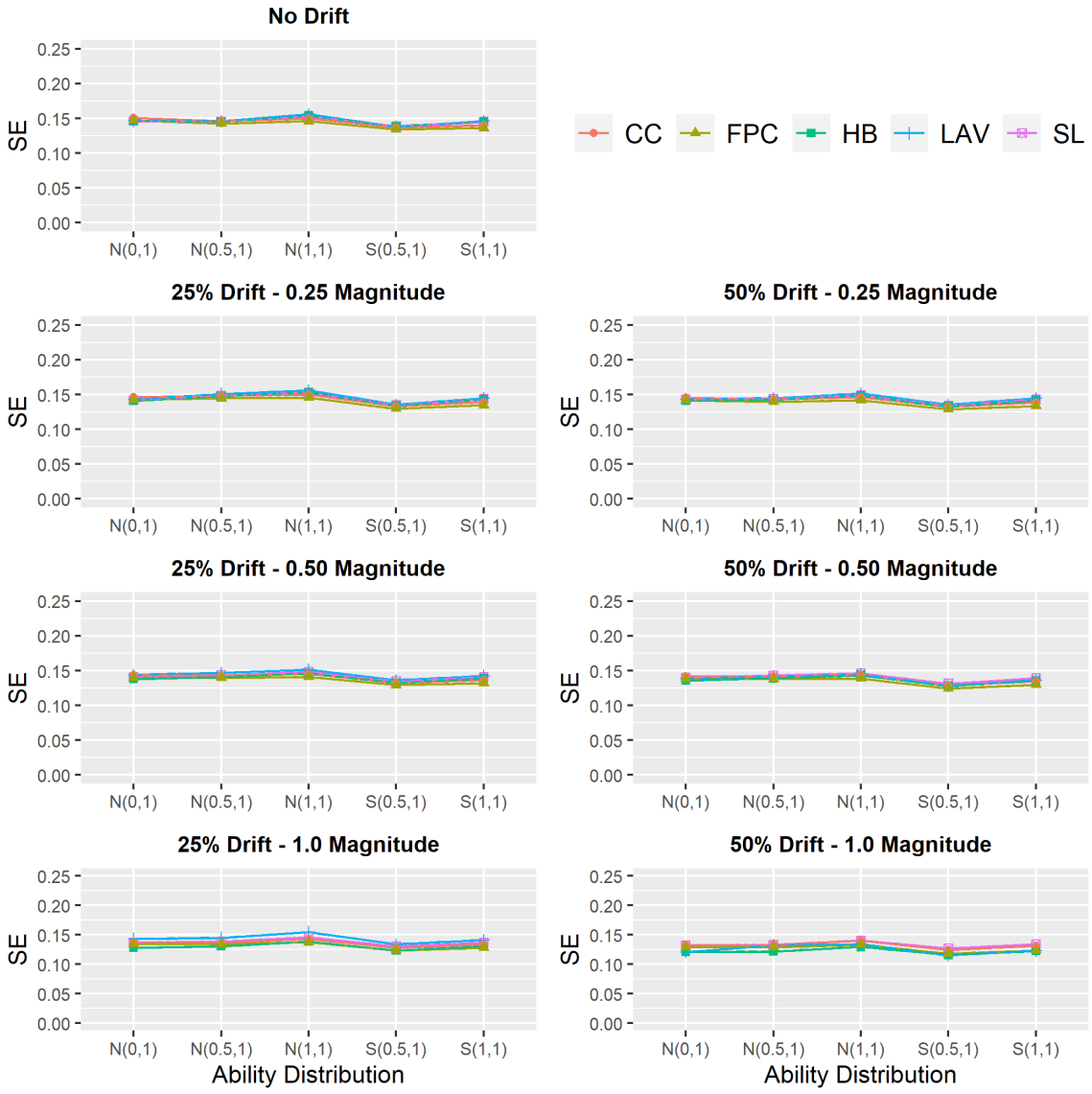

Figure 25. SE Values for Item Estimate b – 3,000 Examinees.

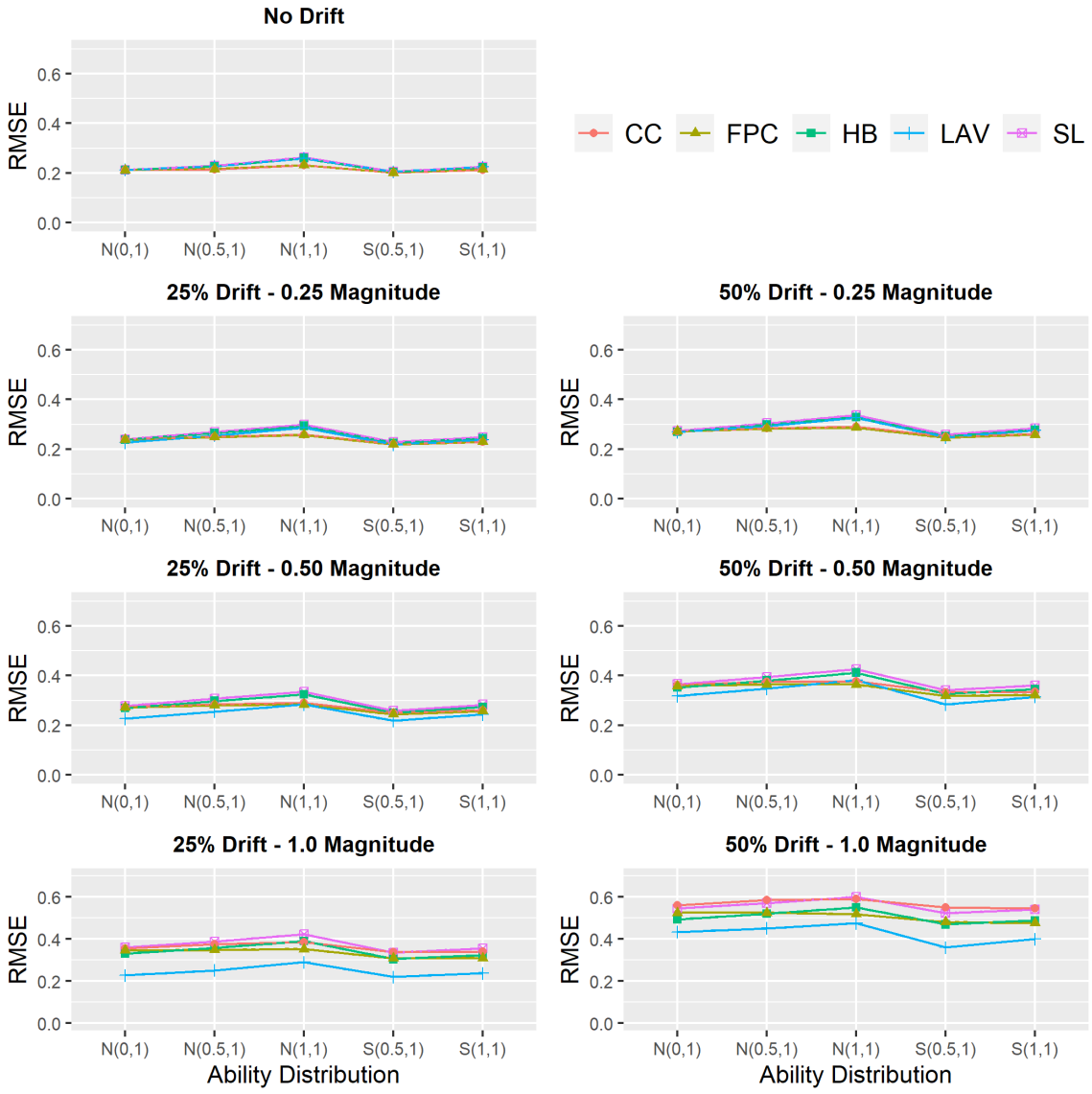

Figure 26. RMSE Values for Item Estimate b – 3,000 Examinees.

Linked Item Parameter Estimate c. Figures 27 – 29 illustrate the bias, SE, and RMSE values for the pseudo-guessing parameter under 1,000 examinees. Figures 30 – 32 illustrate the bias, SE, and RMSE values for the pseudo-guessing parameter under 3,000 examinees. Specific values for each of these three outcomes can be found in Appendix C. These outcomes were also evaluated for all items, common and unique, as found in Appendix D.

Among the 1,000 sample-size conditions, recovery of the pseudo-guessing parameter was nearly identical for the separate calibration methods for all drift conditions. RMSE was nearly identical for the CC and FPC methods. RMSE increased as the mean ability of the population increased. CC and FPC produced slightly smaller values of RMSE for all non N(0,1) conditions, although this difference was negligible. RMSE remained relatively stable as the magnitude of drift increased, although a small decline could be observed.

Among the 3,000 sample-size conditions, RMSE and bias slightly improved relative to the 1,000 sample-size, despite a small increase in SE (although this difference was negligible). RMSE values were practically identical for the separate calibration methods for all drift conditions. RMSE values were approximately the same between CC and FPC. As the mean ability of examinees increased, RMSE increased for all linking methods. As drift magnitude increased, RMSE remained stable for all linking methods.

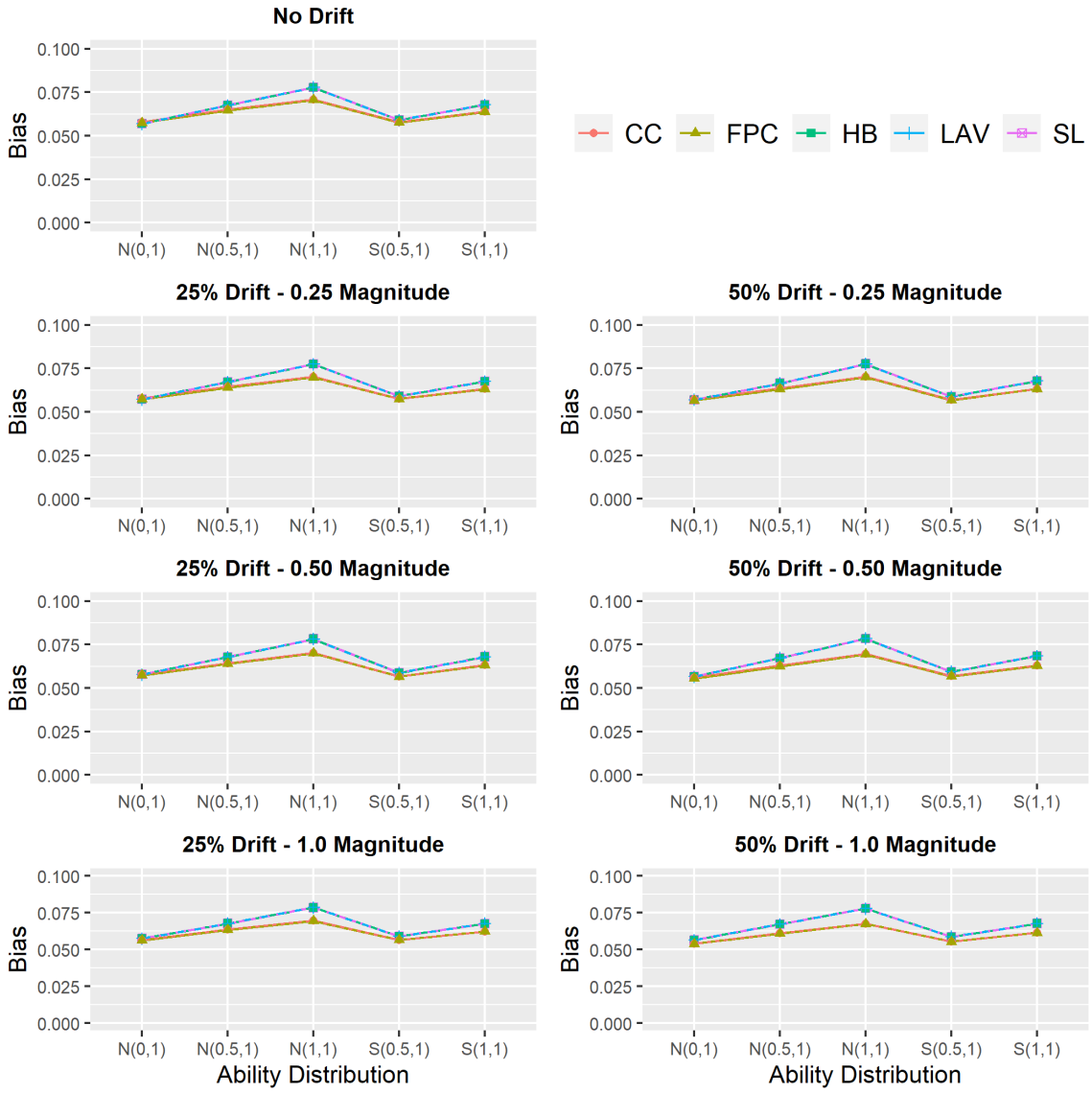

Figure 27. Bias Values for Item Estimate c – 1,000 Examinees.

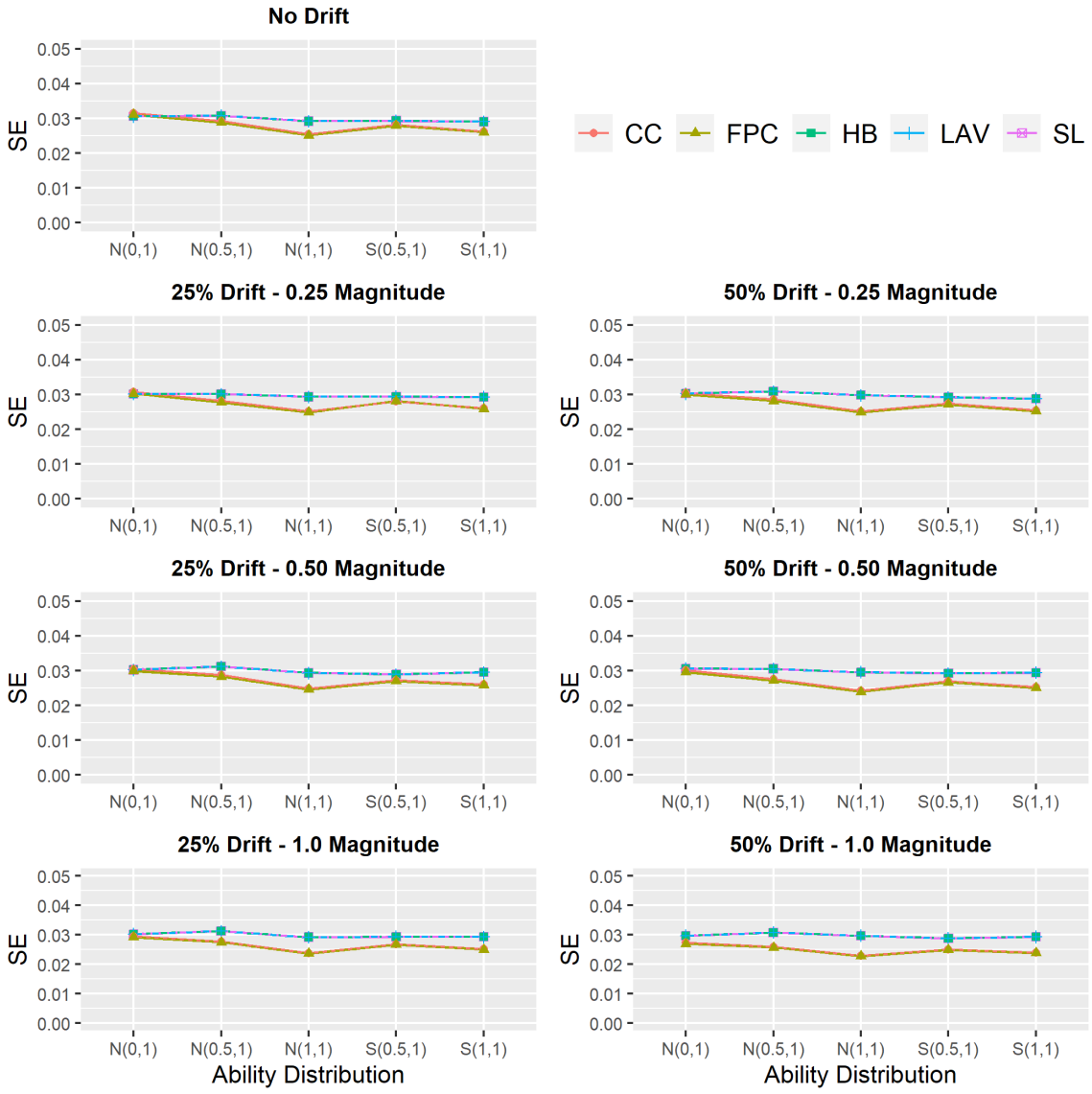

Figure 28. SE Values for Item Estimate c – 1,000 Examinees.

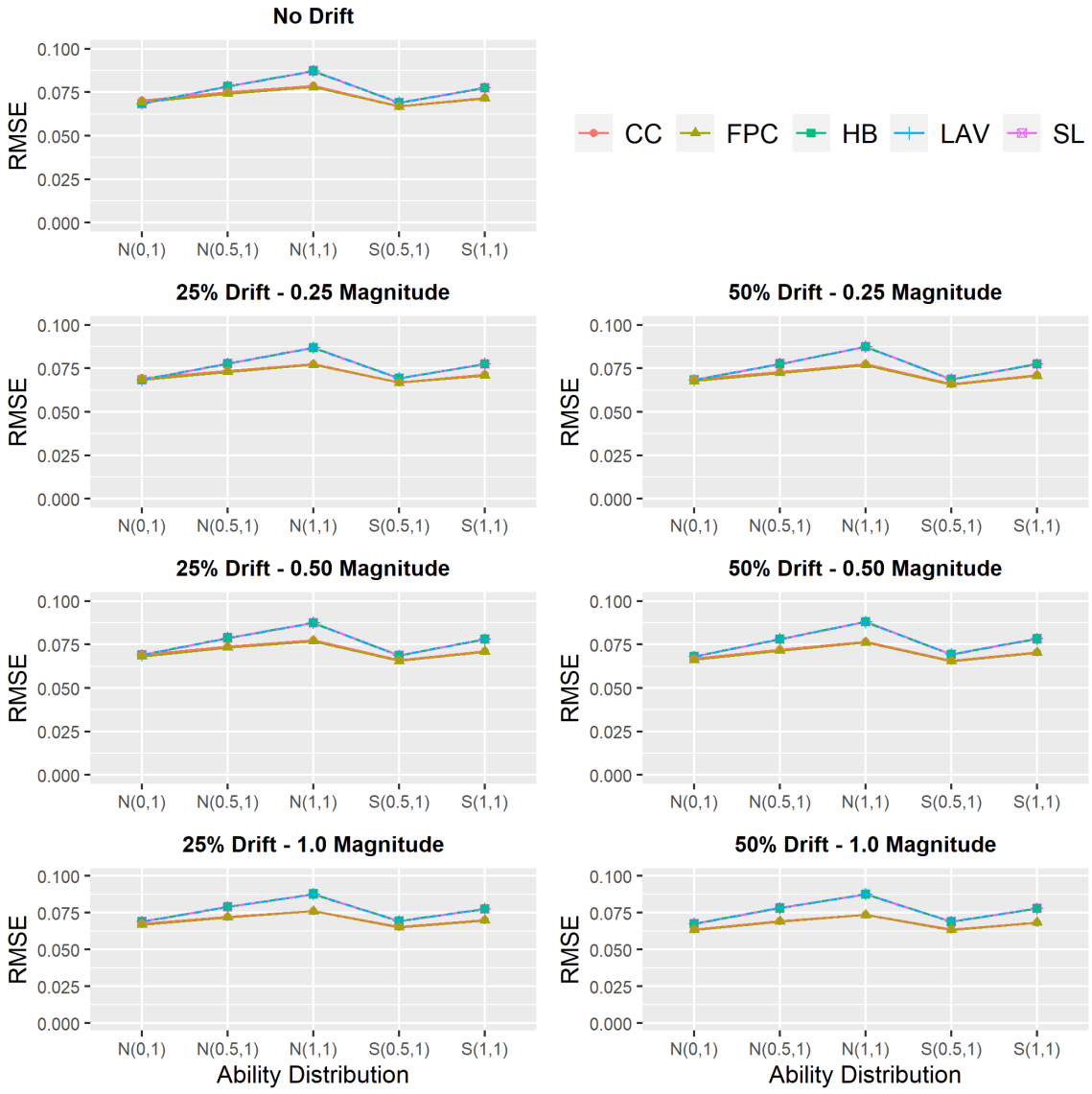

Figure 29. RMSE Values for Item Estimate c – 1,000 Examinees.

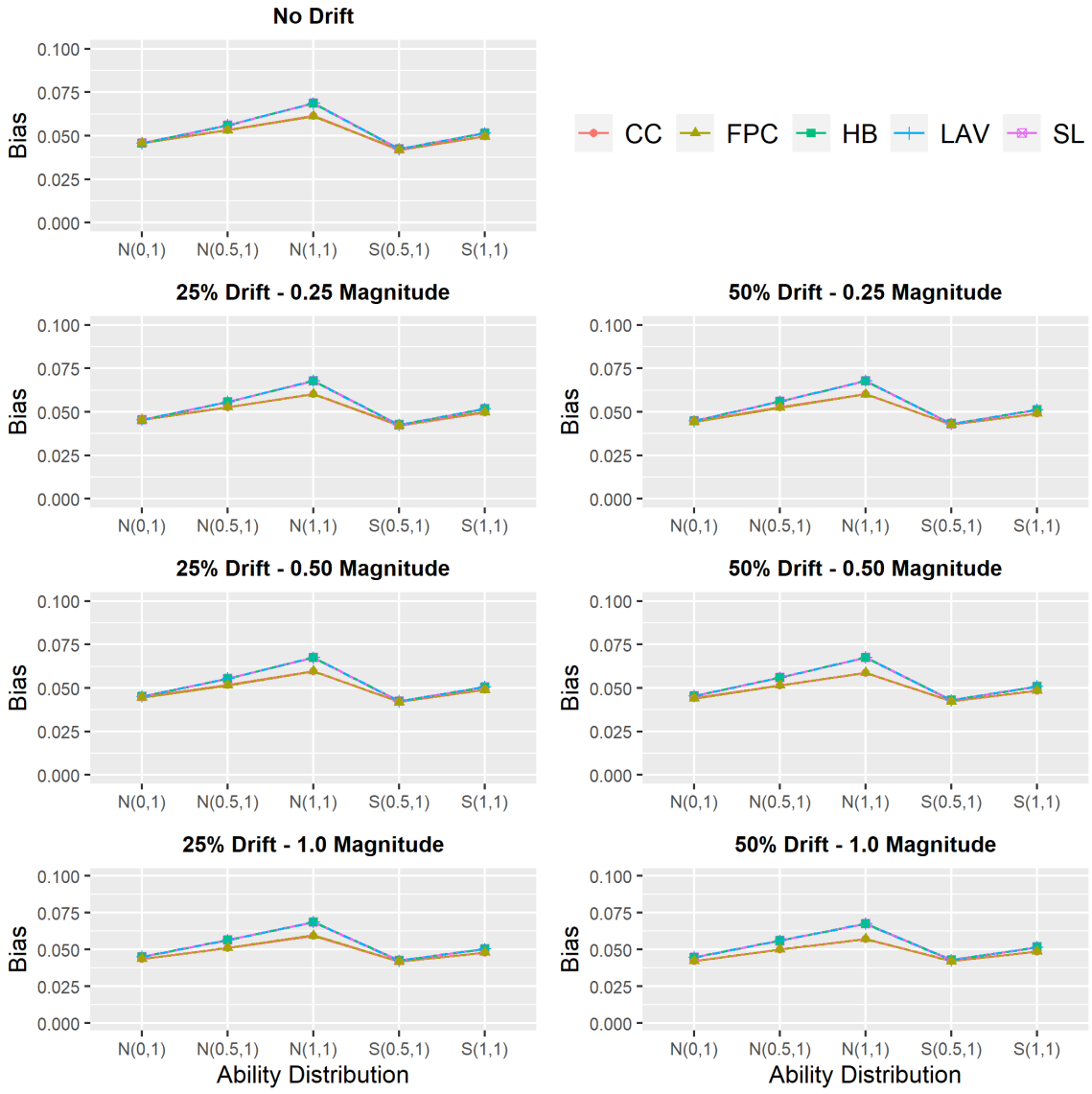

Figure 30. Bias Values for Item Estimate c – 3,000 Examinees.

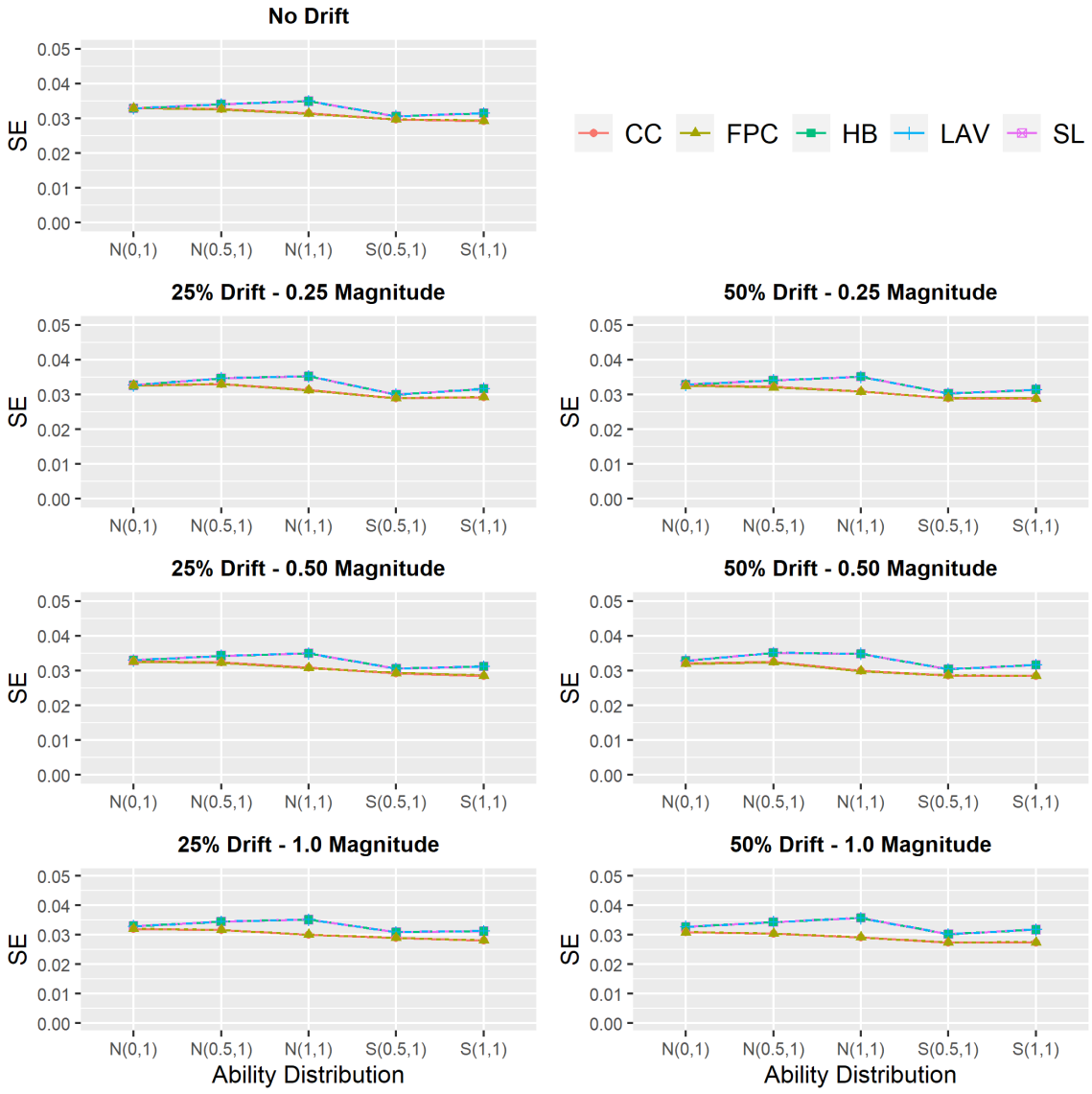

Figure 31. SE Values for Item Estimate c – 3,000 Examinees.

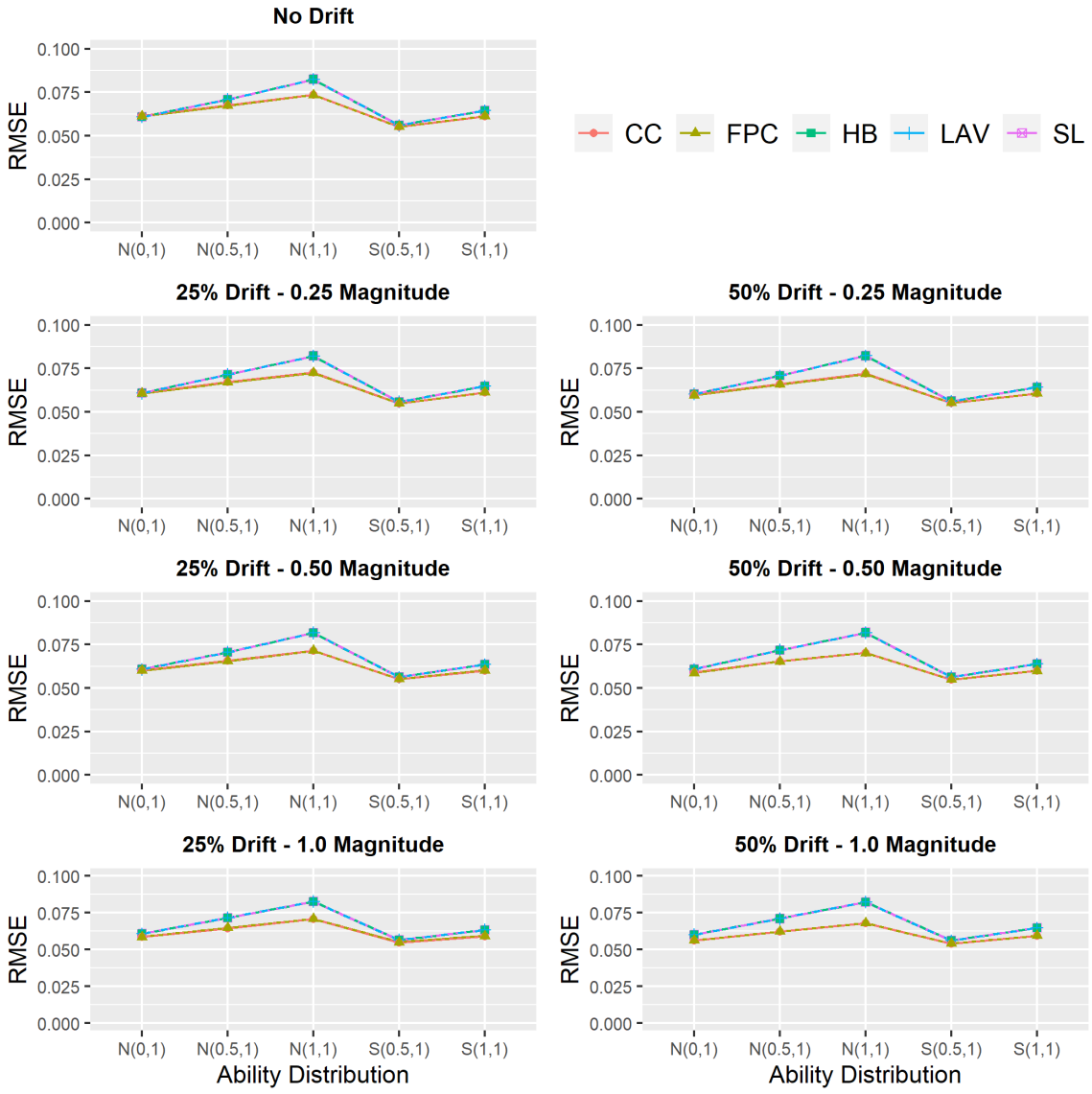

Figure 32. RMSE Values for Item Estimate c – 3,000 Examinees.

Equated Scores. The third research question examined how consequential the effect of IPD was on equated scores using IRT true and observed score equating. For the purposes of evaluation, the difference that matters (DTM; Dorans & Feigenbaum, 1994) threshold was used to determine how well equated scores were recovered from a practical standpoint. DTM is defined by an absolute value of 0.50, the point at which a score would be considered for rounding up to the next integer.

Equated Scores with IRT True Score Equating. Bias, SE, and RMSE were calculated by comparing the estimated scores using IRT true score equating to the criterion equating relationship. The criterion equating relationship used the equated scores obtained from the generating item parameters for the baseline condition. Equated scores below the lower asymptote were ignored so the smallest score available (i.e., 25) for all linking methods, conditions, and sample sizes was chosen. Figures 33 – 35 illustrate the average bias, SE, and RMSE values for equated scores under IRT true score equating for 1,000 examinees. Figures 36 – 40 plot the conditional RMSE values for each score point across the scale for 1,000 examinees. Figures 41 – 43 illustrate the bias, SE, and RMSE values for equated scores under IRT true score equating for 3,000 examinees. Figures 44 – 48 plot the conditional RMSE values for each score point across the scale for 3,000 examinees. Specific values for each of these three outcomes can be found in Appendix E.

Overall, equated scores (true and observed) were most heavily influenced by drift out of all the outcomes. This may occur due to the trickle-down effect that drift has on the linking and equating process. The RMSE from the linking constants combines with the

RMSE from the item parameter estimates to produce inflated RMSE values in the equated scores. In most conditions, findings for bias and RMSE exceeded the DTM threshold, which is a major issue considering one raw score point can be the difference between passing and failing. These findings are described in more detail below.

For the 1,000 sample-size condition, RMSE for the equated scores was lowest when no drift was present. As the mean ability of examinees increased, RMSE increased for all linking methods. However, RMSE exceeded the DTM criterion under all ability distributions. Inspection of the bias values revealed that the separate calibration methods produced bias lower than the DTM for all abilities except N(1,1). CC exceeded the DTM threshold for all ability distributions except N(0,1). FPC did not exceed the DTM threshold for any ability distribution. Overall, the separate calibration methods produced the least bias for the skewed distributions and N(0,1), whereas FPC had the least amount of bias for N(1,1). Although these results were unexpected, Jurich et al. (2012) reported RMSE values for equated true scores near 5 score points for linking between equivalent groups – N(0,1) – without any cheating for the SL, HB, and FPC methods. Hu et al. (2008) reported MSE values for equated true scores that exceeded one score point for SL, CC, and FPC for linking nonequivalent groups of N(0,1) and N(1,1).

When 25% of common items were drifted, the RMSE from all linking methods exceeded the DTM threshold for each ability distribution and drift magnitude. Looking at values of bias, the SL, HB, and CC methods exceeded the DTM for nearly all conditions. The LAV method produced bias near the DTM for most of the ability distributions. FPC produced values of bias at or below DTM for -0.25 magnitude but above DTM for -0.50

and -1.00 magnitudes. No systematic pattern of RMSE nor bias could be discerned for the linking methods as ability increased. For the skewed distributions, RMSE and bias decreased from S(0.5,1) to S(1,1) for all linking methods and conditions with the exception of the LAV for the -1.00 magnitude condition. This decrease might be attributed to the skewed distributions having more examinees with higher probabilities of answering items correctly. As drift magnitude increased, bias and RMSE increased for all linking methods with the exception of the LAV for N(0,1), N(0.5,1), and S(0.5,1). Overall, the LAV method performed exceptionally well, producing the least amount of RMSE and bias for most conditions. FPC performed better than the LAV for a few conditions – N(0.5,1) and N(1,1) with -0.25 drift magnitude, and N(1,1) with -0.50 drift magnitude. SE values for all linking methods were around the 0.50 threshold.

When 50% of common items were drifted, values of RMSE were well above the DTM threshold for all linking methods. Values of bias were near one or higher for all linking methods and conditions. SE was near 0.5 for all linking methods and conditions. There was no systematic pattern of RMSE nor bias for the separate calibration methods as ability increased under the normal distributions. RMSE and bias did decrease for the separate calibration methods as ability increased under the skewed distributions. For CC and FPC, RMSE and bias decreased as ability increased for both normal and skewed distributions. Lower RMSE and bias might be produced at higher ability distributions because examinees are already receiving high scores, with or without drift. As drift magnitude increased, RMSE and bias increased for all linking methods and all conditions. Under -0.25 drift magnitude, all linking methods performed fairly similarly in

terms of RMSE and bias. LAV had the lowest RMSE and bias for most conditions under -0.50 drift magnitude except for N(1,1), which FPC performed the best. For -1.00 magnitude, LAV had the smallest amount of RMSE and bias among all distributions. CC appeared to be most influenced by drift.

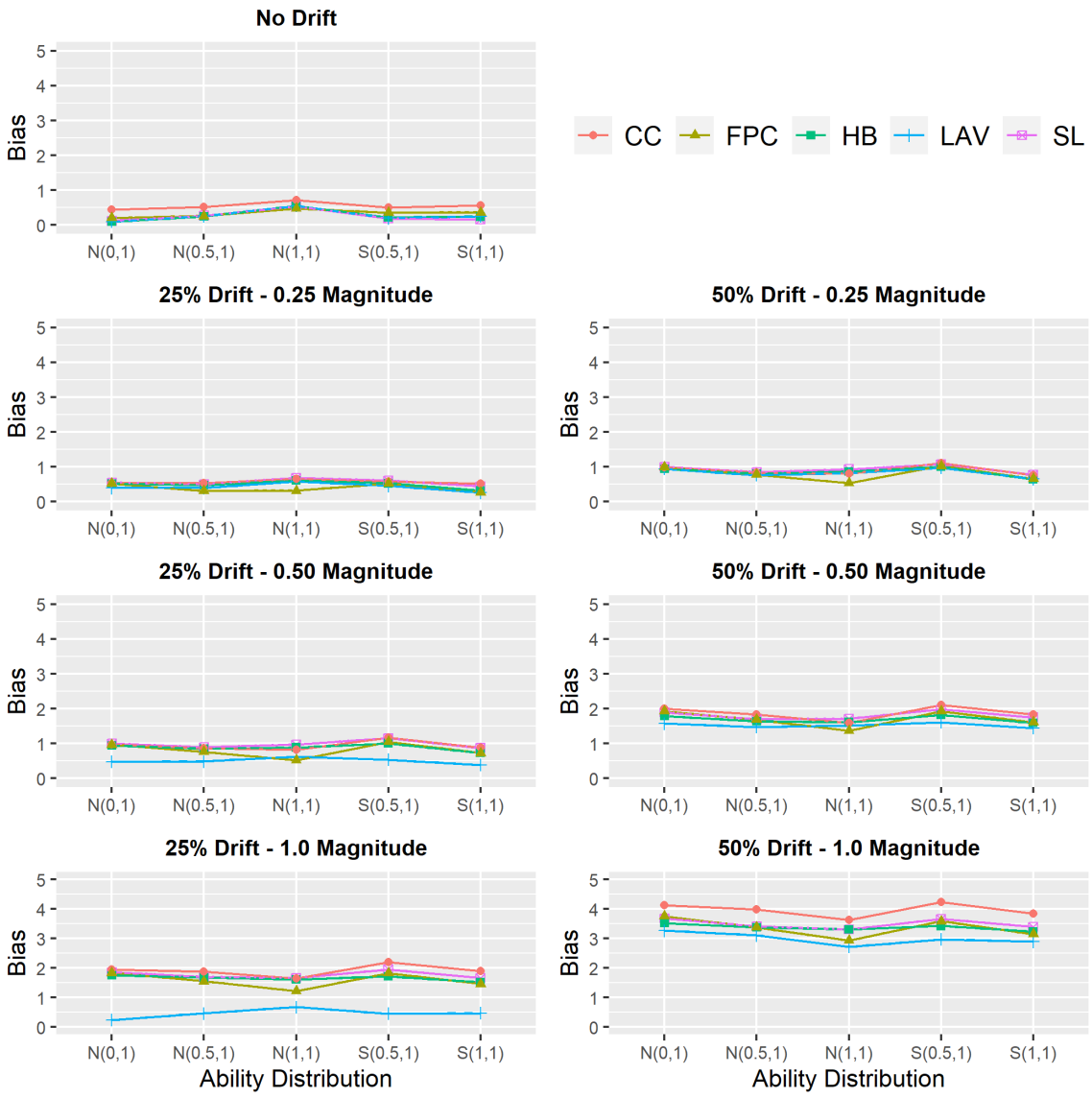

Figure 33. Bias Values for True Scores – 1,000 Examinees.

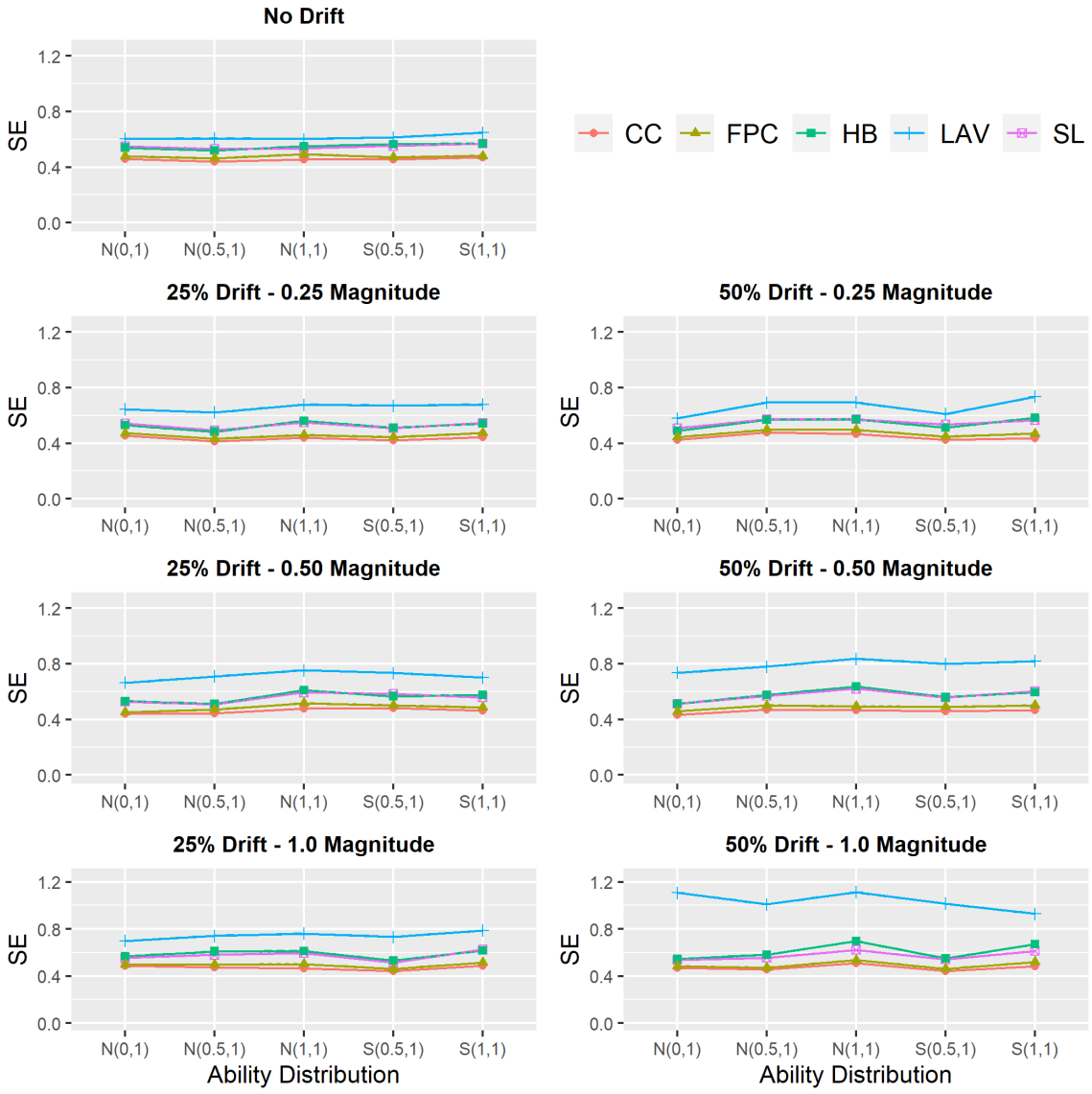

Figure 34. SE Values for True Scores – 1,000 Examinees.

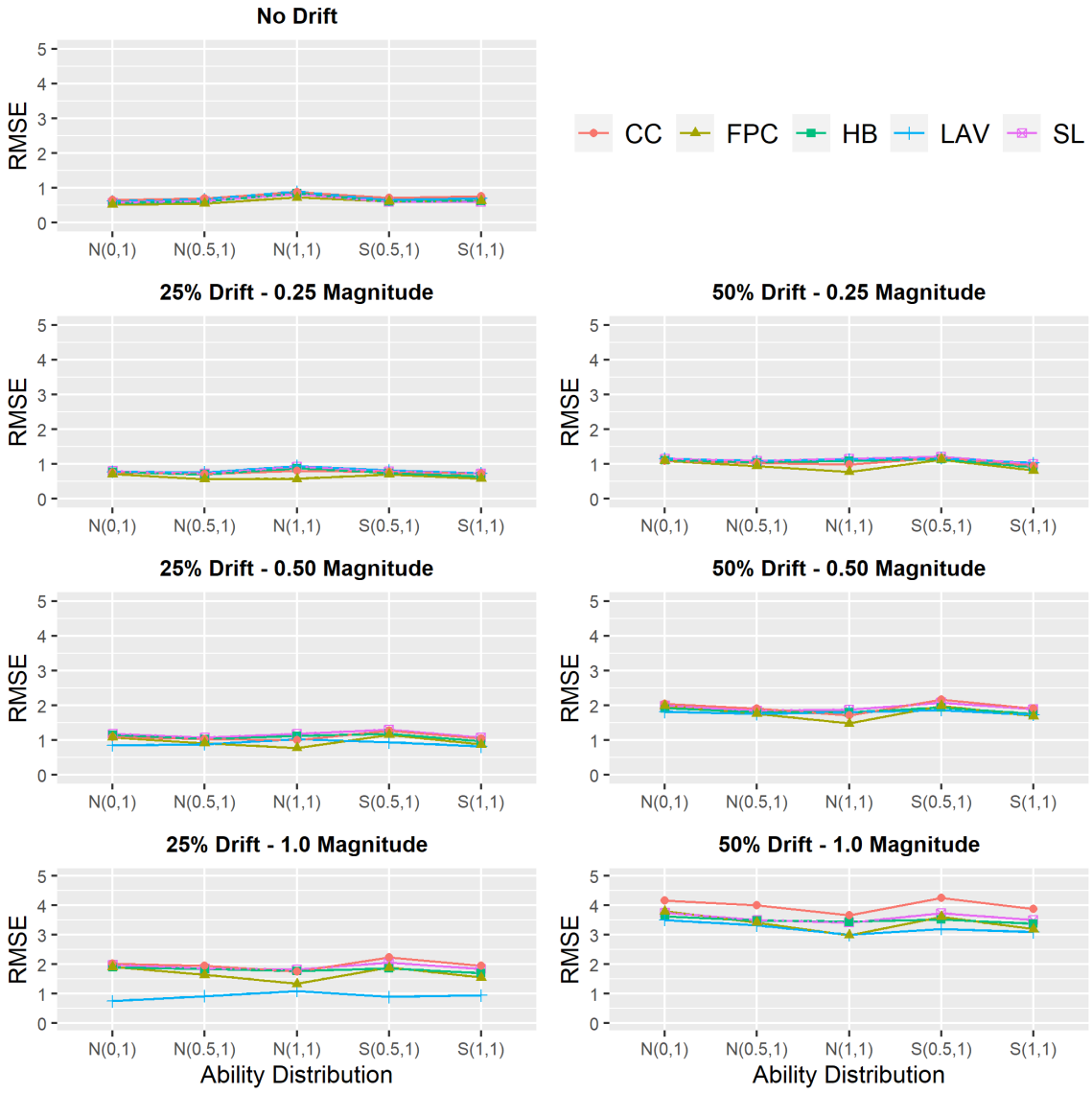

Figure 35. RMSE Values for True Scores – 1,000 Examinees.

Across all linking methods, RMSE decreased as the mean skewed ability increased from S(0.5, 1) to S(1, 1) for all drift conditions except the baseline (no drift). RMSE decreased as the normal ability distributions increased, but this pattern only occurred under the most extreme condition of drift. Under the other conditions of drift, RMSE decreased selectively by each linking method. These findings warranted further inspection, as it was expected that RMSE would increase as the mean ability increased for both normal and skewed ability distributions.

Conditional RMSE values were plotted for each linking method to identify which points along the scale produced the highest RMSE. RMSE values are provided for scores ranging between 25 and 100 because linear interpolation was not used to obtain scores below the sum of the pseudo-guessing parameters for the simulation. Figures 36 – 40 illustrate the conditional RMSE for the 1,000 sample-size of the SL, HB, LAV, CC, and FPC methods, respectively.

For the baseline conditions (top left panels) of each linking method, RMSE tended to be higher at the lower and higher ends of the scale. This is particularly true for the N(1,1) condition. This occurs because there are fewer examinees obtaining scores at these locations, which results in higher RMSE values. For the drift magnitudes of -0.25 and -0.50 (for both 25% and 50% drifted items), RMSE was interspersed evenly in the middle of the scale (i.e., between 20 and 80) where most of the examinee scores are located. However, the RSME for CC (Figure 39) tended to increase between the scores of 60 and 90, where more scores are expected to be compared to the lower and higher ends of the scale. This might explain why CC was most heavily influenced by drift.

For the drift magnitude of -1.0 (bottom row), the SL and HB methods produced the highest RMSE values from the bottom of the scale to approximately 60. The LAV method was less affected by drift for the 25% drifted item, -1.0 magnitude condition (Figure 38, bottom left) as RMSE was distributed uniformly until a score of 80, where RMSE peaked. For the 50% drifted item, -1.0 magnitude condition (Figure 38, bottom right), the LAV was profoundly influenced by drift (similar to the other linking methods), but RMSE peaked between 40 and 50, and also had a second peak around 90. This second peak was also noticeable with the HB method. CC had RMSE values that peaked between 60 and 90, while FPC had RMSE values that were highest between 40 and 80.

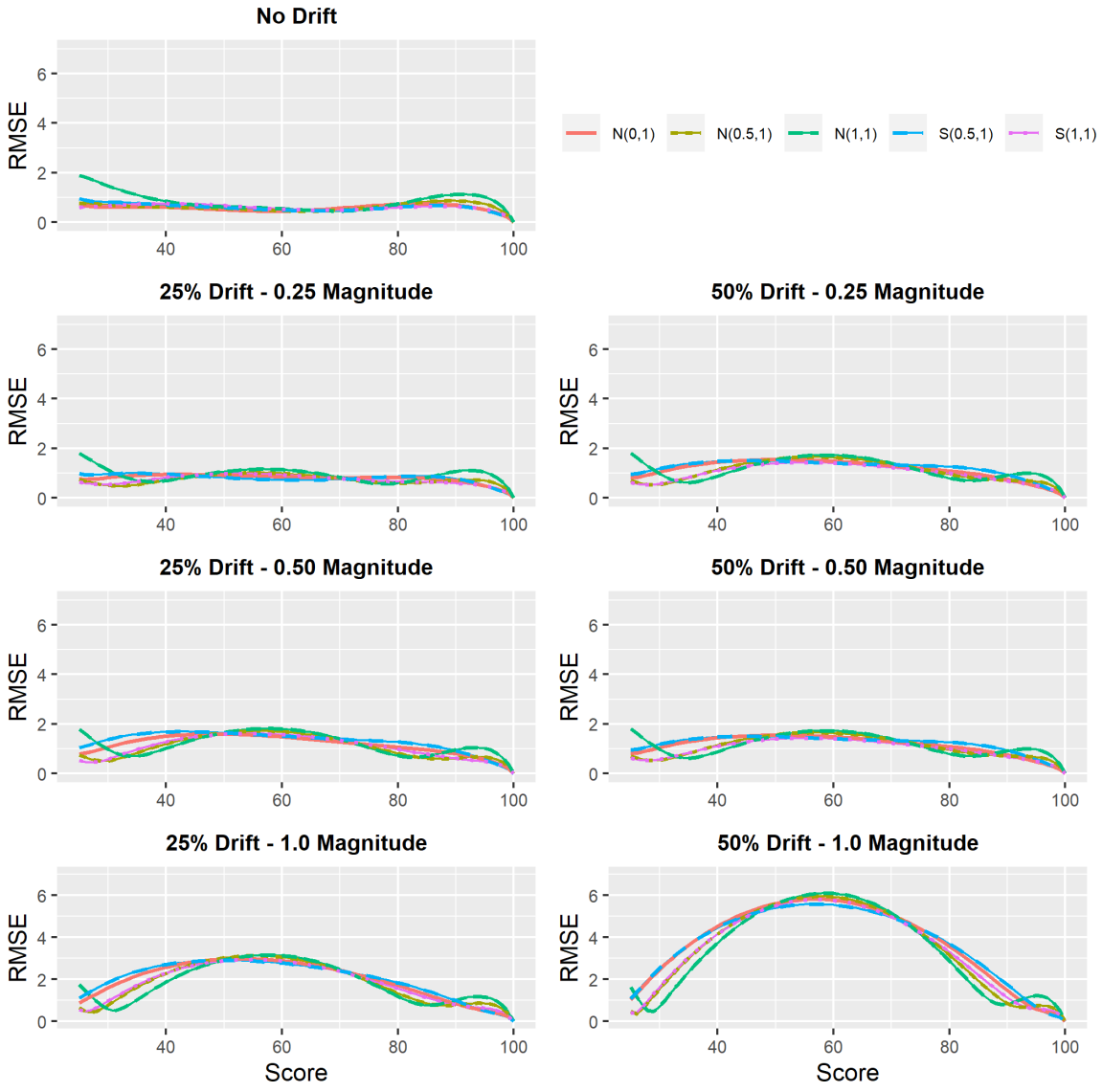

Figure 36. Conditional RMSE for SL True Scores – 1,000 Examinees.

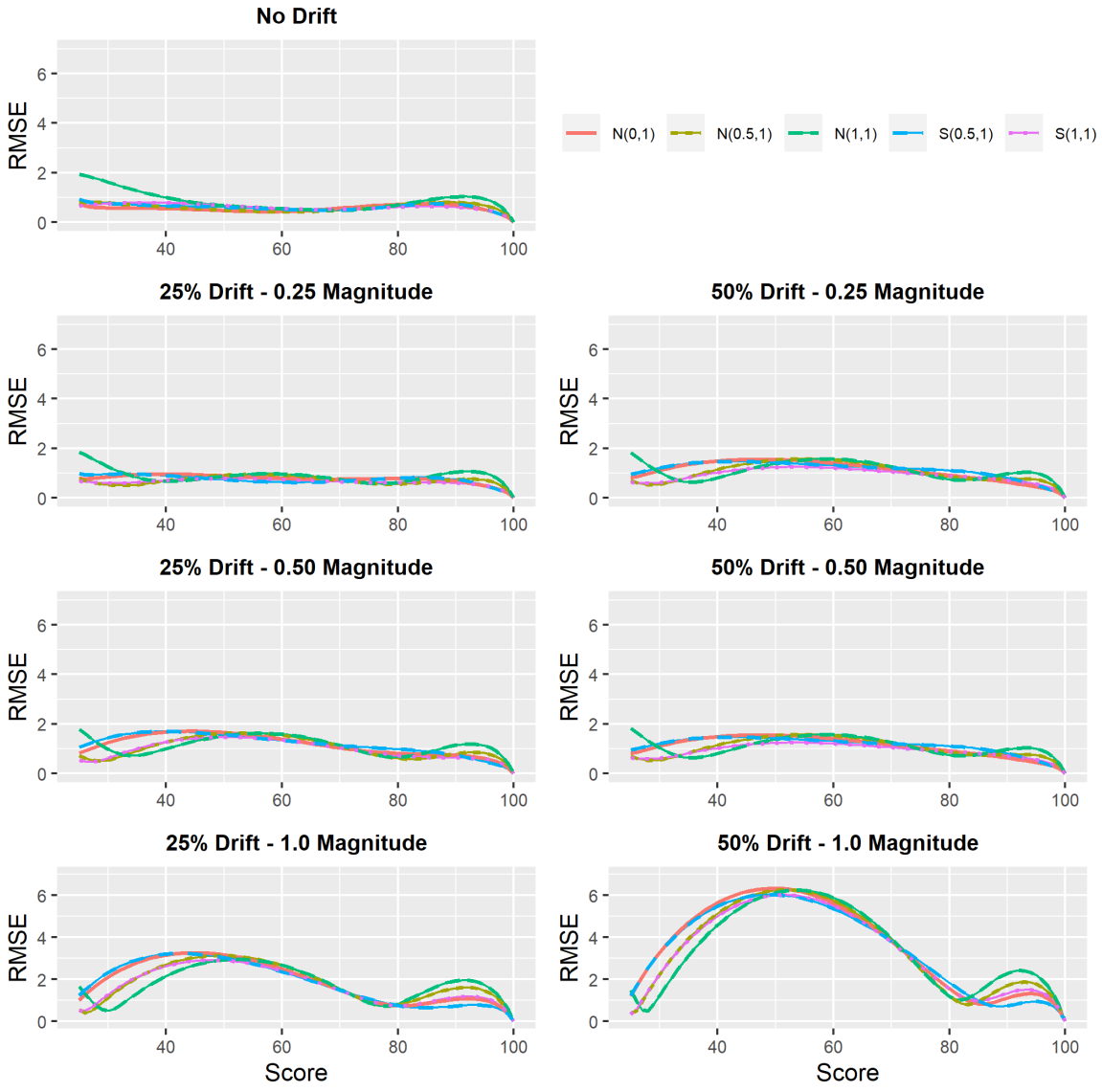

Figure 37. Conditional RMSE for HB True Scores – 1,000 Examinees.

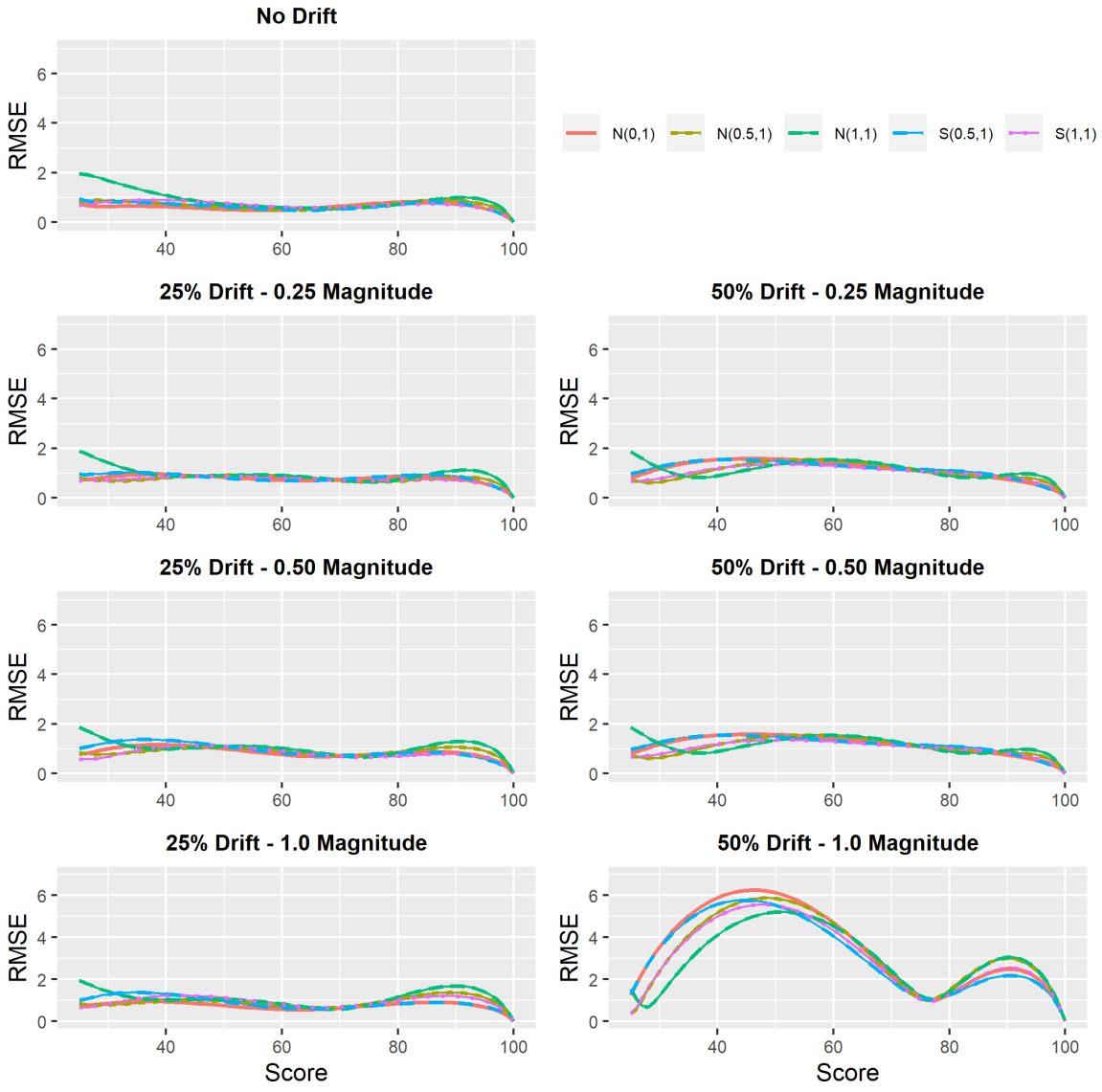

Figure 38. Conditional RMSE for LAV True Scores – 1,000 Examinees.

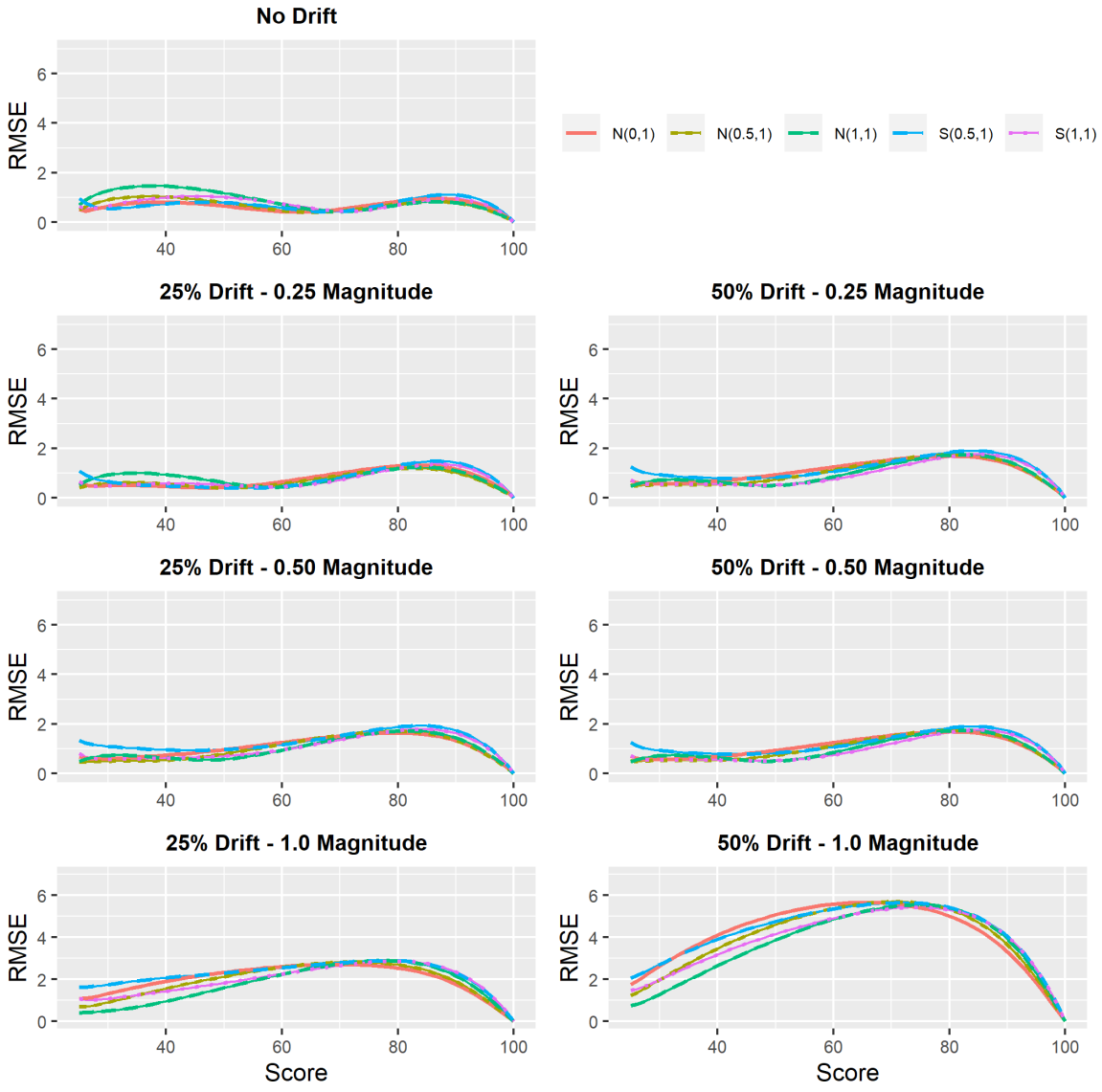

Figure 39. Conditional RMSE for CC True Scores – 1,000 Examinees.

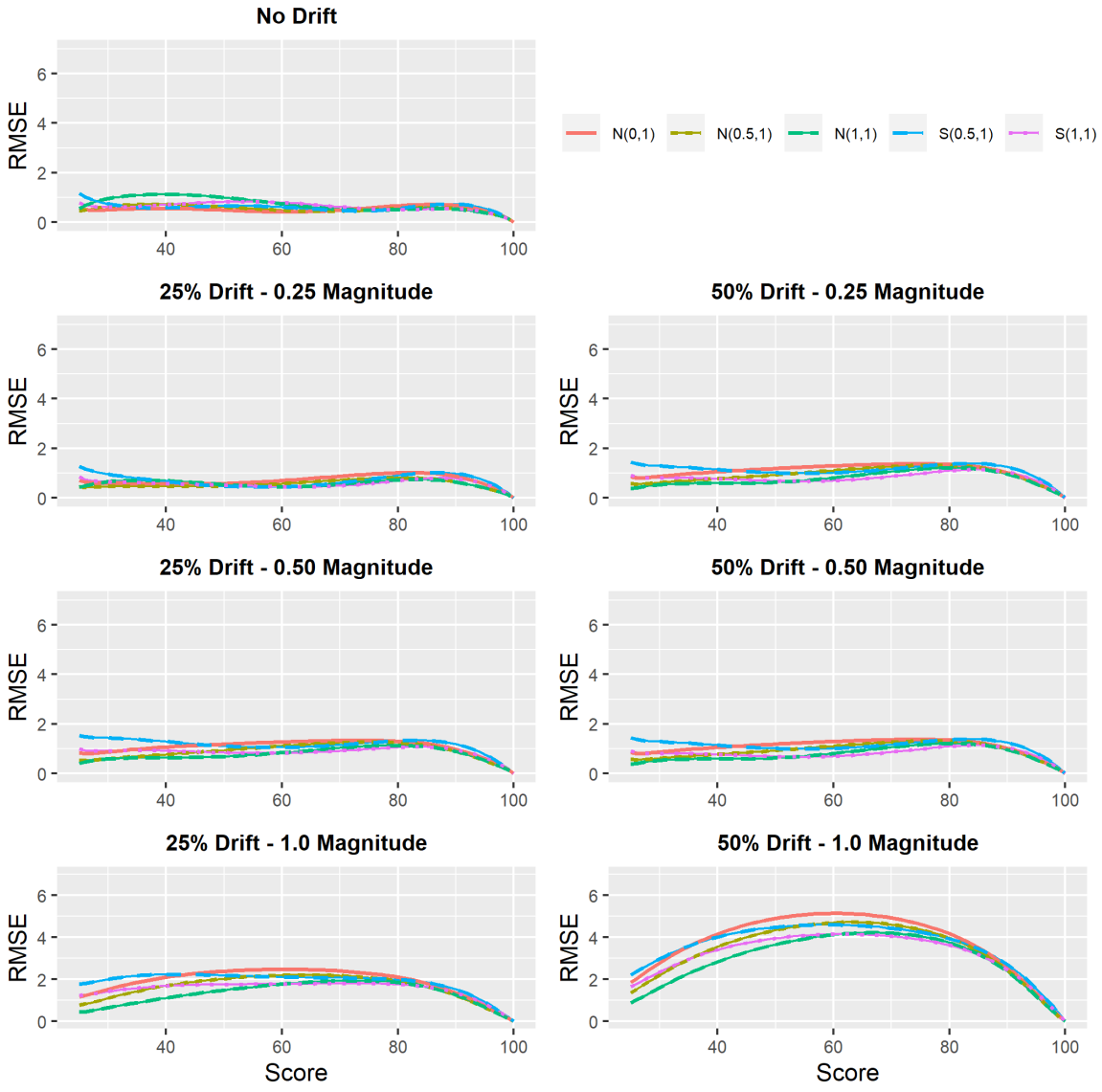

Figure 40. Conditional RMSE for FPC True Scores – 1,000 Examinees.

Compared to the 1,000 sample-size conditions, RMSE was smaller under the 3,000 sample-size conditions when no drift was present. This is reflected by the smaller values of bias and SE. Yet, RMSE values exceeded the DTM threshold for all linking methods under N(1,1) and S(0.5,1) – except for FPC N(1,1). No values of bias for any of the linking methods exceeded the 0.5 threshold. Both RMSE and bias increased for all linking methods as mean ability increased for the normal distributions. However, bias decreased for all linking methods as mean ability increased for the skewed distributions. The separate calibration methods yielded the smallest amount of bias for N(0,1) and S(1,1), whereas FPC yielded the smallest bias for the other ability distributions.

Under the 25% drifted item conditions, RMSE exceeded DTM for nearly all conditions and linking methods. With the exception of the LAV, bias exceeded DTM for the other linking methods for most conditions. Bias for the LAV only exceeded DTM for the N(1,1) condition. For all linking methods, there was no discernable pattern for bias as ability increased for the normal distributions, but bias did decrease for the skewed distributions. As the magnitude of drift increased, bias increased for all linking methods except for LAV. The bias for LAV decreased for N(0,1) and the skewed distributions. Overall, the LAV method performed the best despite having slightly elevated SE values.

For the 50% drifted item conditions, RMSE and bias was near or exceeded one for all linking methods. As the mean ability increased for the normal distributions, bias decreased for CC and FPC, although there was no systematic pattern of bias for the separate calibration methods. As mean ability increased for the skewed distributions, bias decreased for nearly all linking methods and conditions. When drift magnitude increased,

bias increased for all linking methods and conditions. The LAV performed the best among all linking methods even with elevated SE levels. CC produced the smallest SE values although appeared the most susceptible linking method to drift.

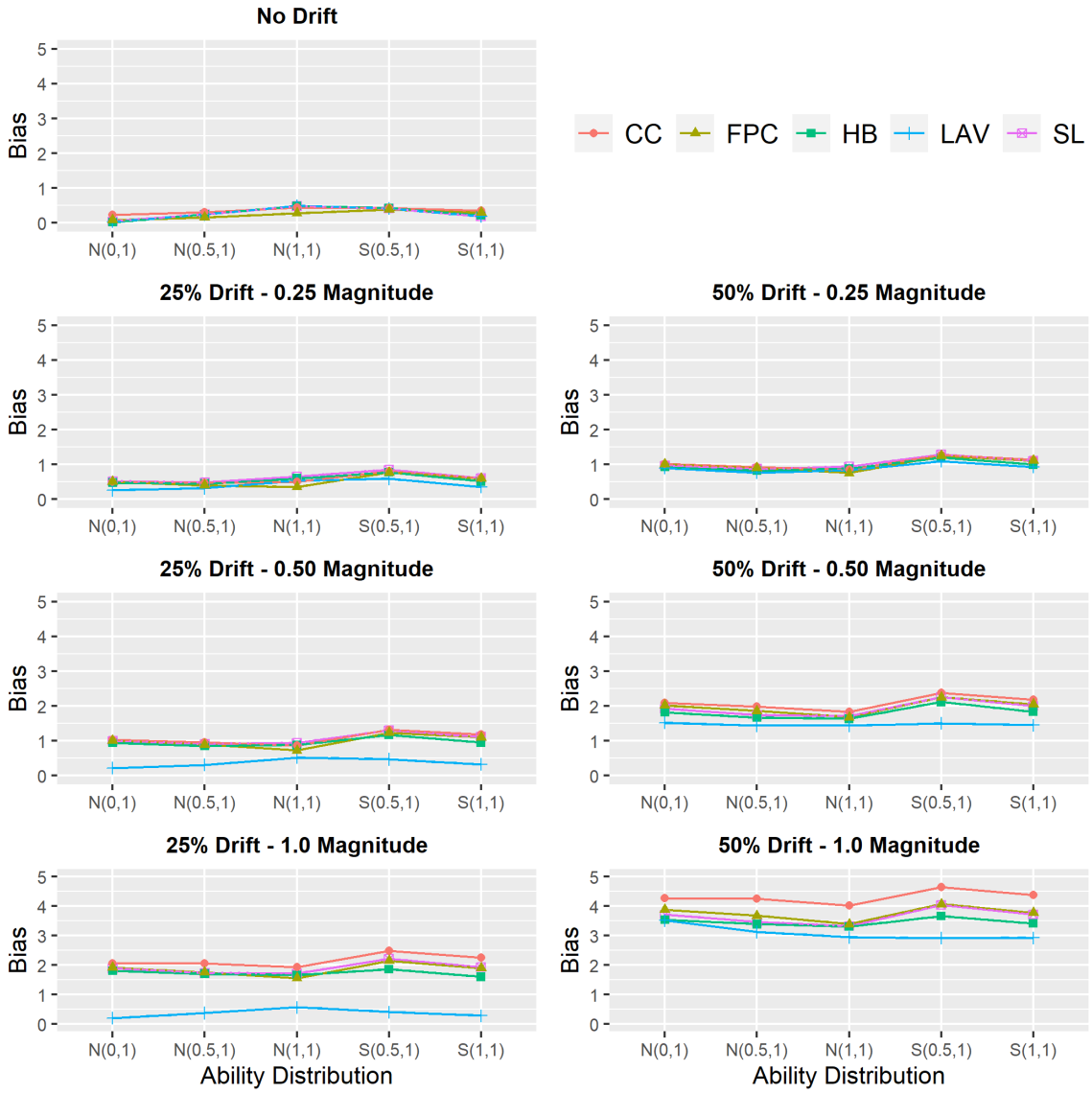

Figure 41. Bias Values for True Scores – 3,000 Examinees.

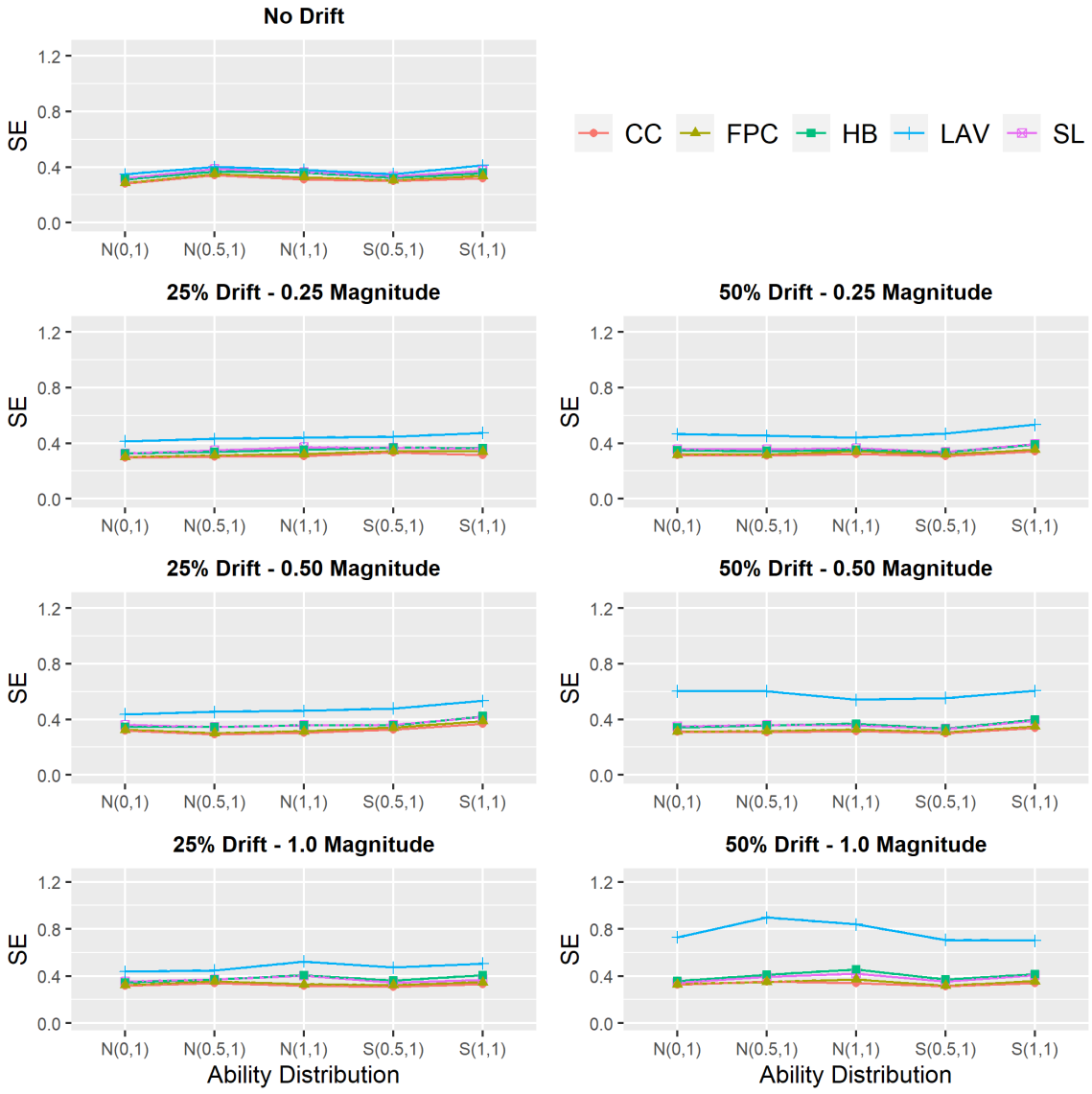

Figure 42. SE Values for True Scores – 3,000 Examinees.

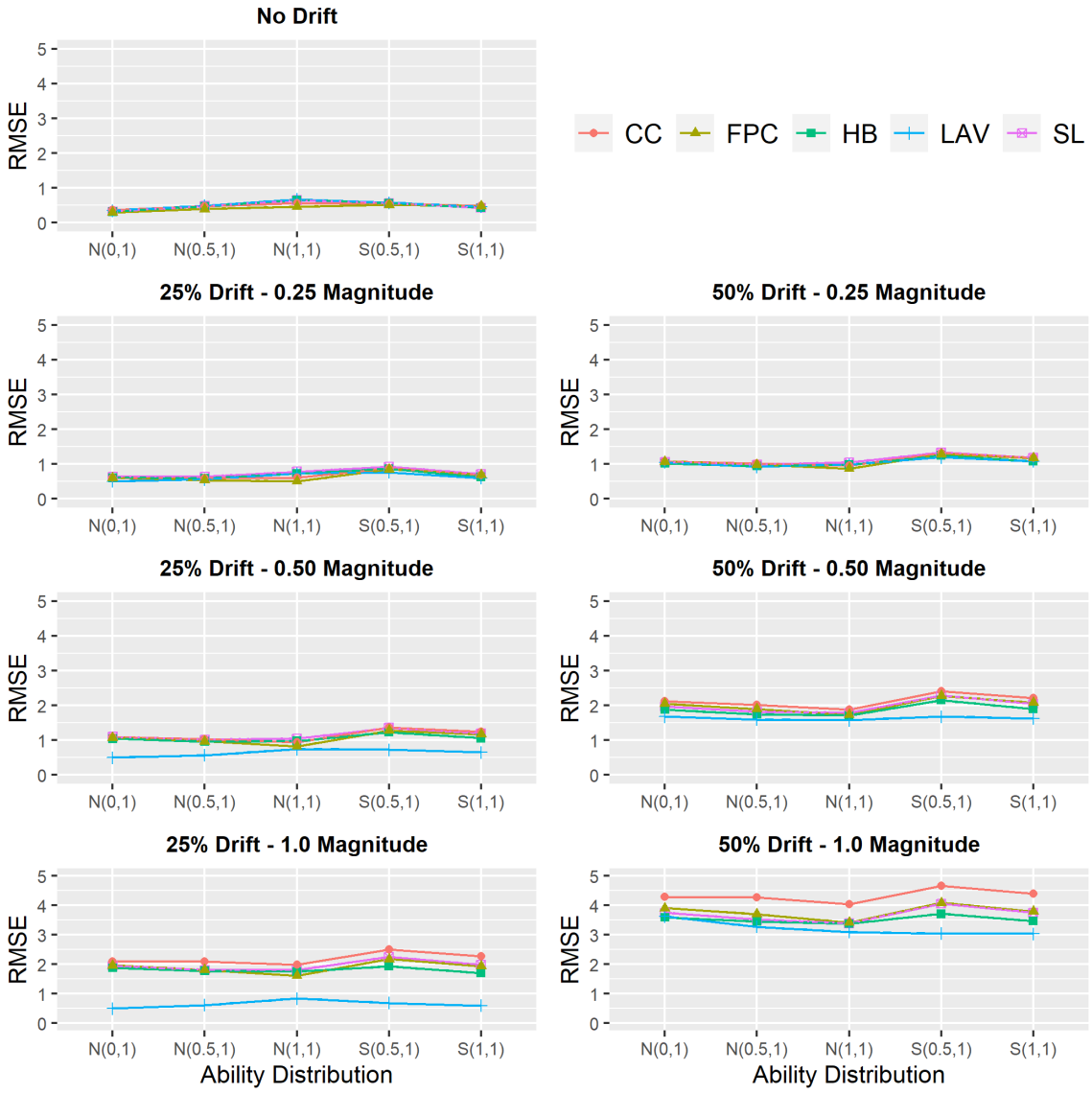

Figure 43. RMSE Values for True Scores – 3,000 Examinees.

Similar to the 1,000 sample-size conditions, RMSE decreased as the mean skewed ability increased from S(0.5, 1) to S(1, 1) for all linking methods and drift conditions including no drift. RMSE decreased as the normal ability distributions increased, but this pattern only consistently occurred for the most extreme condition of drift.

Conditional RMSE values were plotted for each linking method to identify which points along the scale produced the highest RMSE. RMSE values are provided for scores ranging between 25 and 100 because linear interpolation was not used to obtain scores below the sum of the pseudo-guessing parameters for the simulation. Figures 44 – 48 illustrate the conditional RMSE for the 3,000 sample-size of the SL, HB, LAV, CC, and FPC methods, respectively.

For all conditions (top left panels) of each linking method, RMSE tended to be higher at the lower and higher ends of the scale. This is particularly true for the N(1,1) and S(0.5, 1) conditions. The S(0.5, 1) condition (blue line) is noticeably higher for the 3,000 sample-size conditions compared to the 1,000 sample-size conditions for all linking methods. This helps to explain why RMSE was higher for S(0.5, 1) than for S(1, 1). The pattern of findings for the remaining drift conditions is similar to that found with the 1,000 sample-size conditions; however, the RMSE for S(0.5, 1) is also elevated for each drift condition throughout most points of the scale.

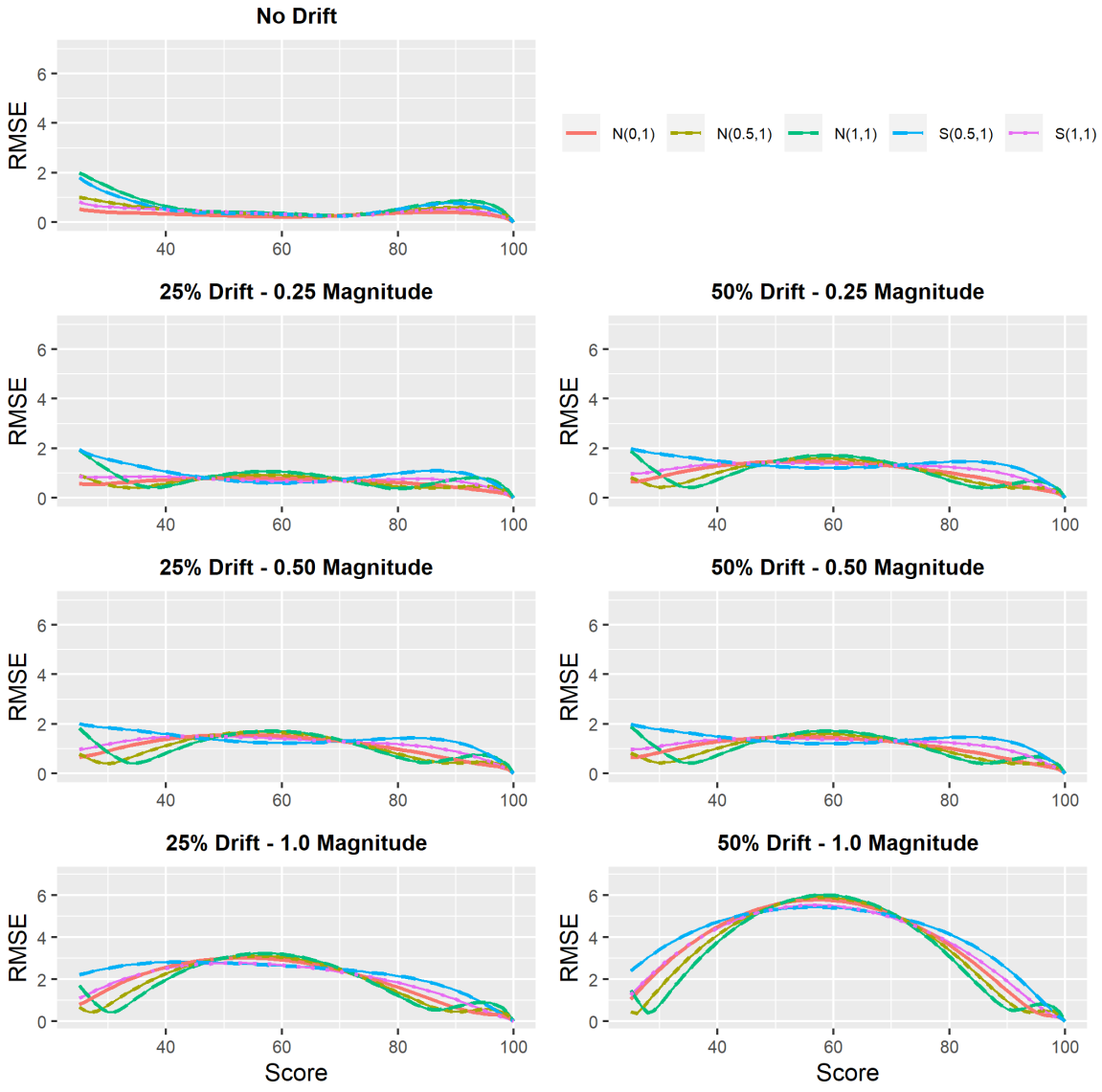

Figure 44. Conditional RMSE for SL True Scores – 3,000 Examinees.

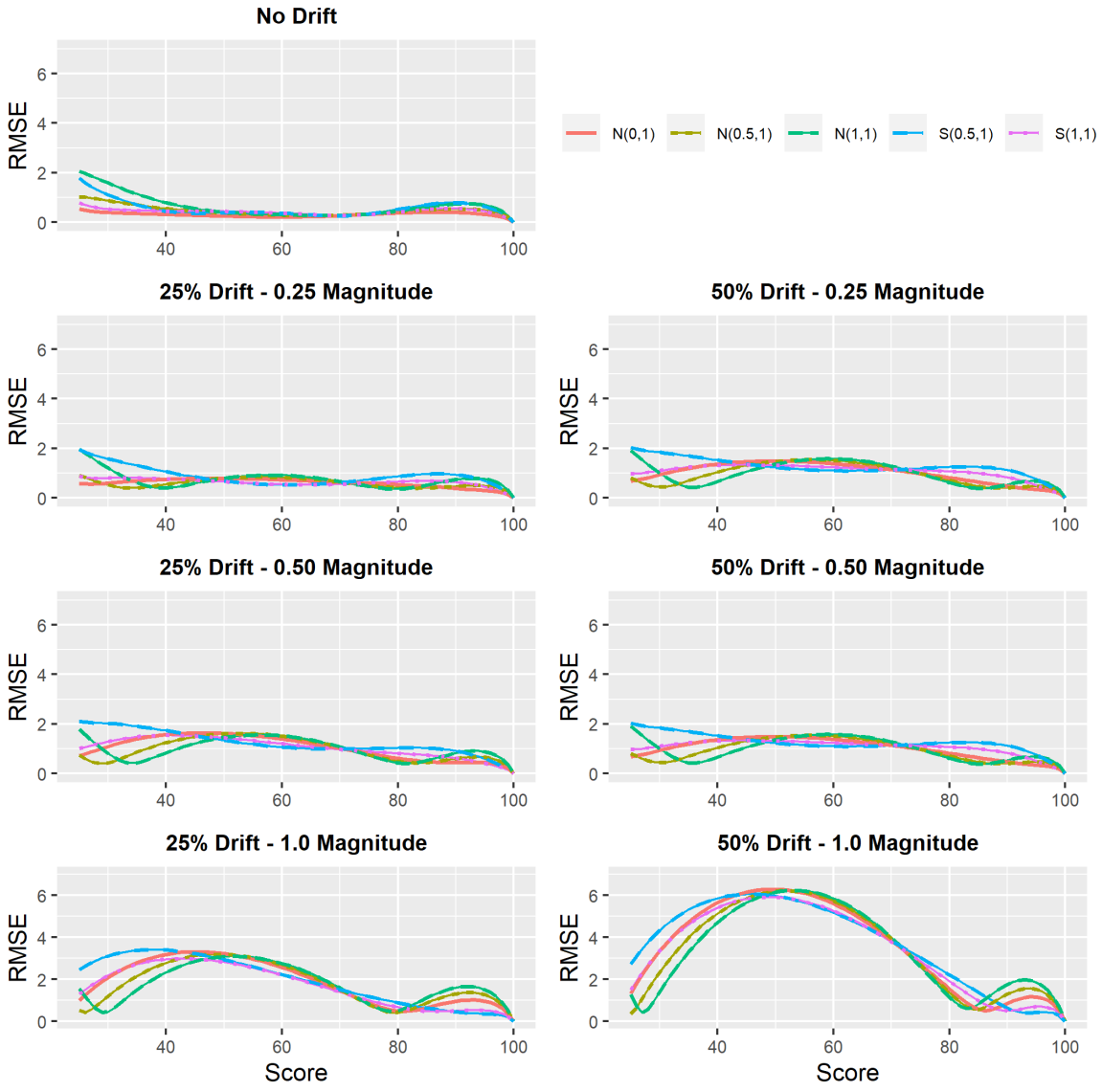

Figure 45. Conditional RMSE for HB True Scores – 3,000 Examinees.

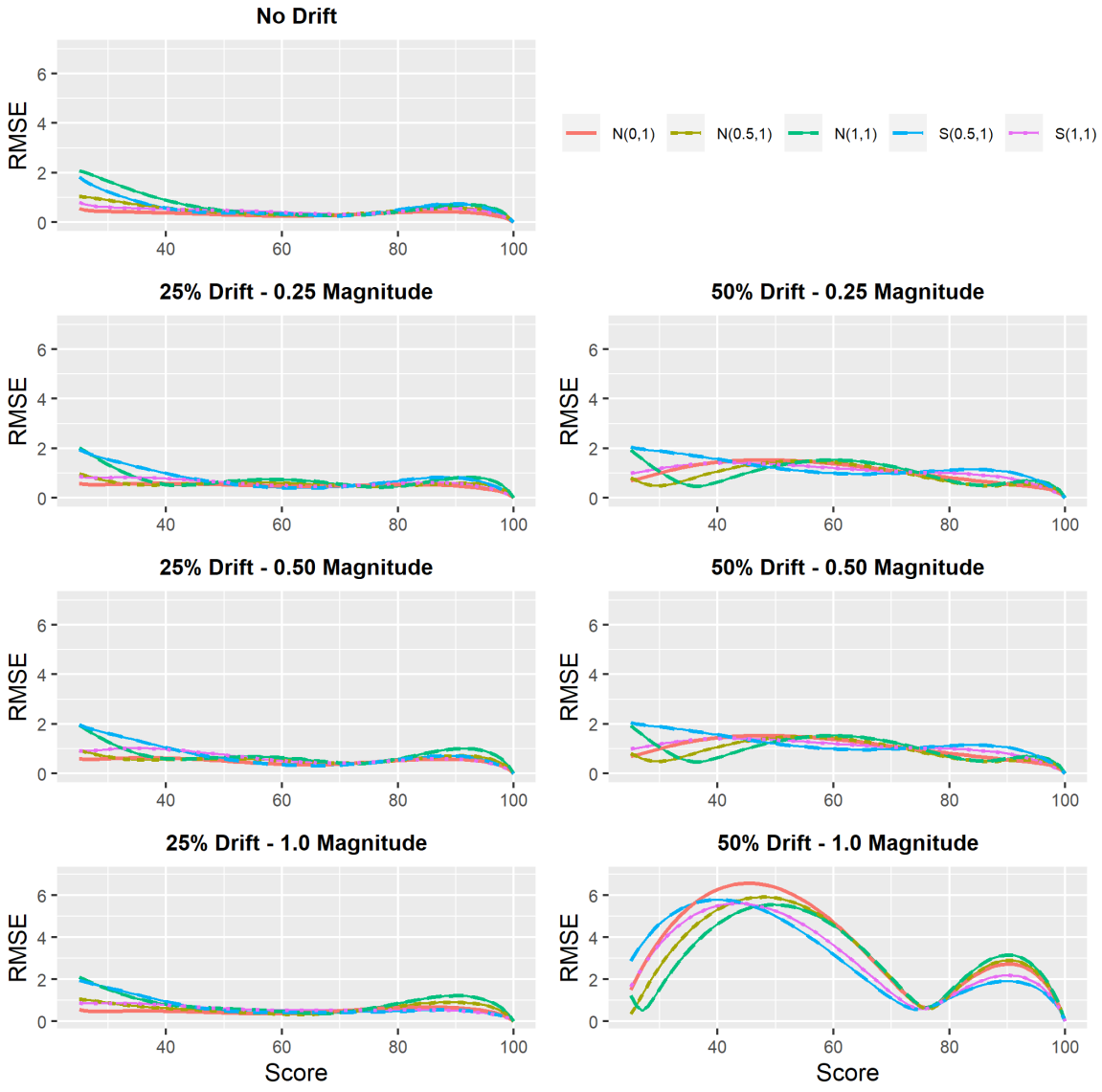

Figure 46. Conditional RMSE for LAV True Scores – 3,000 Examinees.

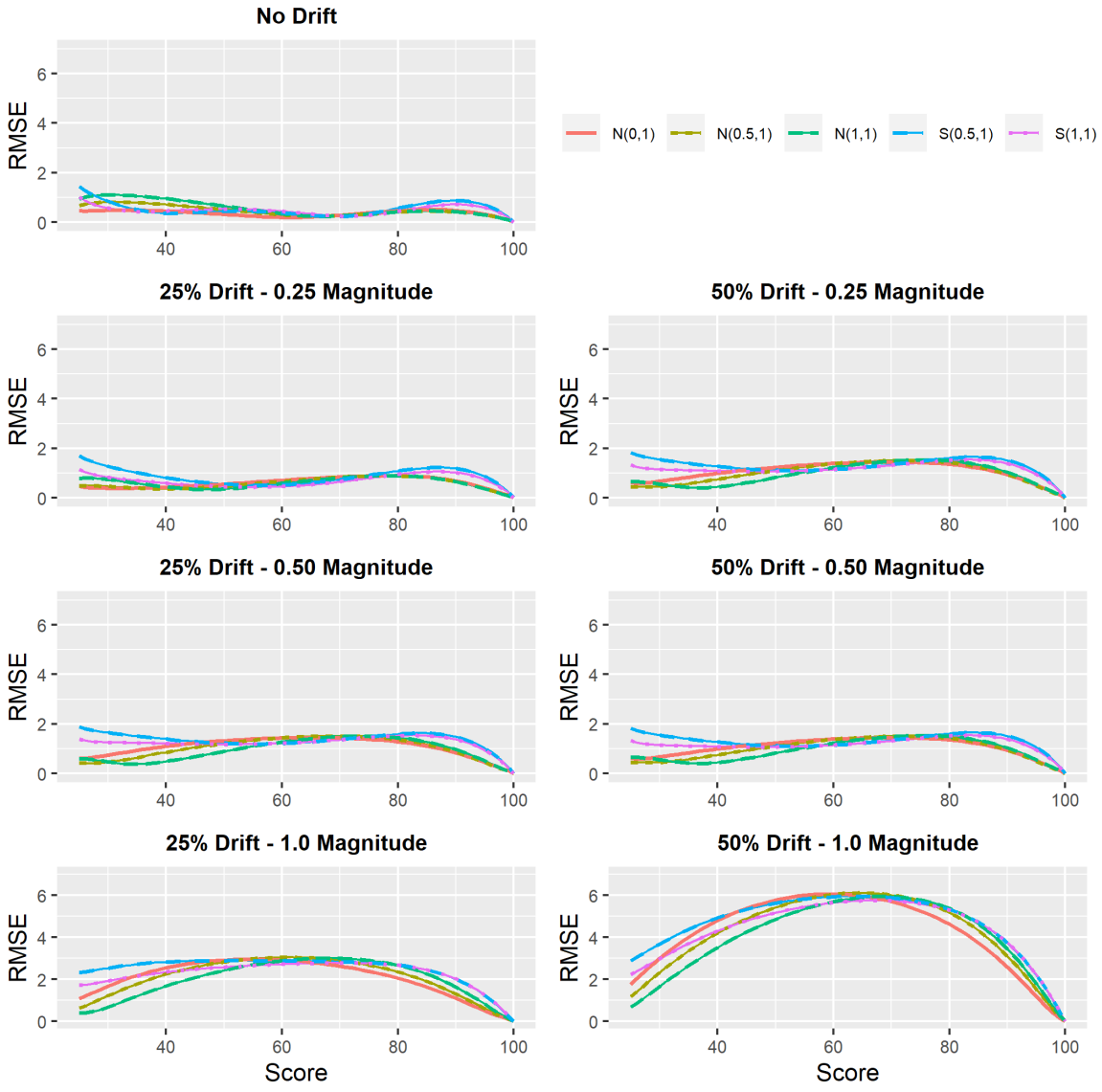

Figure 47. Conditional RMSE for CC True Scores – 3,000 Examinees.

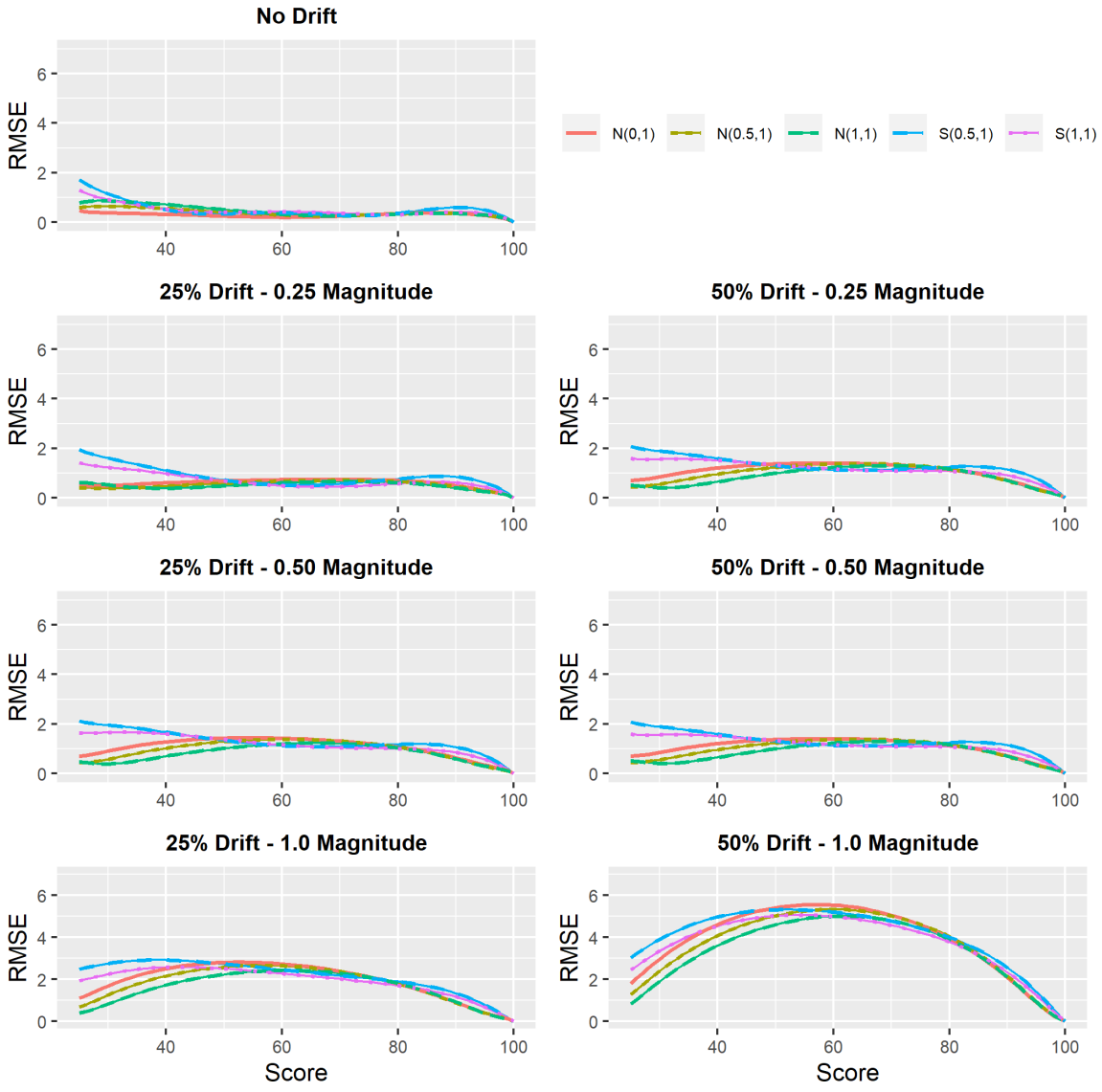

Figure 48. Conditional RMSE for FPC True Scores – 3,000 Examinees.

Equated Scores with IRT Observed Score Equating. Bias, SE, and RMSE were calculated by comparing the estimated scores using IRT observed score equating to the criterion equating relationship. The criterion equating relationship used the equated scores obtained from the generating item parameters for the baseline condition. Figures 49 – 51 illustrate the average bias, SE, and RMSE values for equated scores under IRT observed score equating for 1,000 examinees. Figures 52 – 56 plot the conditional RMSE values for 1,000 examinees. Figures 57 – 59 illustrate average bias, SE, and RMSE values, while Figures 60 – 64 plot the conditional RMSE values for 3,000 examinees. Specific values for each of these three outcomes can be found in Appendix E.

Overall, observed scores followed the same pattern witnessed with IRT true score equating. The LAV and FPC methods produced the lowest RMSE values for most conditions. Findings for all conditions are presented below.

For the 1,000 sample-size conditions, when no drift was present, RMSE exceeded the DTM for nearly all linking methods and conditions. Bias values mostly remained under 0.5 for all linking methods except for the N(1,1) distribution. SE values were larger than bias values for the separate calibration methods and FPC, and similar to bias values for CC. RMSE and bias increased as the mean ability increased under the normal distributions for all linking methods. As the mean ability increased under the skewed distributions, RMSE and bias increased for HB, LAV, and CC, but decreased for SL and FPC. The separate calibration methods returned the smallest amount of RMSE and bias for all ability distributions except for N(1,1), which belonged to FPC.

When 25% of items drifted, RMSE exceeded the DTM for all linking methods and conditions (this was also true for 50% items drifted). There was no systematic pattern of RMSE or bias for any of the linking methods as ability increased for the normal distributions. All linking methods displayed a decrease in RMSE and bias as ability increased for the skewed distributions. As the magnitude of drift increased, RMSE and bias increased for all linking methods except for LAV. The LAV method showed decreases in RMSE and bias for the N(0,1), N(0.5,1), and S(0.5,1) as the magnitude of drift increased from -0.50 to -1.00. The LAV yielded the smallest amounts of RMSE and bias for nearly all conditions. FPC performed better than LAV under N(0.5,1) and N(1,1) when drift magnitude was -0.25, and under N(1,1) with a drift magnitude of -0.50. These findings are consistent with those found for equated true scores.

When 50% of items drifted, RMSE and bias tended to decrease as ability increased for the normal distributions. RMSE and bias decreased for all linking methods when ability increased for the skewed distributions. As drift magnitude increased, RMSE and bias increased for all linking methods. Under the -0.25-drift magnitude, all linking methods performed similarly. When drift magnitude increased to -0.50 and -1.00, the LAV produced values of RMSE bias smaller than other linking methods for most conditions. Performance of the LAV is consistent with those observed for equated true scores.

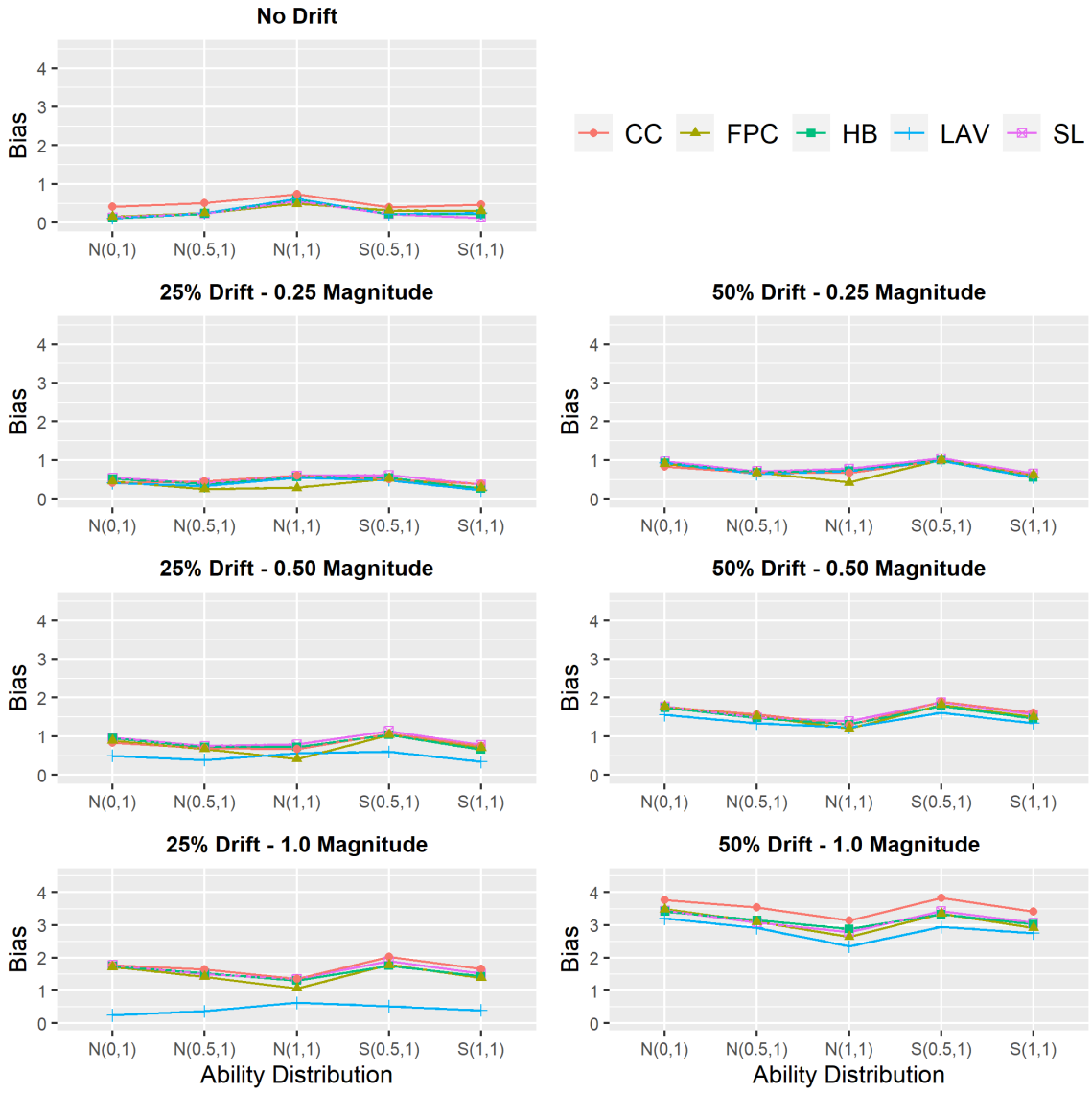

Figure 49. Bias Values for Observed Scores – 1,000 Examinees.

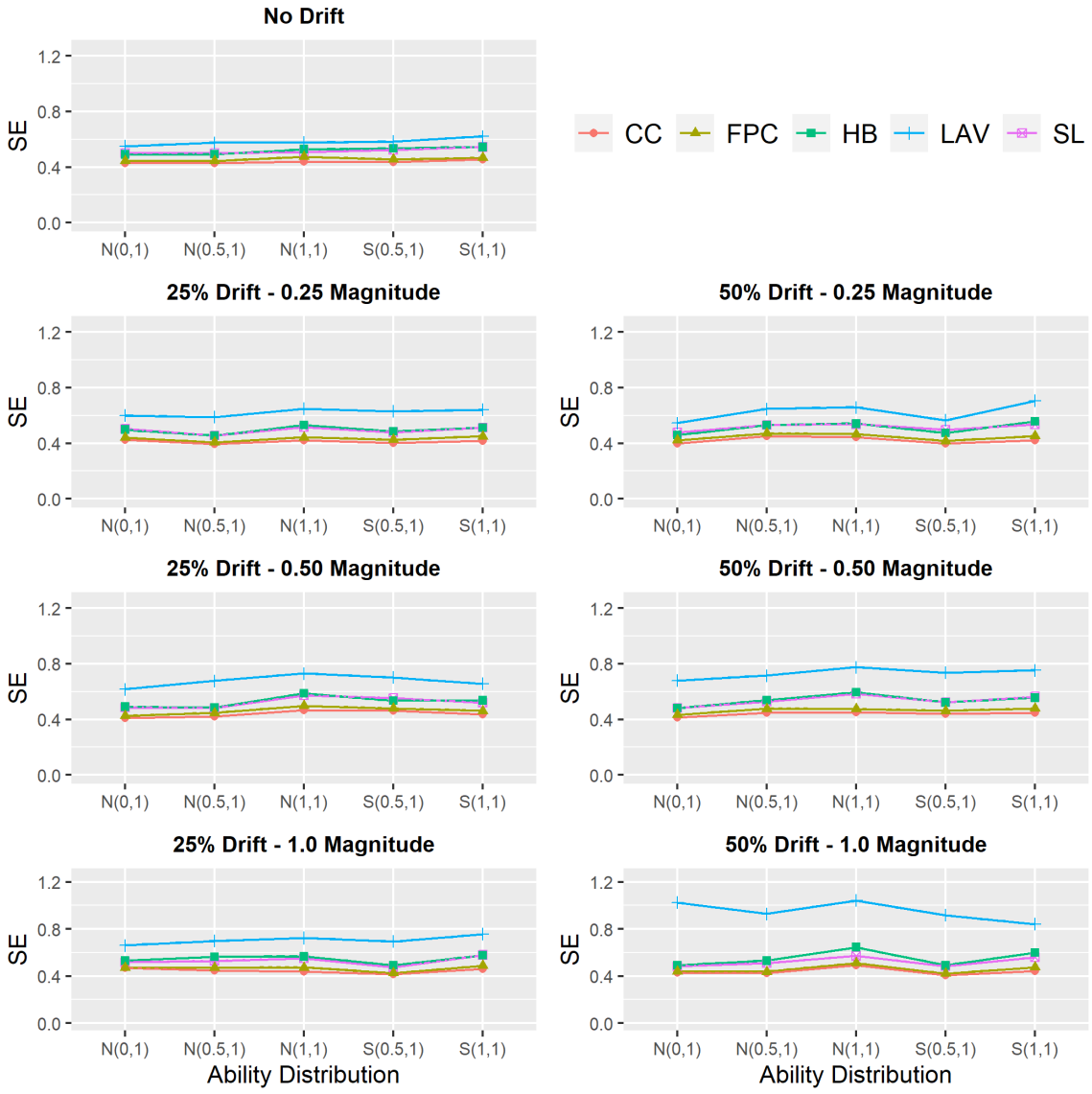

Figure 50. SE Values for Observed Scores – 1,000 Examinees.

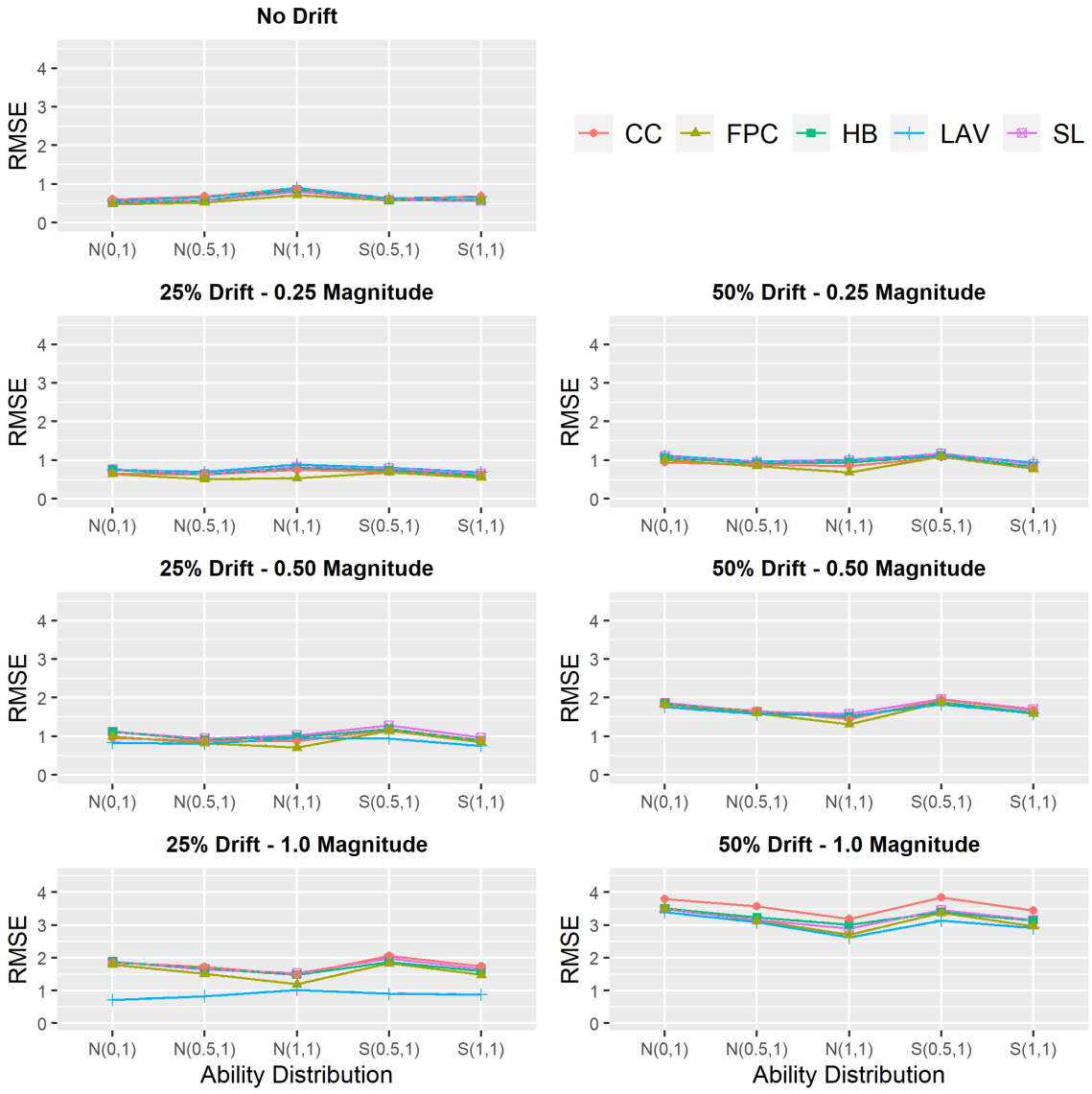

Figure 51. RMSE Values for Observed Scores – 1,000 Examinees.

Similar to equated true scores, RMSE from observed scores decreased as the mean skewed ability increased from S(0.5, 1) to S(1, 1) for all drift conditions except the baseline (no drift). RMSE decreased as the normal ability distributions increased, but this pattern only occurred under the most extreme condition of drift. Under the other conditions of drift, RMSE decreased selectively by each linking method.

Conditional RMSE values were plotted for each linking method to identify which points along the scale produced the highest RMSE. Figures 52 – 56 illustrate the conditional RMSE for equated observed scores under the 1,000 sample-size for the SL, HB, LAV, CC, and FPC methods, respectively. These figures are nearly identical to the conditional RMSE plots for equated true scores. RMSE tended to be higher at the lower and higher ends of the scale for the baseline conditions. RMSE was distributed rather evenly in the middle of the scale as drift increased. The 50% drifted -1.0 magnitude condition resulted in elevated RMSE values for all linking methods, although it was distributed in different parts of the scale. RMSE was greatest in the middle of the scale for SL, HB, and LAV. The LAV and HB methods also had small spikes of RMSE at the higher end of the scale. FPC and CC exhibited the largest RMSE values towards the higher end of the scale.

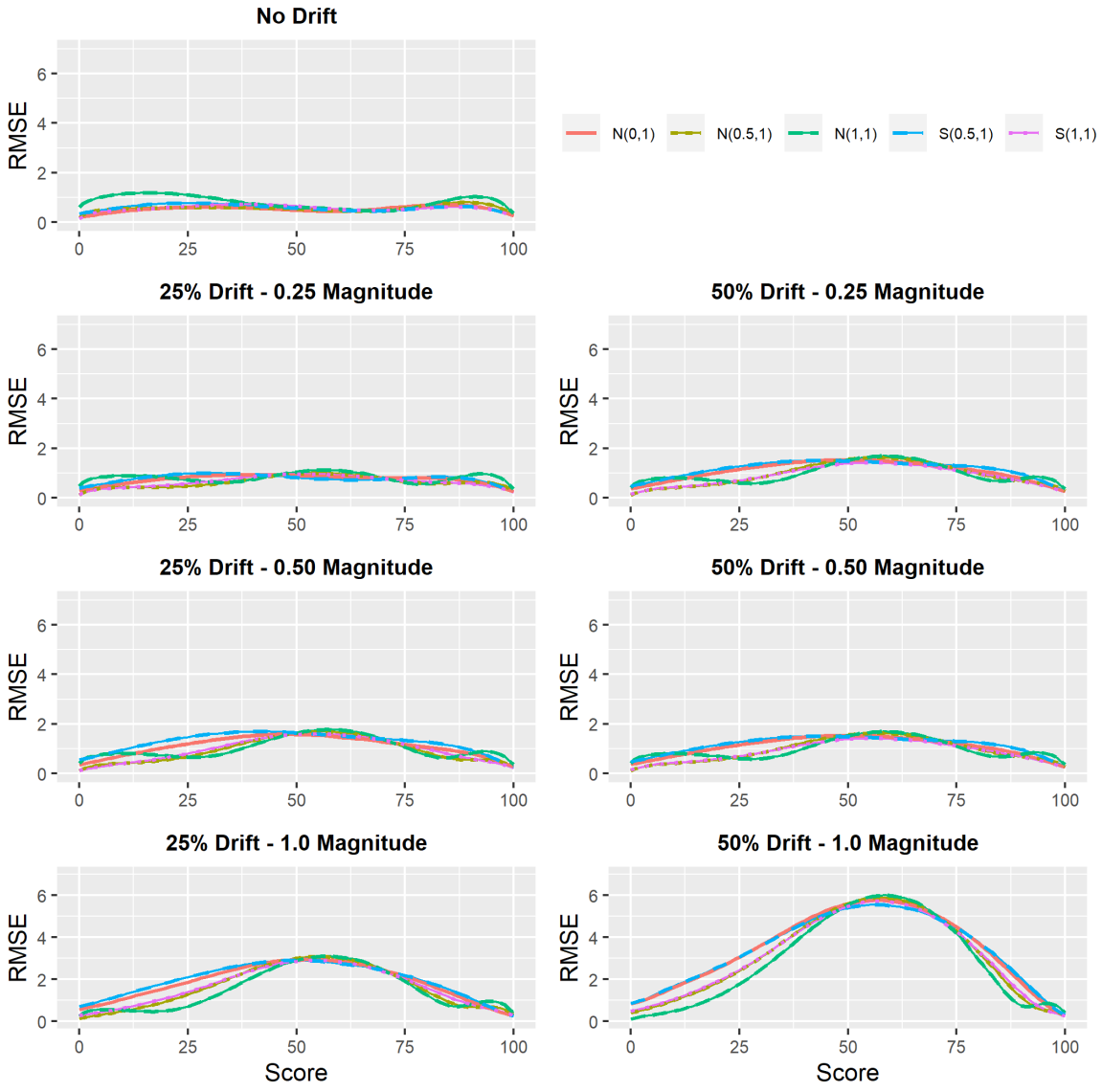

Figure 52. Conditional RMSE for SL Observed Scores – 1,000 Examinees.

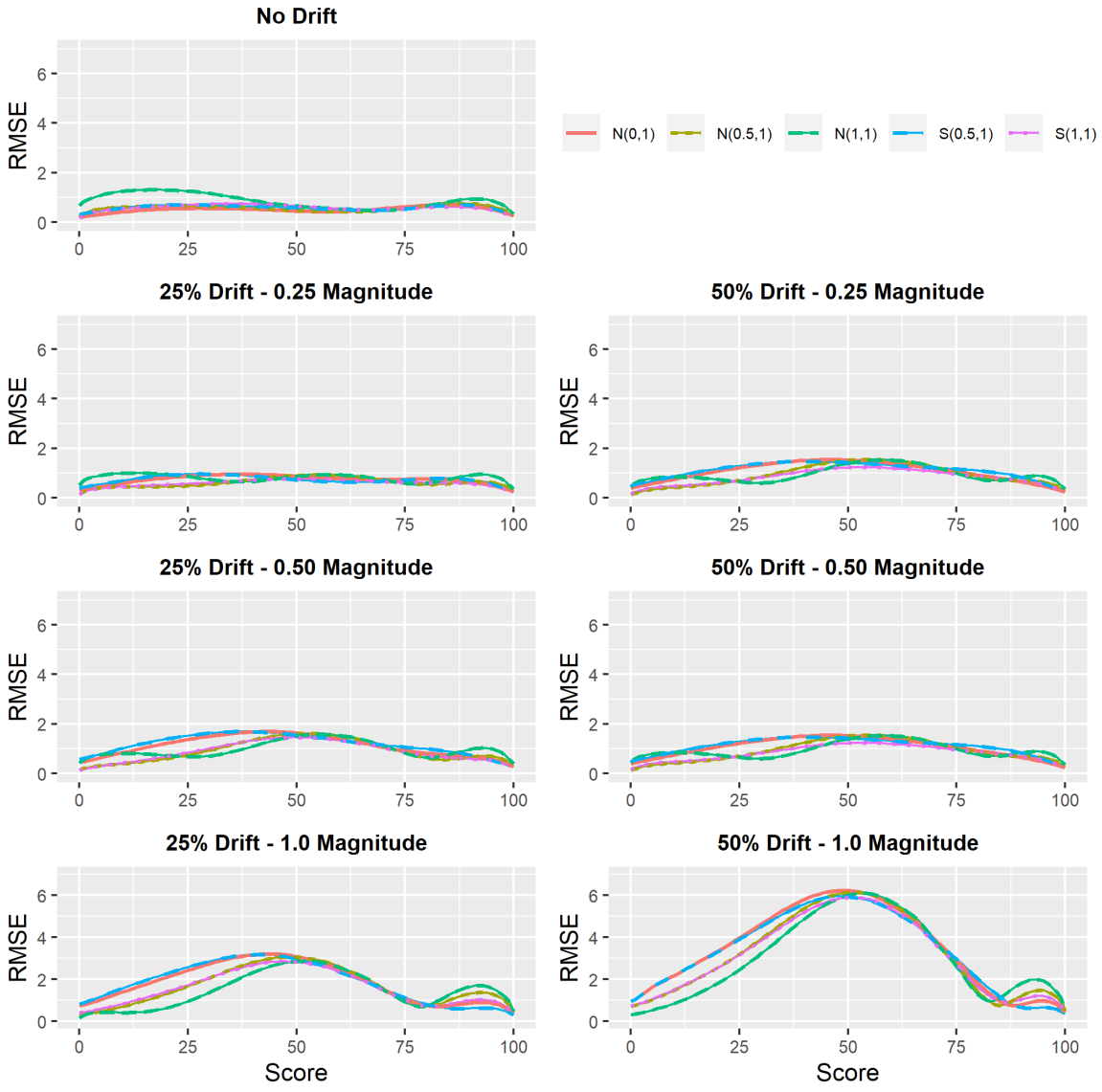

Figure 53. Conditional RMSE for HB Observed Scores – 1,000 Examinees.

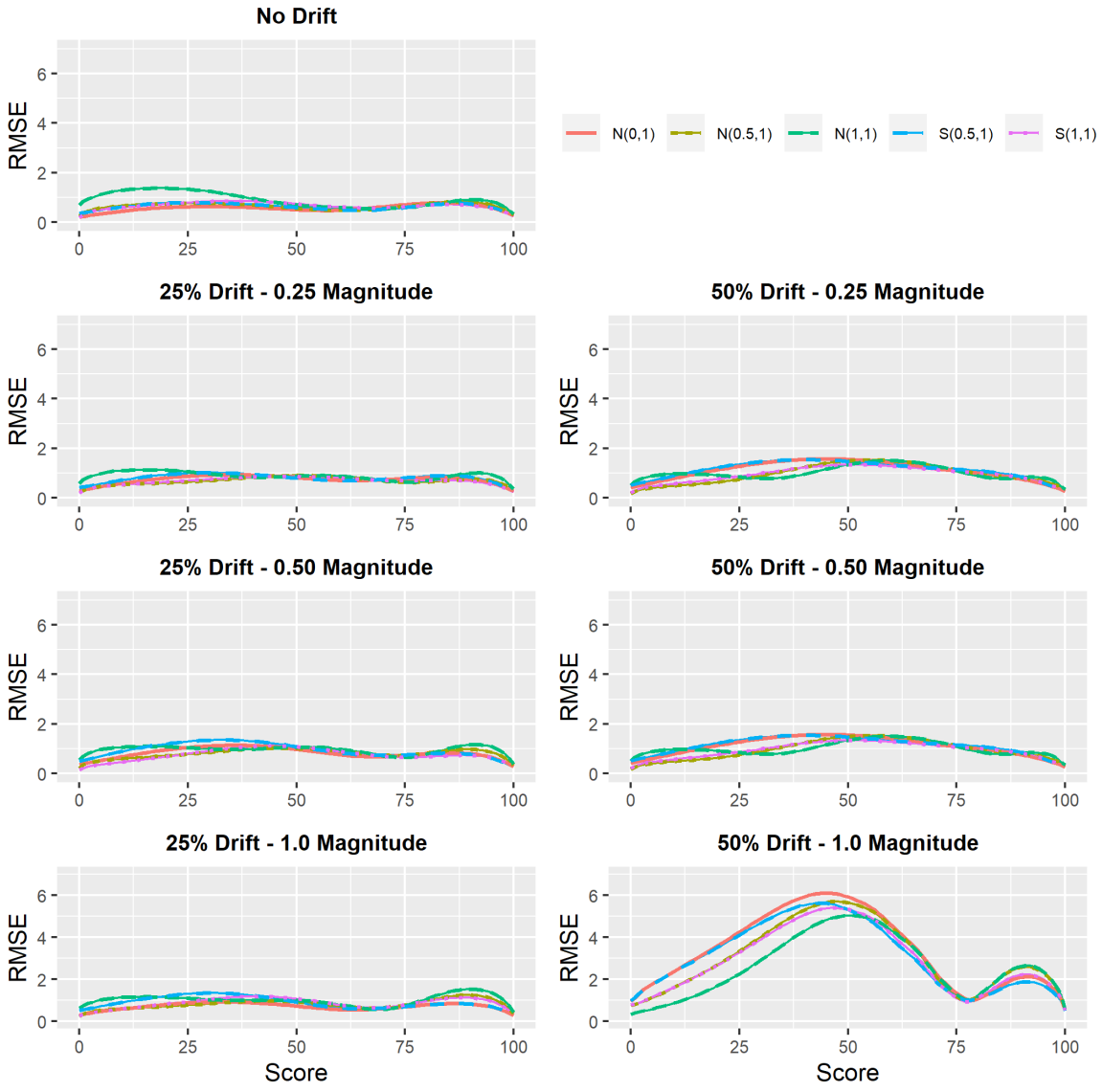

Figure 54. Conditional RMSE for LAV Observed Scores – 1,000 Examinees.

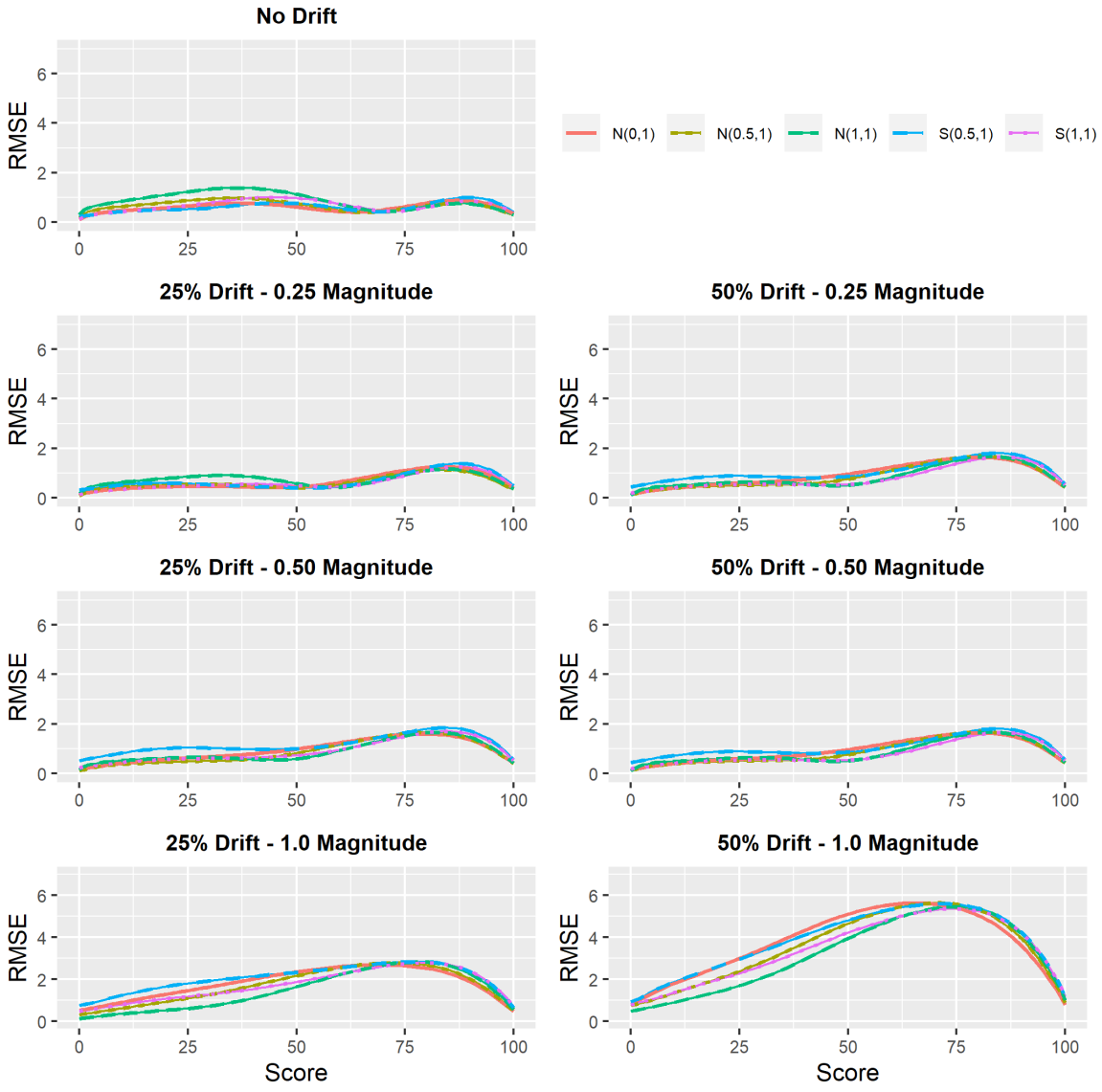

Figure 55. Conditional RMSE for CC Observed Scores – 1,000 Examinees.

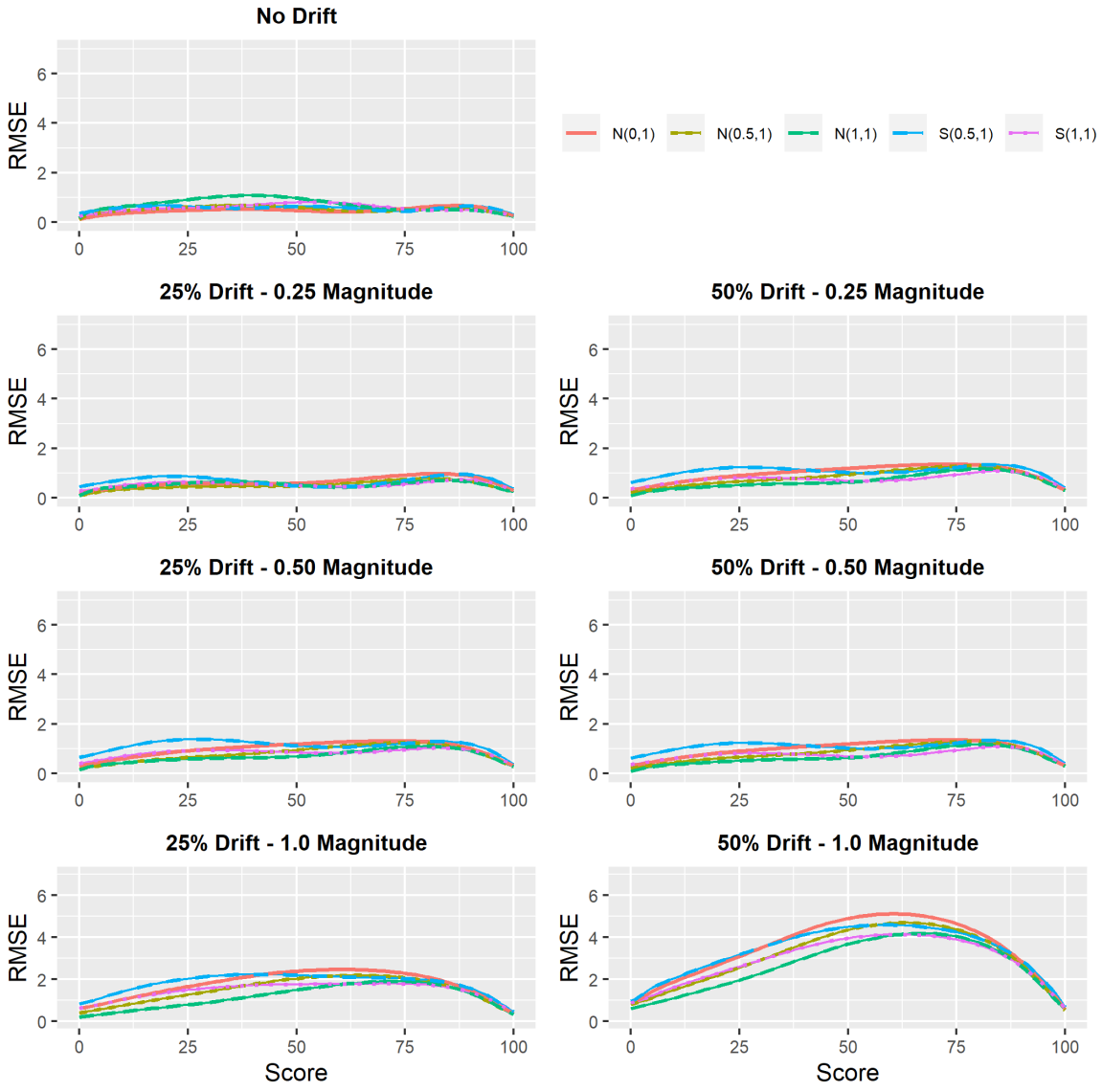

Figure 56. Conditional RMSE for FPC Observed Scores – 1,000 Examinees.

Compared to the 1,000 sample-size conditions, RMSE and bias were smaller under lower magnitudes of drift (none or 25% drifted items), but higher under the highest magnitudes of drift (50% drifted items, -0.50 and -1.00 magnitude) SE decreased for all linking methods and conditions. RMSE exceeded DTM for all linking methods under N(1,1) and S(0.5,1), with the exception of FPC, which had an RMSE of .487 for N(1,1). Interestingly, FPC exceeded the DTM for S(1,1), whereas the other linking methods were below the DTM.

When 25% of items drifted, RMSE exceeded the DTM threshold for most conditions and linking methods (also true for 50% of items drifted). RMSE and bias decreased as the mean ability of the skewed distributions increased for all linking methods. This was only true sometimes for the normal distributions. As drift magnitude increased, RMSE and bias also increased for all linking methods except for LAV conditions N(0,1) and the skewed distributions. All linking methods had similar values of RMSE and similar values of bias when drift magnitude was -0.25. When drift magnitude increased to -0.50 and -1.00, LAV yielded the smallest amounts of RMSE and bias.

When 50% of items drifted, RMSE and bias tended to decrease as candidate mean ability increased for both normal and skewed distributions. As drift magnitude increased, RMSE and bias also increased for all linking methods. At -0.25 drift magnitude, all linking methods performed similarly. At -0.50 and -1.00 magnitudes of drift, the LAV method produced the smallest amounts of RMSE and bias. CC was most influenced at the highest magnitudes of drift.

Similar to IRT true score equating, the LAV method yielded the highest values of SE for IRT observed score equating under all conditions and sample sizes. This also affected the performance of the LAV when evaluating RMSE. For the 1,000-candidate condition, when the magnitude of drift was -1.00 under 25% and 50% items drifted, the LAV still recovered equated observed scores the best. Under the remaining drift conditions, all methods performed similarly. The FPC method produced slightly smaller RMSE values than the other linking methods, particularly under N(1,1). When the sample size increased to 3,000, LAV performed the best for conditions where the magnitude of drift was -0.50 and -1.00 and the percentage of items drifted was 25% and 50%. When no drift was present, or when drift magnitude was -0.25, all methods performed similarly.

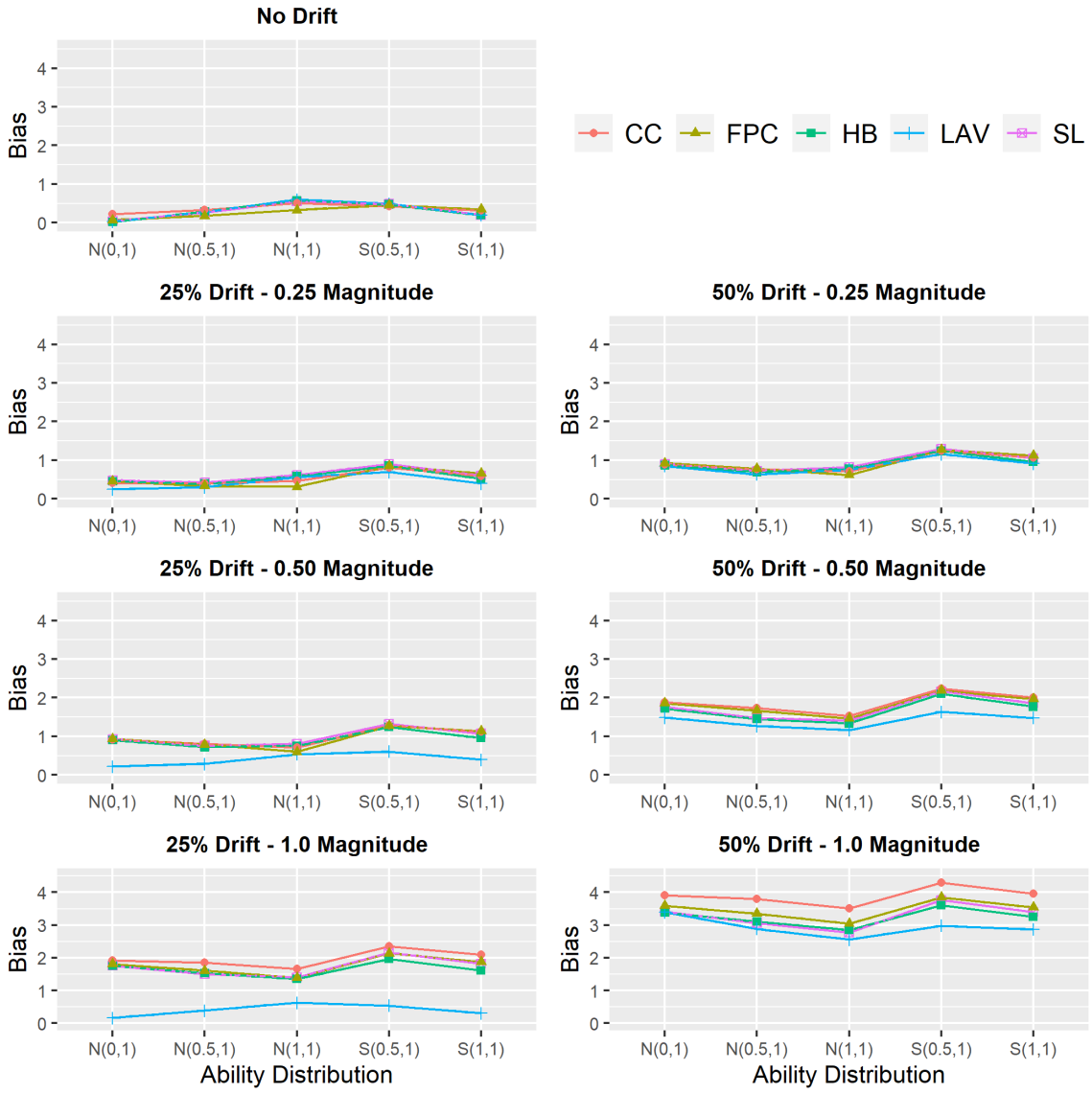

Figure 57. Bias Values for Observed Scores – 3,000 Examinees.

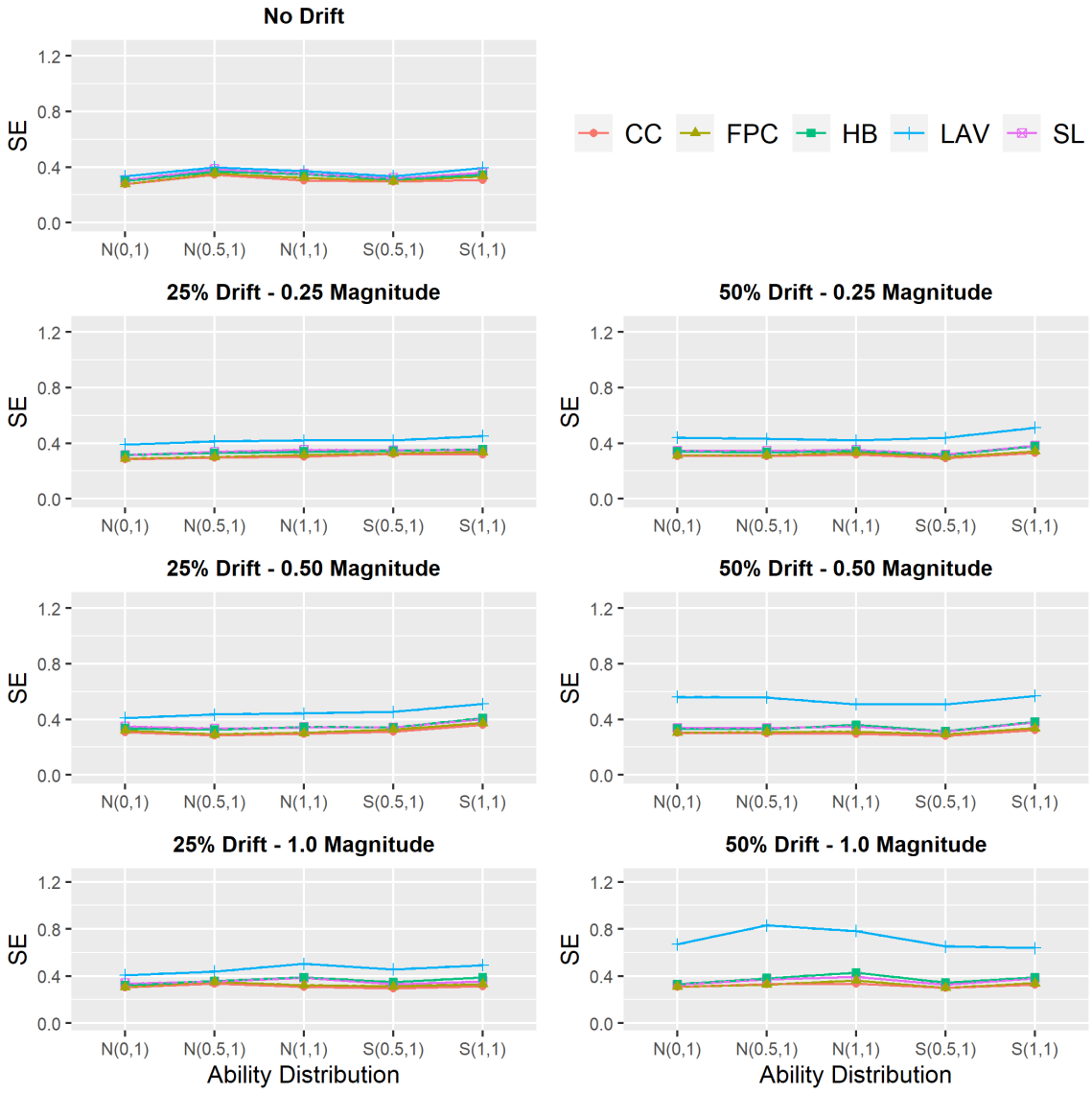

Figure 58. SE Values for Observed Scores – 3,000 Examinees.

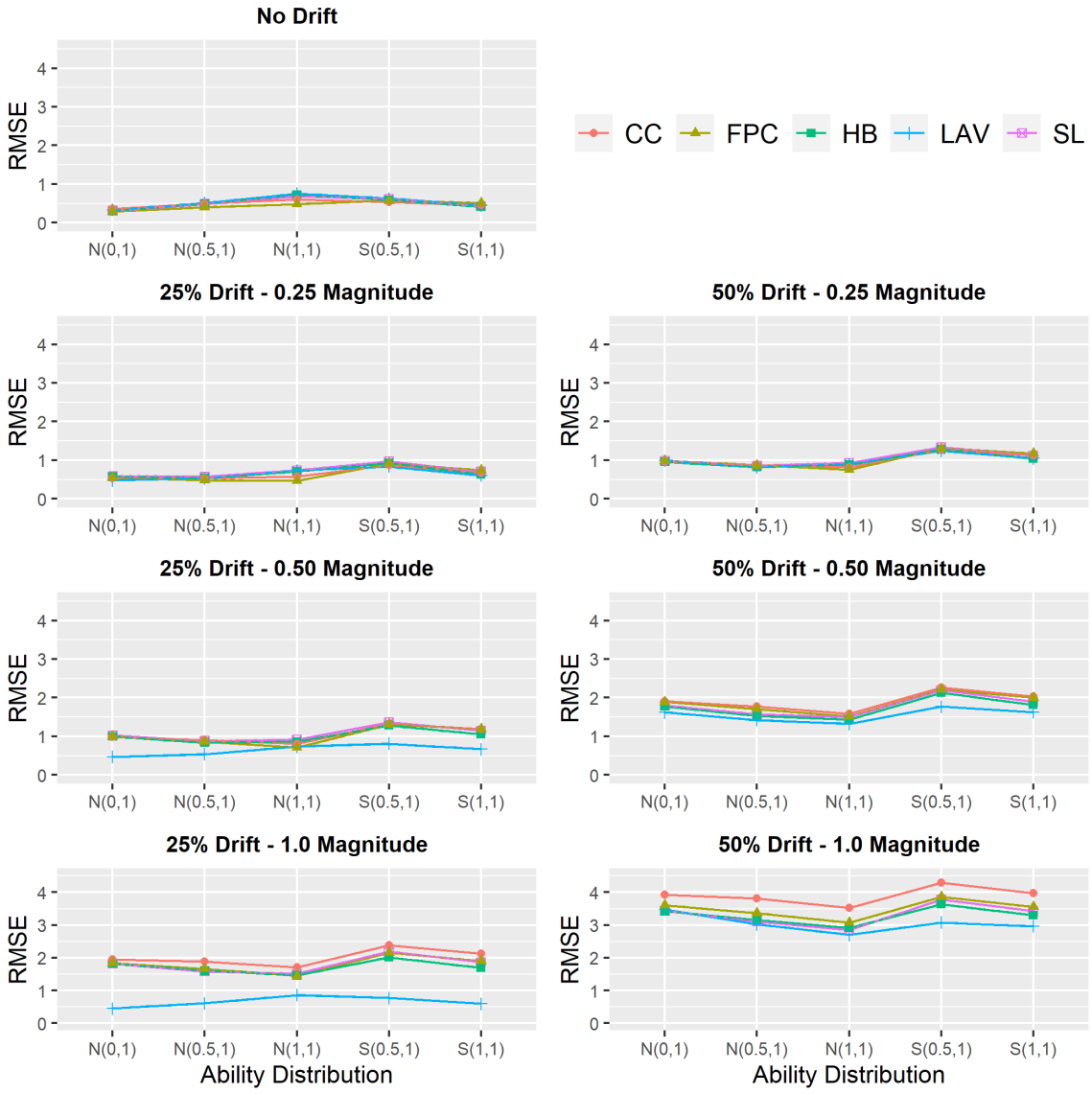

Figure 59. RMSE Values for Observed Scores – 3,000 Examinees.

Conditional RMSE values were plotted for each linking method to identify why RMSE may have decreased as the mean ability of the normal and skewed distributions increased. Figures 60 – 64 illustrate the conditional RMSE for the 3,000 sample-size of the SL, HB, LAV, CC, and FPC methods, respectively. The patterns from these figures are similar to the other conditional RMSE true and observed score plots. The RMSE values for the equated true and observed scores exceed the DTM threshold for the drift and non-drift conditions, which is similar to other studies (e.g., Hu et al., 2008; Jurich et al., 2012). However, using a weighted RMSE would provide better equating results because the RMSE's at the lower and higher ends of the scale would not be emphasized less than the scores in the middle of the scale, where most of the scores are located.

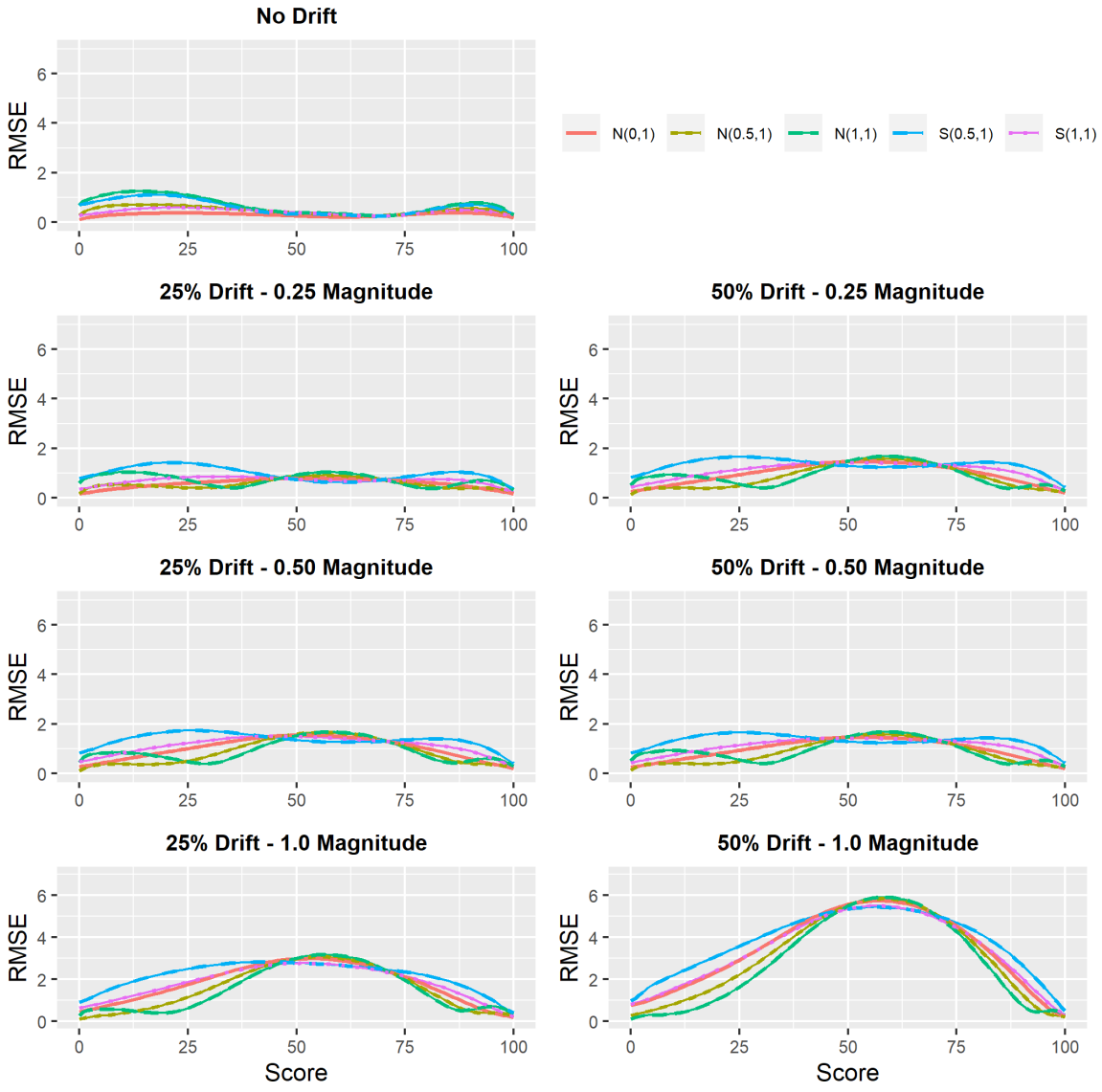

Figure 60. Conditional RMSE for SL Observed Scores – 3,000 Examinees.

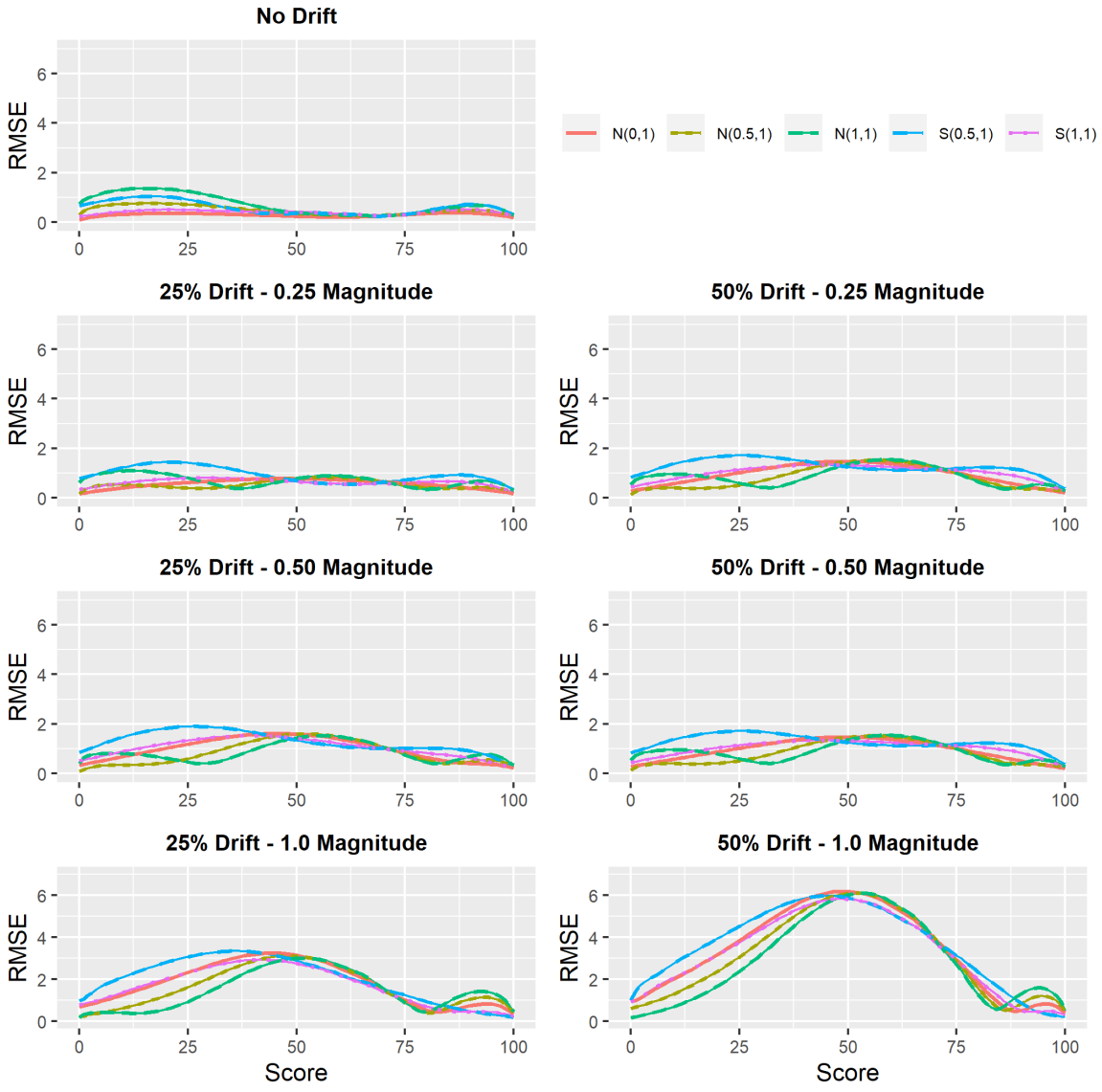

Figure 61. Conditional RMSE for HB Observed Scores – 3,000 Examinees.

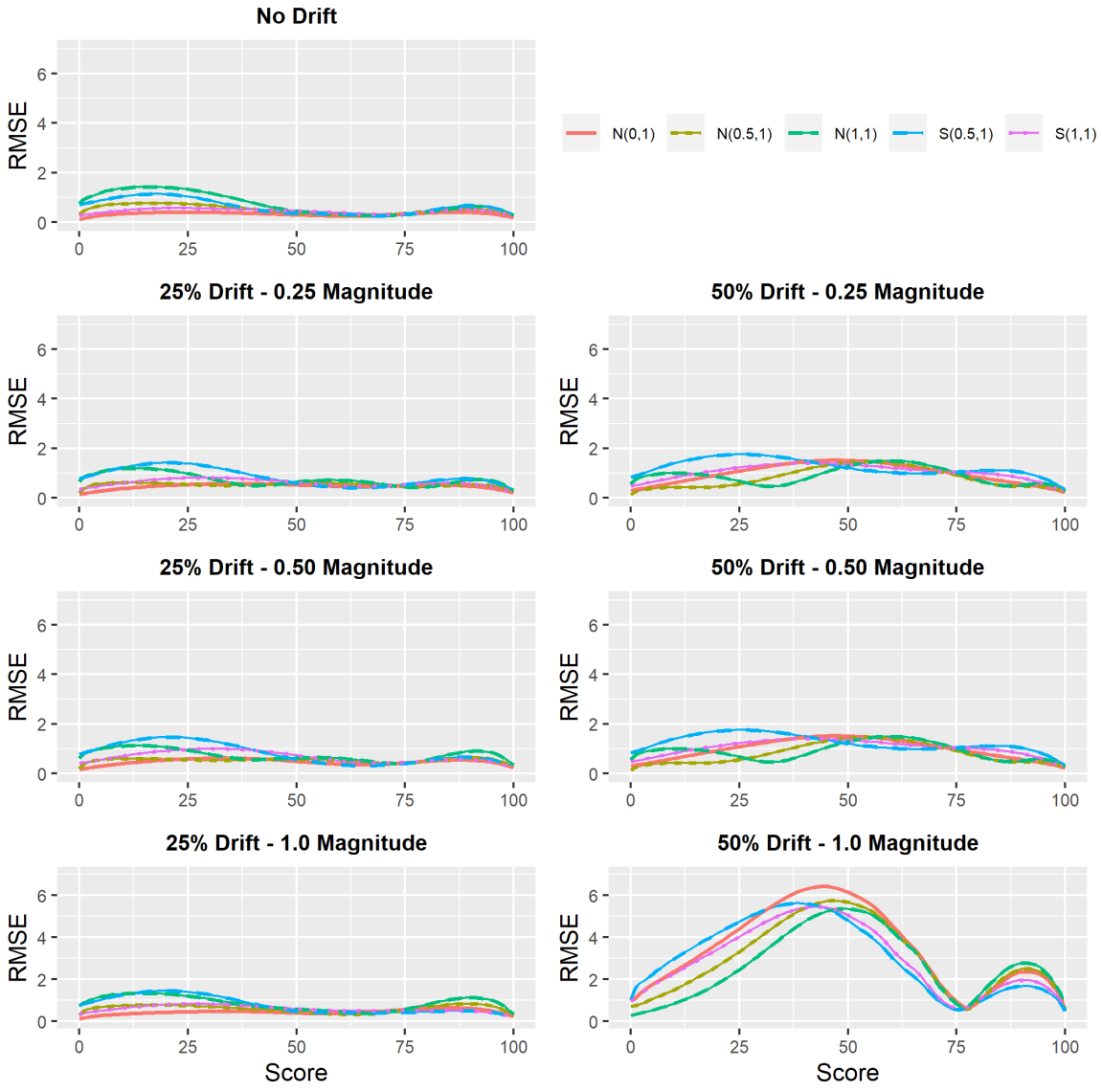

Figure 62. Conditional RMSE for LAV Observed Scores – 3,000 Examinees.

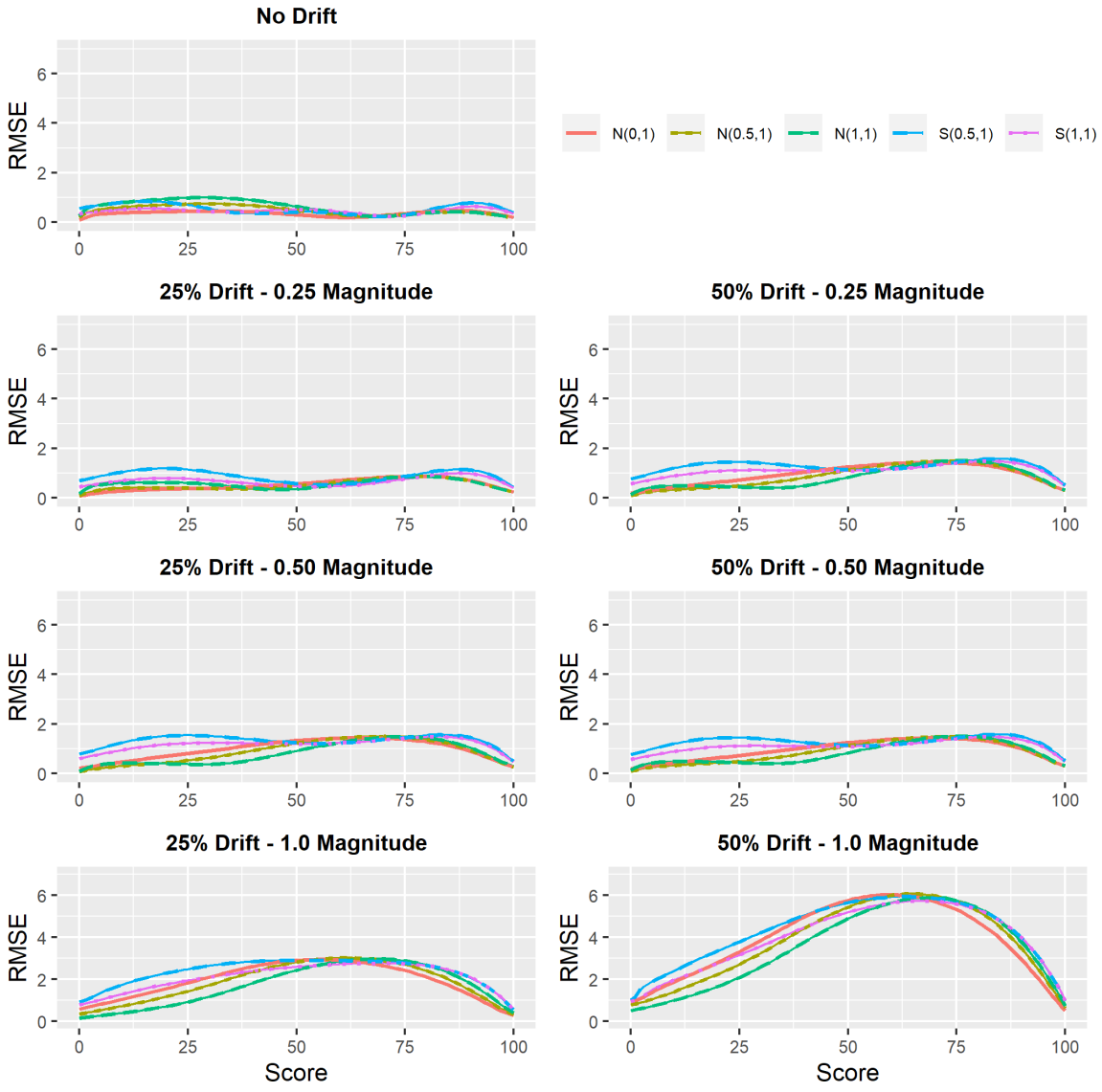

Figure 63. Conditional RMSE for CC Observed Scores – 3,000 Examinees.

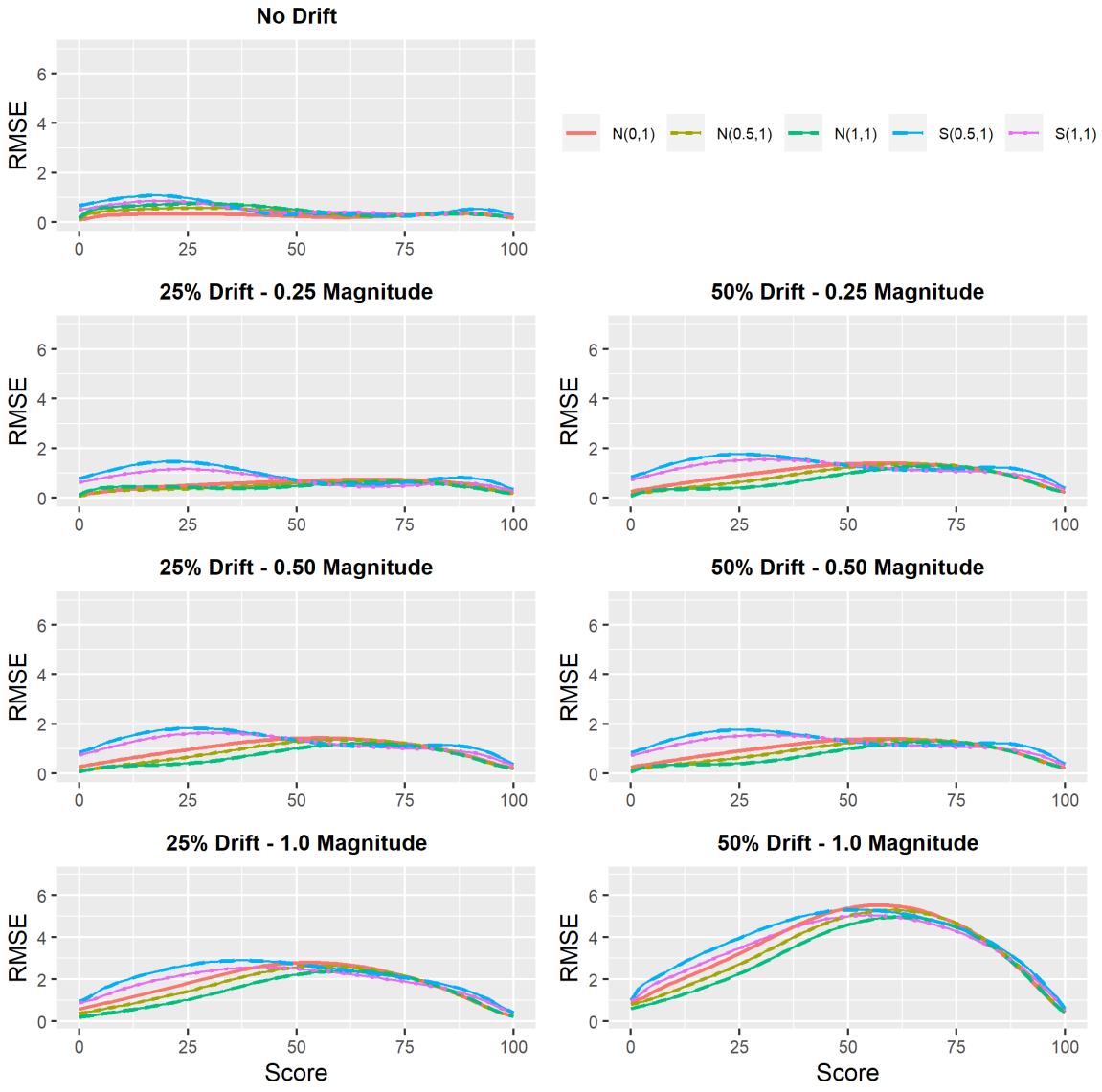

Figure 64. Conditional RMSE for FPC Observed Scores – 3,000 Examinees.

Classification Accuracy. The fourth research question examined the extent to which IPD affects classification accuracy rates. Bias, SE, and RMSE were calculated by comparing the classification accuracy rates using the phi coefficient from each linking method to the true classification criterion, which is defined as the proportion of examinees classified as pass-pass or fail-fail for both the true and observed θ status'. There was a total of five true classification criterion, one for each ability distribution. The five true classification criterion rates, along with the estimated classification accuracy rates can be seen in Tables 10 and 11, for the 1,000 and 3,000 sample-size conditions, respectively. Figures 65 – 67 illustrate the bias, SE, and RMSE values for classification accuracy with 1,000 examinees. Figures 68 – 70 illustrate the bias, SE, and RMSE values for classification accuracy with 3,000 examinees. Specific values for each of these three outcomes can be found in Appendix F.

The success rate of classification accuracy and consistency is partially predicated upon how well the item parameters are recovered. All linking methods performed similarly in their estimation of classification accuracy and consistency. If examinees were concentrated at the lower points of the scale, then classification accuracy and consistency may have decreased further because examinees would move towards the cut score, where decision about pass and fail could fluctuate more.

For both sample sizes, the same general conclusions can be drawn since the classification rates were very similar for all linking methods. Under all conditions of drift, classification rates were slightly underestimated. No discernable pattern for RMSE could be observed as the ability distributions increased. Only under the most extreme

conditions of drift, could some difference in RMSE be observed between the linking methods. For the 25% drifted common items and -1.00 magnitude, the LAV method performed slightly better in terms of RMSE than the other linking methods, which all performed similarly. The smaller RMSE was due to the smaller amounts of bias. This was also true for the 50% drifted common items and -1.00 magnitude of drift condition. These results can be attributed to LAV's accurate recovery of the difficulty parameter for the same conditions. The SE was similar for all linking methods under all conditions of drift. The one exception was in the 50% drifted common items and -1.00 magnitude of drift condition, where the LAV method had slightly higher SE values. However, the LAV performed the best in terms of RMSE despite larger SE due to small bias for the -1.00 magnitude of drift under the 25% and 50% drifted item conditions.

Table 10

Classification Accuracy Rates – 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
True Classification Criterion			**0.872**	**0.897**	**0.934**	**0.899**	**0.934**
SL	None	None	0.869	0.894	0.932	0.894	0.932
	25%	-0.25	0.868	0.892	0.929	0.893	0.930
		-0.50	0.866	0.888	0.926	0.889	0.927
		-1.00	0.857	0.878	0.917	0.880	0.919
	50%	-0.25	0.866	0.888	0.926	0.889	0.927
		-0.50	0.856	0.877	0.916	0.879	0.918
		-1.00	0.823	0.842	0.887	0.847	0.891
HB	None	None	0.869	0.894	0.932	0.894	0.933
	25%	-0.25	0.868	0.892	0.929	0.893	0.931
		-0.50	0.867	0.889	0.927	0.890	0.928
		-1.00	0.863	0.883	0.922	0.886	0.924
	50%	-0.25	0.866	0.889	0.927	0.890	0.928
		-0.50	0.859	0.879	0.918	0.882	0.920
		-1.00	0.832	0.852	0.895	0.856	0.899
LAV	None	None	0.869	0.894	0.932	0.894	0.932
	25%	-0.25	0.869	0.892	0.930	0.893	0.931
		-0.50	0.869	0.893	0.930	0.893	0.931
		-1.00	0.870	0.894	0.931	0.895	0.932
	50%	-0.25	0.866	0.889	0.927	0.890	0.928
		-0.50	0.862	0.882	0.920	0.884	0.922
		-1.00	0.847	0.869	0.911	0.875	0.914
CC	None	None	0.865	0.892	0.932	0.893	0.932
	25%	-0.25	0.865	0.891	0.930	0.892	0.931
		-0.50	0.864	0.889	0.928	0.889	0.929
		-1.00	0.858	0.881	0.922	0.881	0.922
	50%	-0.25	0.864	0.889	0.929	0.890	0.929
		-0.50	0.858	0.881	0.923	0.882	0.923
		-1.00	0.828	0.849	0.899	0.849	0.899
FPC	None	None	0.868	0.894	0.933	0.895	0.933
	25%	-0.25	0.867	0.892	0.931	0.893	0.932
		-0.50	0.866	0.890	0.929	0.891	0.930
		-1.00	0.859	0.883	0.925	0.884	0.925
	50%	-0.25	0.865	0.890	0.929	0.891	0.930
		-0.50	0.857	0.882	0.923	0.883	0.924
		-1.00	0.830	0.856	0.906	0.858	0.906

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Table 11

Classification Accuracy Rates – 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
True Classification Criterion			**0.872**	**0.897**	**0.934**	**0.899**	**0.934**
SL	None	None	0.871	0.895	0.933	0.896	0.934
	25%	-0.25	0.870	0.893	0.930	0.894	0.931
		-0.50	0.867	0.889	0.927	0.891	0.928
		-1.00	0.858	0.879	0.918	0.882	0.921
	50%	-0.25	0.867	0.889	0.927	0.891	0.928
		-0.50	0.857	0.878	0.918	0.881	0.920
		-1.00	0.825	0.844	0.890	0.850	0.895
HB	None	None	0.871	0.895	0.933	0.896	0.934
	25%	-0.25	0.870	0.893	0.930	0.895	0.932
		-0.50	0.869	0.891	0.928	0.893	0.930
		-1.00	0.864	0.885	0.923	0.887	0.926
	50%	-0.25	0.868	0.890	0.927	0.892	0.929
		-0.50	0.860	0.881	0.919	0.884	0.922
		-1.00	0.833	0.853	0.897	0.859	0.902
LAV	None	None	0.871	0.895	0.933	0.896	0.934
	25%	-0.25	0.871	0.894	0.931	0.896	0.933
		-0.50	0.872	0.895	0.932	0.896	0.933
		-1.00	0.872	0.896	0.933	0.897	0.934
	50%	-0.25	0.868	0.891	0.928	0.893	0.930
		-0.50	0.866	0.887	0.924	0.891	0.927
		-1.00	0.849	0.871	0.913	0.885	0.920
CC	None	None	0.869	0.895	0.933	0.896	0.934
	25%	-0.25	0.868	0.893	0.931	0.894	0.932
		-0.50	0.866	0.890	0.928	0.891	0.929
		-1.00	0.858	0.878	0.919	0.880	0.920
	50%	-0.25	0.866	0.890	0.928	0.891	0.929
		-0.50	0.857	0.879	0.920	0.881	0.921
		-1.00	0.821	0.839	0.889	0.841	0.890
FPC	None	None	0.871	0.896	0.933	0.897	0.934
	25%	-0.25	0.869	0.893	0.931	0.895	0.932
		-0.50	0.867	0.890	0.929	0.892	0.930
		-1.00	0.859	0.882	0.922	0.884	0.924
	50%	-0.25	0.867	0.890	0.928	0.892	0.930
		-0.50	0.857	0.880	0.921	0.882	0.922
		-1.00	0.827	0.850	0.898	0.853	0.900

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

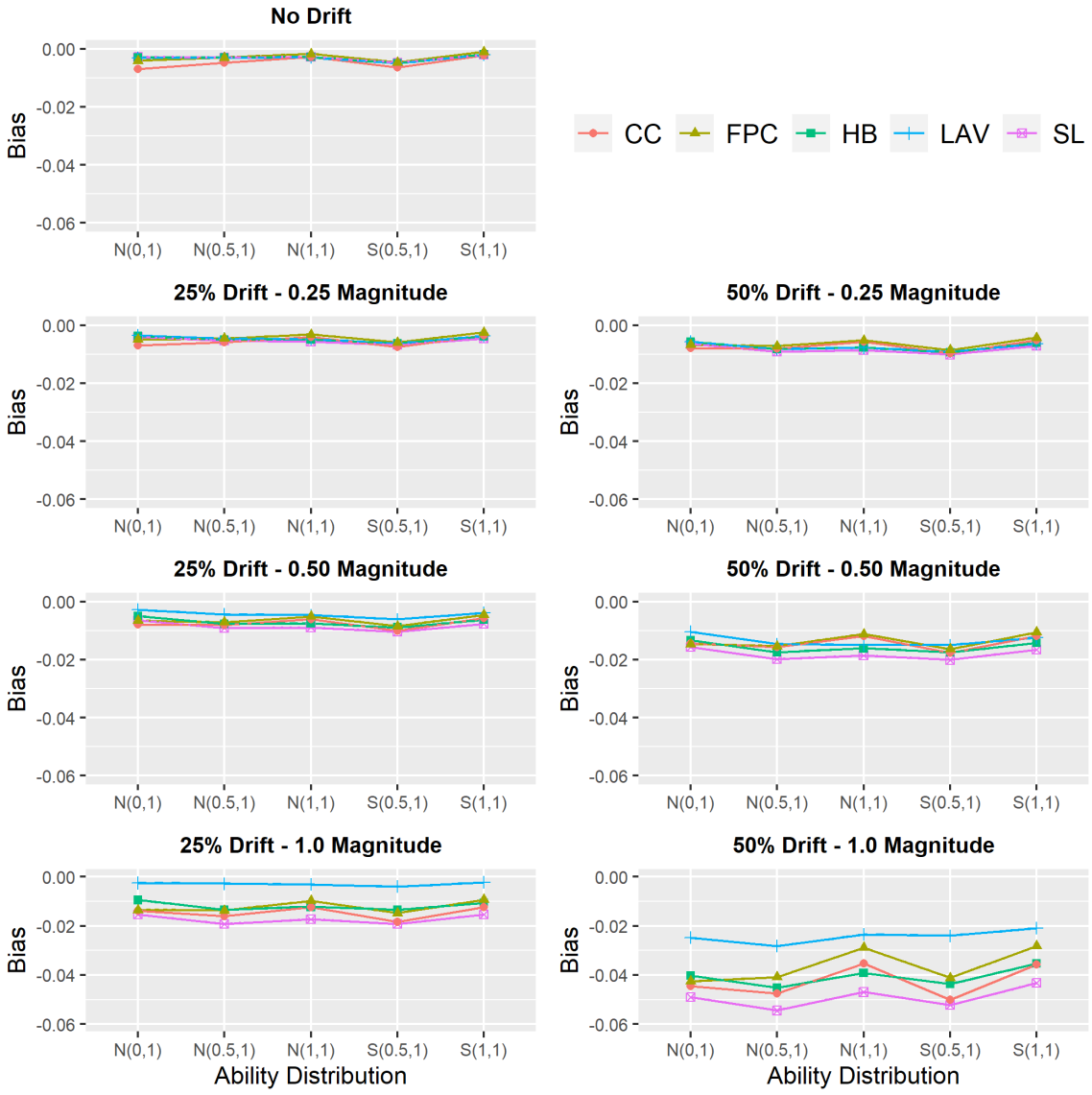

Figure 65. Bias Values for Classification Accuracy – 1,000 Examinees.

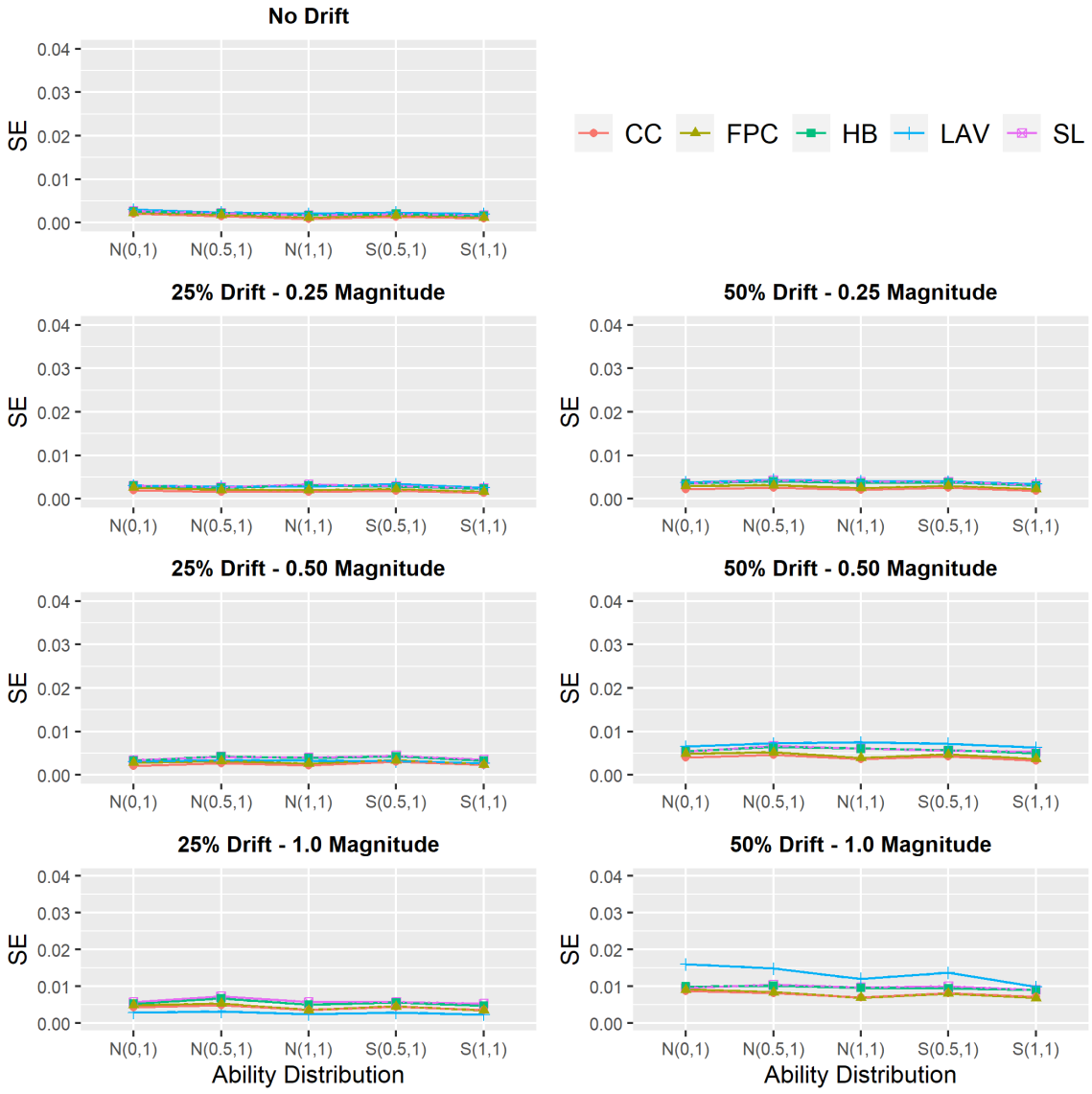

Figure 66. SE Values for Classification Accuracy – 1,000 Examinees.

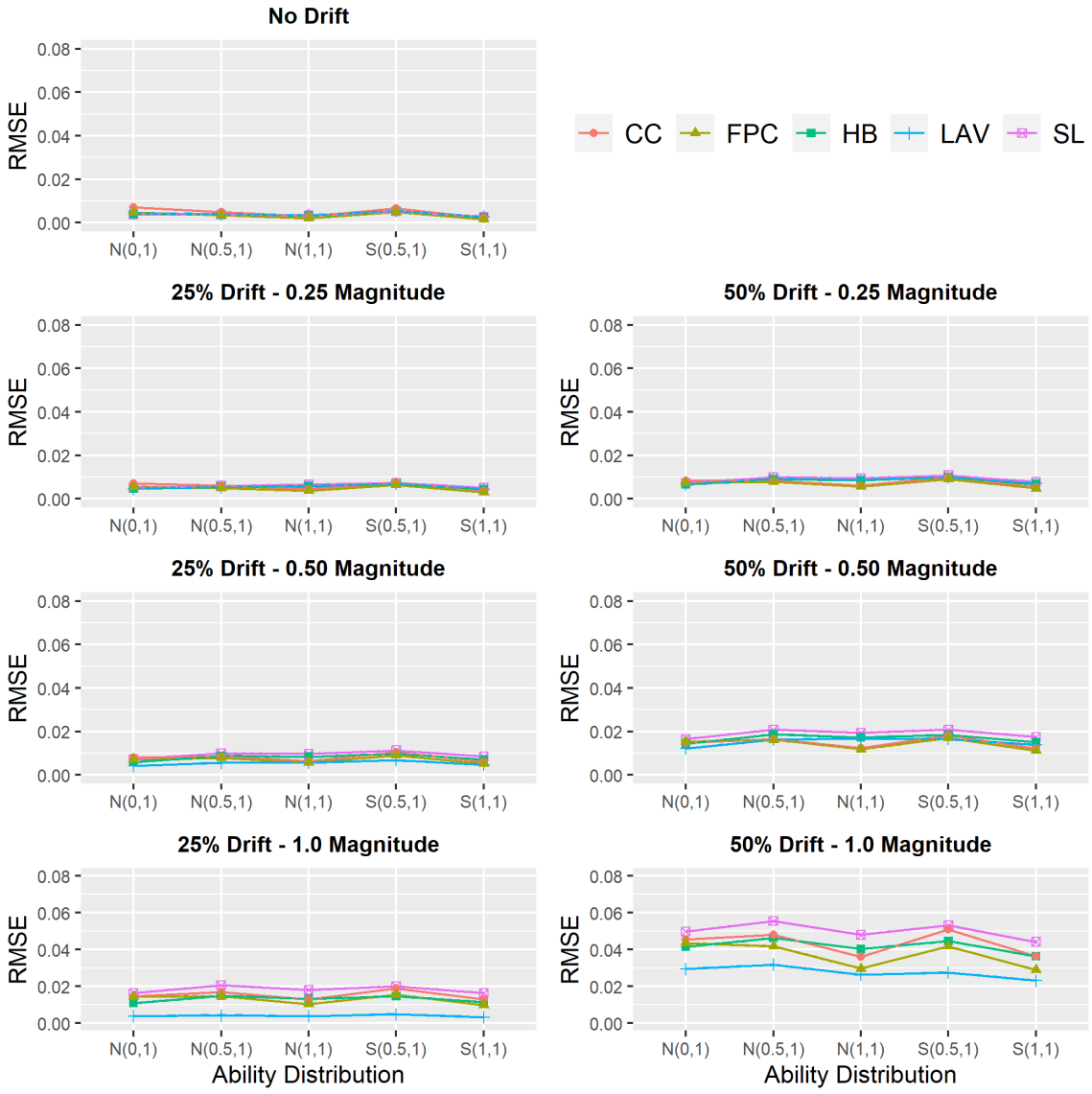

Figure 67. RMSE Values for Classification Accuracy – 1,000 Examinees.

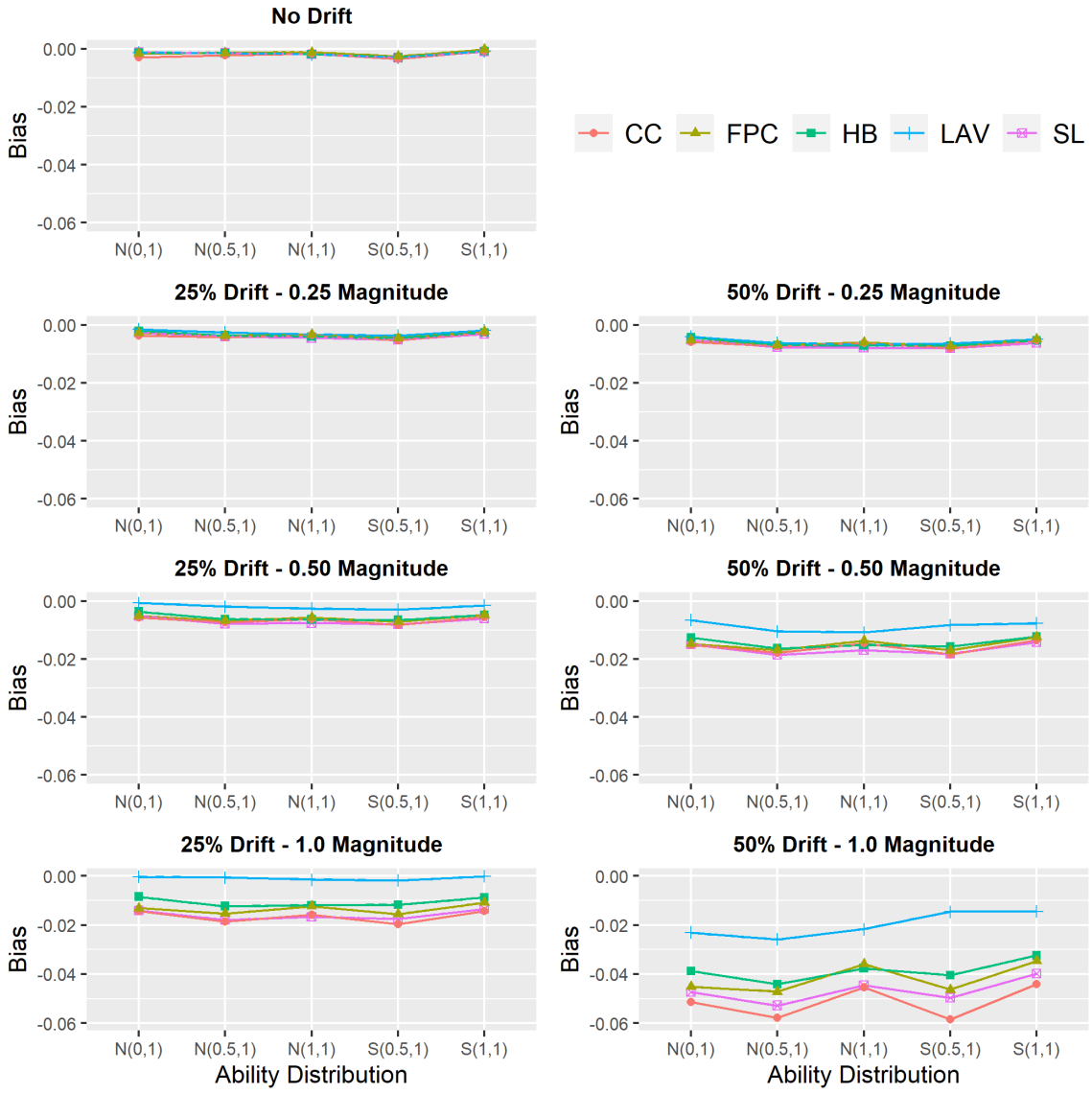

Figure 68. Bias Values for Classification Accuracy – 3,000 Examinees.

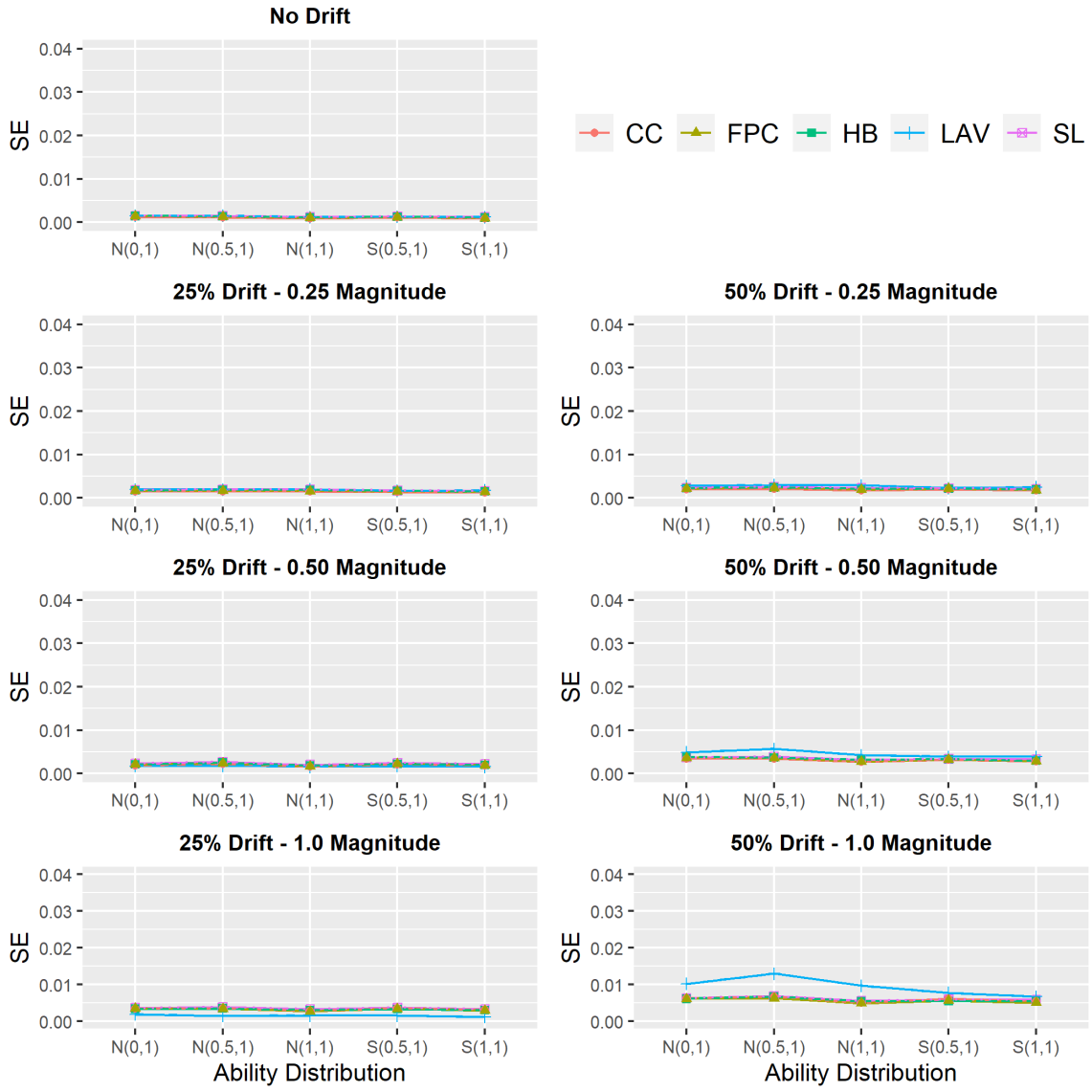

Figure 69. SE Values for Classification Accuracy – 3,000 Examinees.

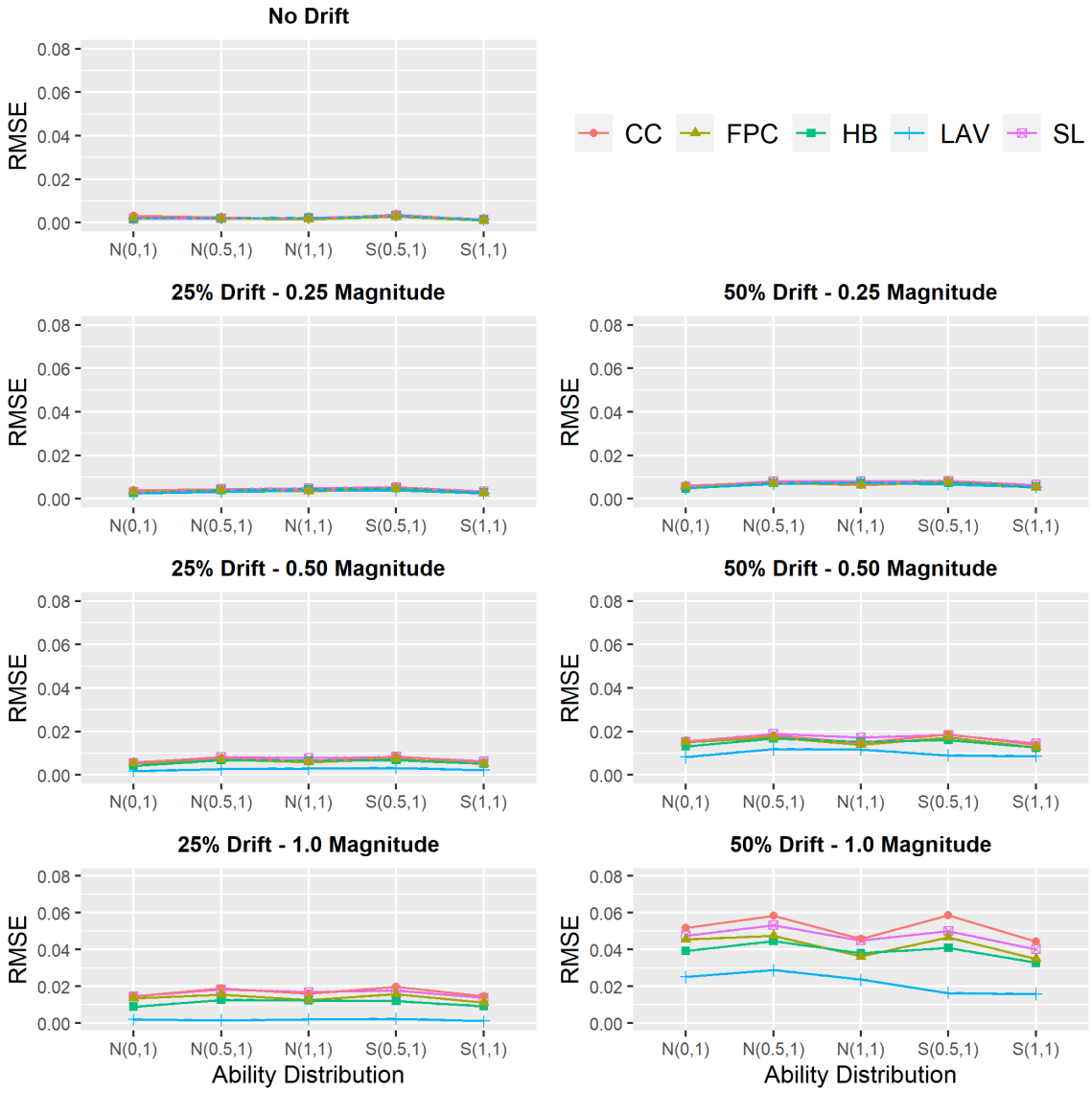

Figure 70. RMSE Values for Classification Accuracy – 3,000 Examinees.

Classification Consistency. The fifth research question examined the extent to which IPD affects classification consistency rates. Bias, SE, and RMSE were calculated by comparing the classification consistency rates using the phi coefficient from each linking method to the true classification criterion, which is defined as the proportion of examinees classified as pass-pass or fail-fail for two independent administrations of a test. There was a total of five true classification criterion, one for each ability distribution. The five true classification criterion rates, along with the estimated classification consistency rates can be seen in Tables 12 and 13, for the 1,000 and 3,000 sample-size conditions, respectively. Figures 71 – 73 illustrate the bias, SE, and RMSE values for classification consistency with 1,000 examinees. Figures 74 – 76 illustrate the bias, SE, and RMSE values for classification consistency with 3,000 examinees. Specific values for each of these three outcomes can be found in Appendix G.

For both sample sizes, each linking method produced similar classification consistency rates for each of the drift conditions. In most instances, consistency rates were slightly underestimated. All linking methods performed similarly in terms of RMSE. However, some observations could be made when inspecting bias. No discernable pattern for bias could be observed as the ability distributions increased. Only under the most extreme condition of drift (50% drifted items, -1.00 drift magnitude), could some difference in bias be observed between the linking methods. For N(0,1), consistency was slightly overestimated by the separate calibration methods, and slightly underestimated by CC and FPC. However, SL and FPC produced the least amount of bias for N(0,1). For the remaining ability distributions, the LAV method produced the

smallest values of bias. The SE values were very similar for all of the linking methods at all levels of drift. The RMSE values followed a similar pattern that was present among the findings for bias.

Table 12

Classification Consistency Rates – 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
True Classification Criterion			**0.872**	**0.897**	**0.934**	**0.897**	**0.934**
SL	None	None	0.868	0.892	0.930	0.893	0.932
	25%	-0.25	0.868	0.890	0.927	0.891	0.928
		-0.50	0.868	0.888	0.924	0.889	0.925
		-1.00	0.870	0.886	0.919	0.886	0.920
	50%	-0.25	0.868	0.887	0.924	0.888	0.925
		-0.50	0.869	0.884	0.918	0.884	0.919
		-1.00	0.876	0.882	0.910	0.880	0.910
HB	None	None	0.867	0.892	0.930	0.893	0.932
	25%	-0.25	0.868	0.890	0.927	0.891	0.929
		-0.50	0.871	0.890	0.925	0.890	0.927
		-1.00	0.877	0.892	0.925	0.893	0.926
	50%	-0.25	0.869	0.888	0.924	0.889	0.926
		-0.50	0.872	0.887	0.920	0.887	0.921
		-1.00	0.887	0.892	0.918	0.891	0.919
LAV	None	None	0.867	0.892	0.930	0.893	0.932
	25%	-0.25	0.868	0.891	0.928	0.892	0.929
		-0.50	0.870	0.892	0.928	0.892	0.929
		-1.00	0.869	0.894	0.930	0.896	0.932
	50%	-0.25	0.869	0.888	0.924	0.890	0.926
		-0.50	0.874	0.889	0.921	0.889	0.923
		-1.00	0.892	0.900	0.925	0.900	0.927
CC	None	None	0.860	0.889	0.931	0.891	0.932
	25%	-0.25	0.860	0.886	0.928	0.888	0.929
		-0.50	0.860	0.883	0.924	0.885	0.925
		-1.00	0.861	0.878	0.917	0.880	0.918
	50%	-0.25	0.860	0.883	0.924	0.885	0.925
		-0.50	0.860	0.878	0.917	0.879	0.918
		-1.00	0.864	0.868	0.901	0.869	0.903
FPC	None	None	0.865	0.892	0.932	0.894	0.934
	25%	-0.25	0.865	0.889	0.929	0.891	0.931
		-0.50	0.865	0.887	0.926	0.889	0.927
		-1.00	0.866	0.884	0.921	0.886	0.923
	50%	-0.25	0.865	0.887	0.926	0.888	0.928
		-0.50	0.865	0.882	0.920	0.884	0.921
		-1.00	0.870	0.877	0.910	0.878	0.911

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Table 13

Classification Consistency Rates – 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
True Classification Criterion			**0.872**	**0.897**	**0.934**	**0.897**	**0.934**
SL	None	None	0.870	0.895	0.932	0.896	0.933
	25%	-0.25	0.870	0.892	0.928	0.893	0.930
		-0.50	0.871	0.890	0.926	0.891	0.927
		-1.00	0.874	0.888	0.921	0.888	0.922
	50%	-0.25	0.871	0.890	0.925	0.891	0.927
		-0.50	0.872	0.886	0.920	0.886	0.921
		-1.00	0.879	0.884	0.912	0.882	0.912
HB	None	None	0.870	0.894	0.932	0.896	0.933
	25%	-0.25	0.871	0.892	0.929	0.894	0.931
		-0.50	0.874	0.892	0.927	0.893	0.929
		-1.00	0.881	0.895	0.926	0.895	0.928
	50%	-0.25	0.872	0.891	0.926	0.892	0.928
		-0.50	0.875	0.889	0.922	0.890	0.924
		-1.00	0.889	0.894	0.920	0.893	0.920
LAV	None	None	0.870	0.894	0.931	0.896	0.933
	25%	-0.25	0.871	0.893	0.930	0.895	0.932
		-0.50	0.872	0.895	0.931	0.897	0.933
		-1.00	0.872	0.896	0.932	0.899	0.935
	50%	-0.25	0.872	0.891	0.926	0.893	0.928
		-0.50	0.879	0.894	0.926	0.896	0.929
		-1.00	0.897	0.903	0.929	0.905	0.932
CC	None	None	0.867	0.893	0.932	0.895	0.934
	25%	-0.25	0.867	0.890	0.928	0.892	0.930
		-0.50	0.868	0.888	0.925	0.890	0.926
		-1.00	0.870	0.884	0.917	0.885	0.919
	50%	-0.25	0.867	0.887	0.925	0.889	0.927
		-0.50	0.868	0.882	0.918	0.884	0.920
		-1.00	0.874	0.874	0.902	0.876	0.904
FPC	None	None	0.869	0.895	0.933	0.897	0.935
	25%	-0.25	0.869	0.892	0.929	0.894	0.931
		-0.50	0.870	0.890	0.926	0.892	0.929
		-1.00	0.873	0.887	0.922	0.889	0.924
	50%	-0.25	0.870	0.889	0.926	0.892	0.928
		-0.50	0.871	0.885	0.920	0.888	0.922
		-1.00	0.876	0.880	0.910	0.883	0.913

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

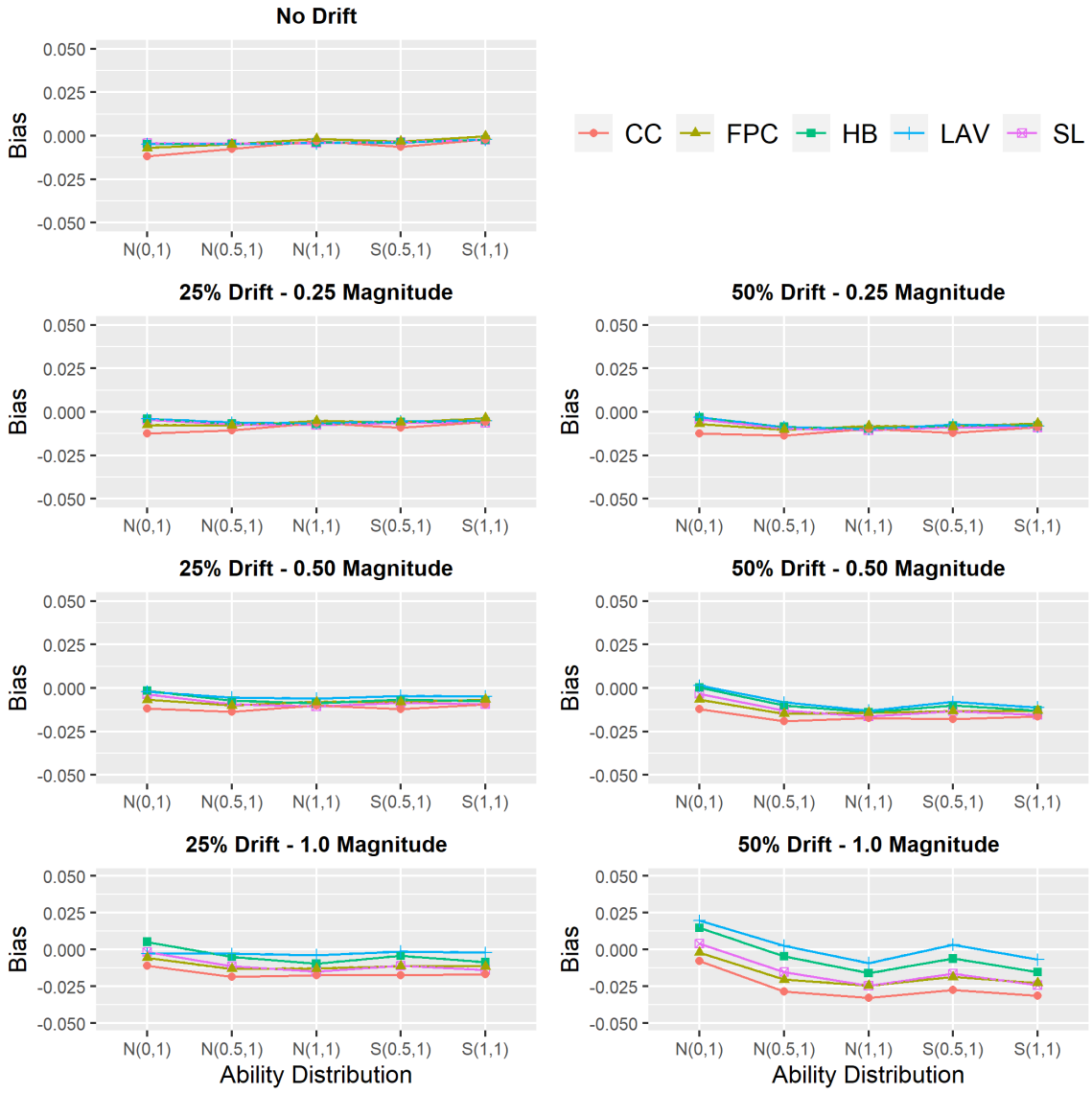

Figure 71. Bias Values for Classification Consistency – 1,000 Examinees.

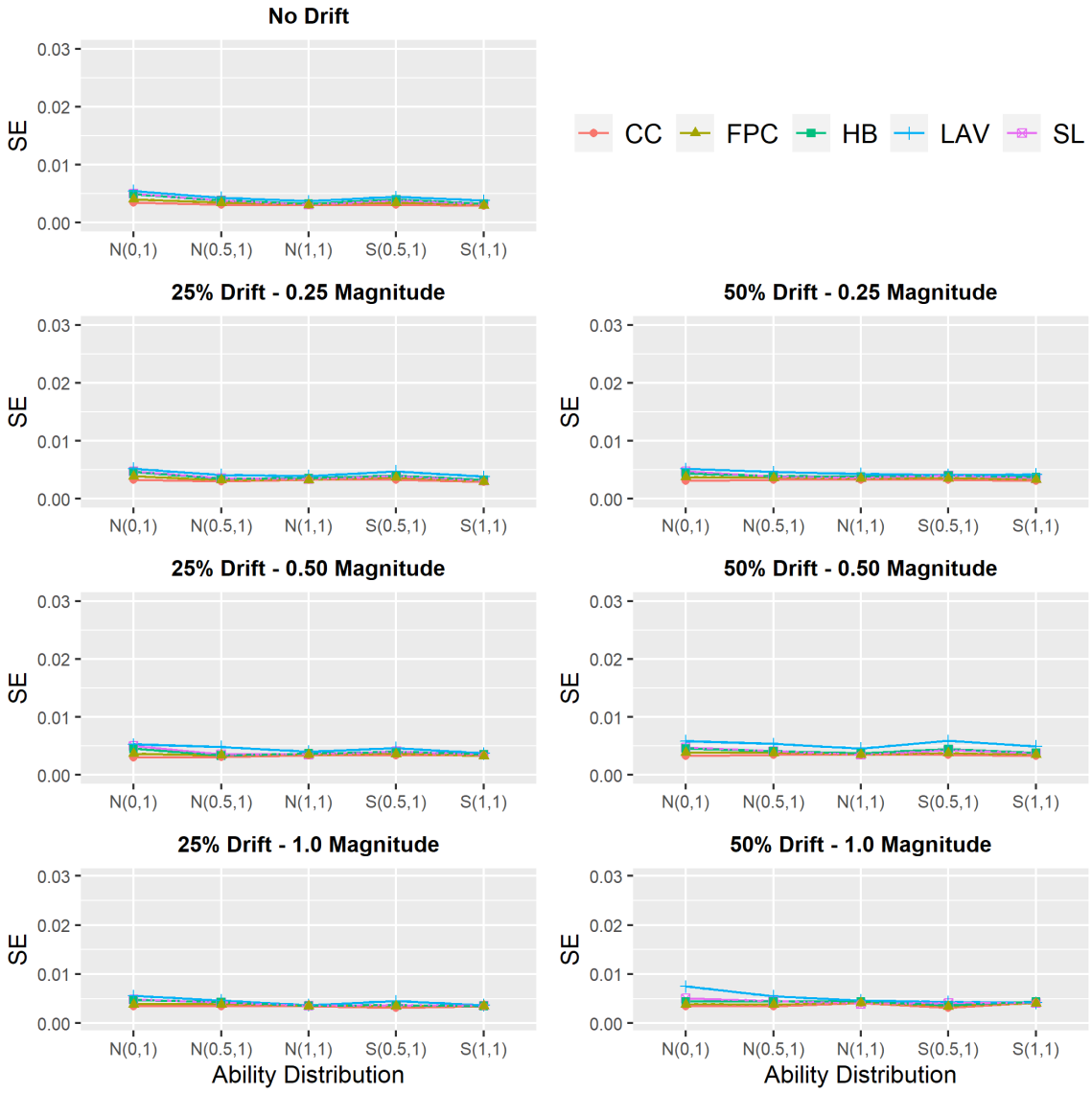

Figure 72. SE Values for Classification Consistency – 1,000 Examinees.

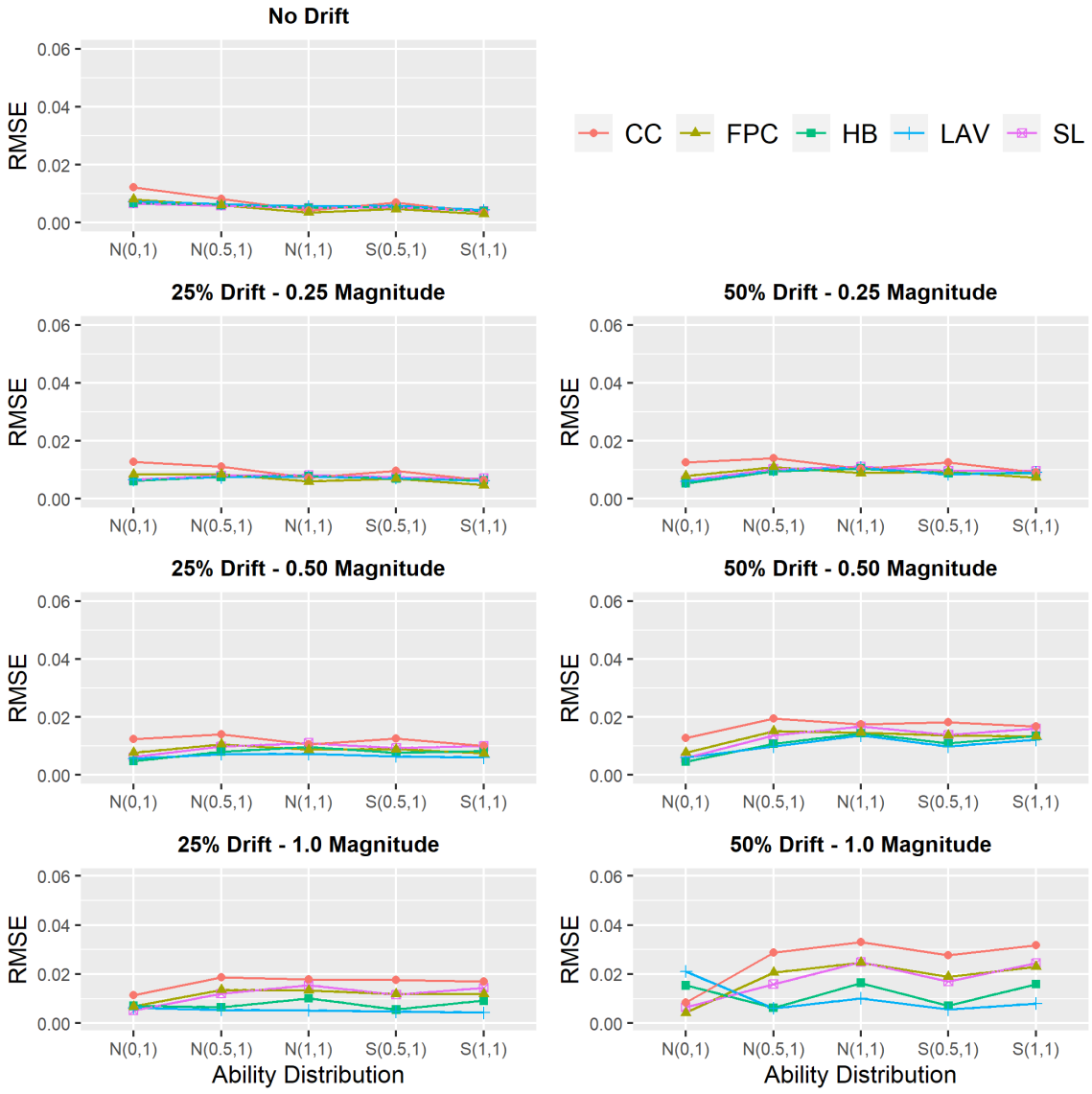

Figure 73. RMSE Values for Classification Consistency – 1,000 Examinees.

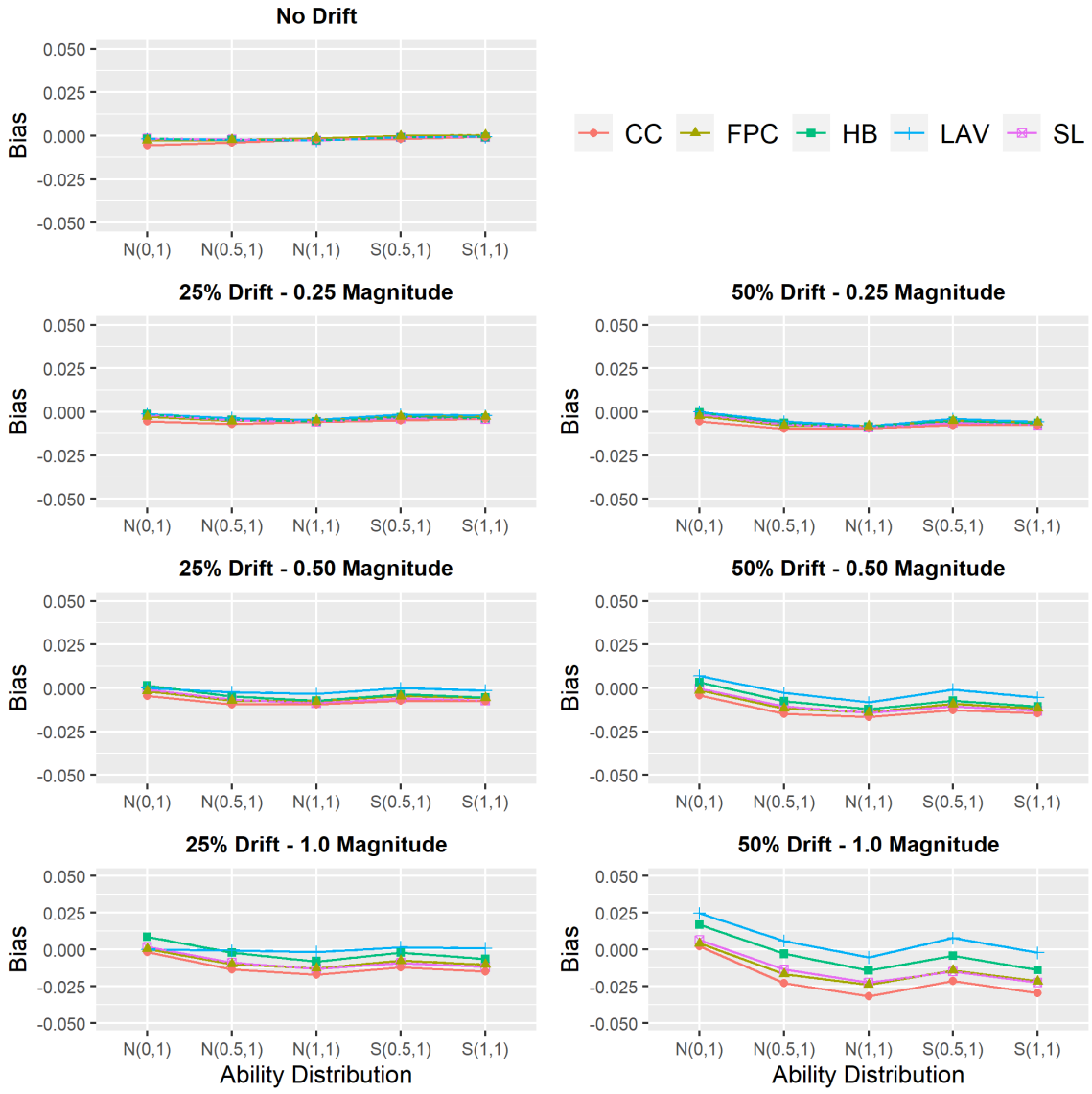

Figure 74. Bias Values for Classification Consistency – 3,000 Examinees.

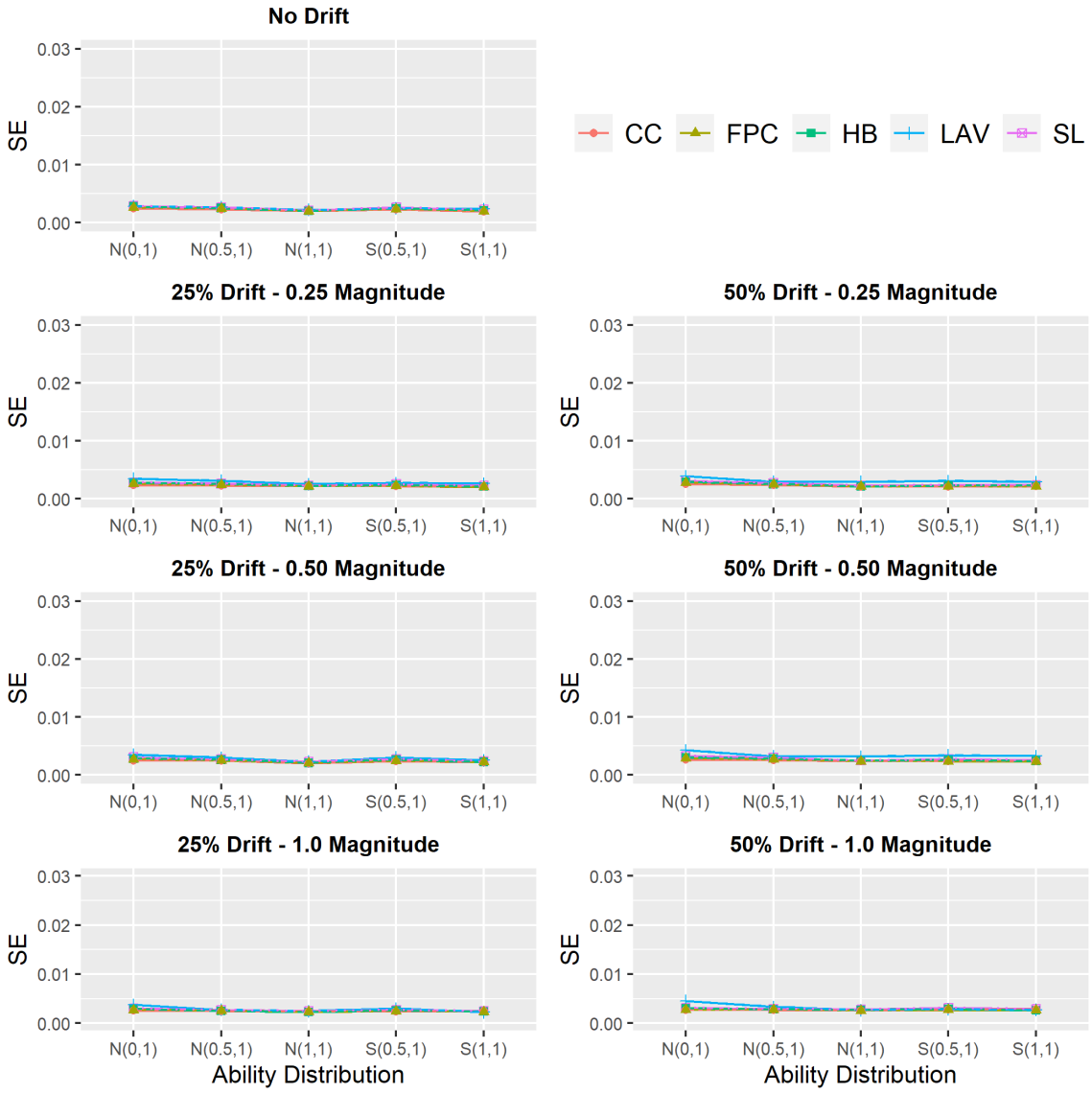

Figure 75. SE Values for Classification Consistency – 3,000 Examinees.

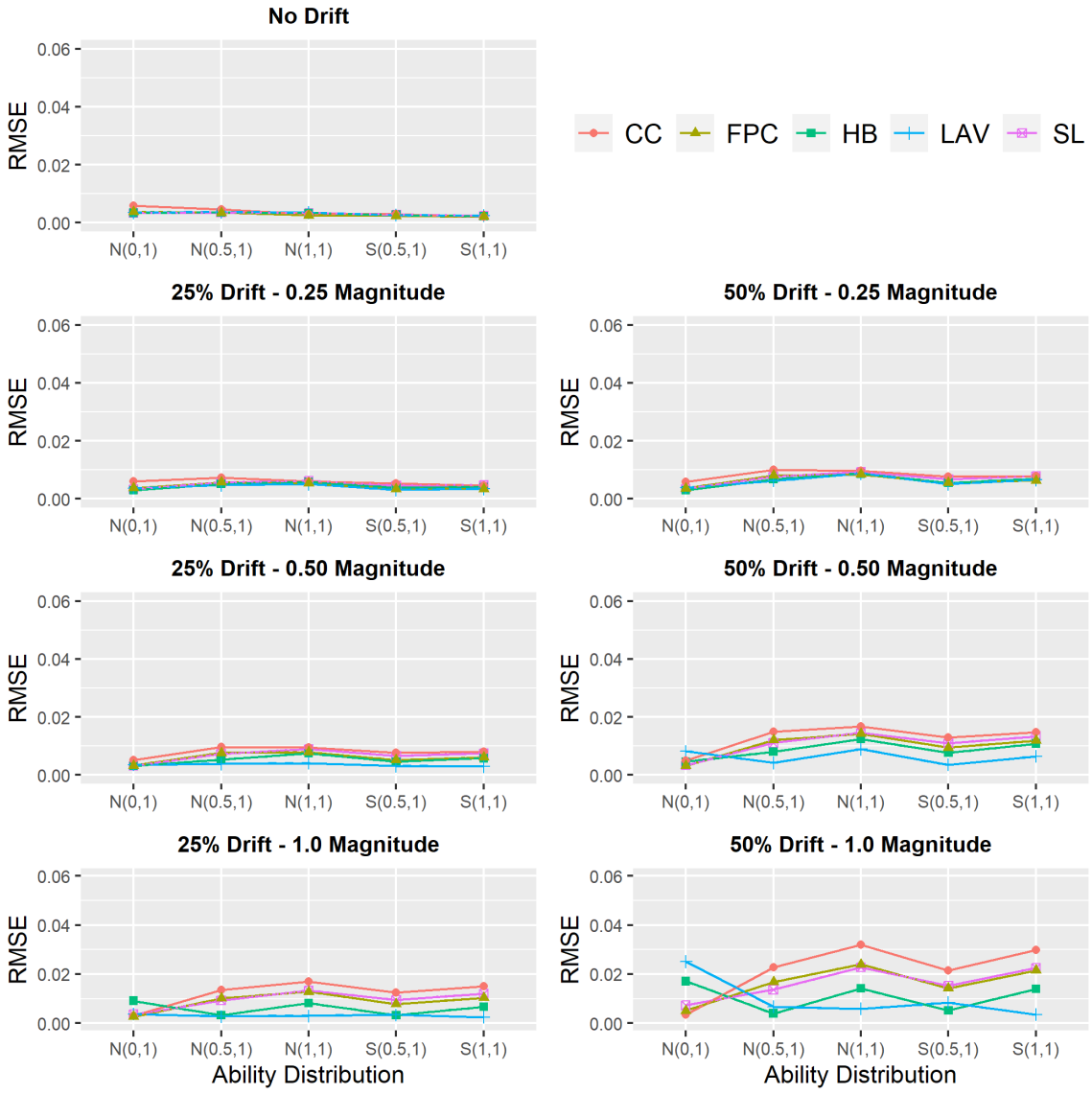

Figure 76. RMSE Values for Classification Consistency – 3,000 Examinees.

Linking Method Comparison. The simulation results presented in the preceding sections are summarized below. Tables 14 and 15 represent the linking methods that yielded the smallest value of RMSE for a particular condition. In some instances, multiple methods performed similarly, and differences could not be separated. These findings can be used by practitioners and researchers to determine which linking method might be the most useful when confronted with drift. Interpretation of these findings can be found in the discussion section.

Table 14

Summary of Simulation Results – 1,000 Examinees

	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0,1)	N(0.5,1)	N(1,1)	S(0.5,1)	S(1,1)
Linking Constant A	None	None	SC	SC	FPC	SC	SC
	25%	-0.25	SC	SC/FPC	FPC	SC	SC/FPC
		-0.50	SC/FPC	FPC	FPC	FPC	FPC
		-1.00	FPC	FPC	FPC	FPC	FPC
	50%	-0.25	SC	FPC	FPC	SC/FPC	SC/FPC
		-0.50	SL/FPC	FPC	FPC	FPC	FPC
		-1.00	FPC/CC	FPC	FPC	FPC	FPC
Linking Constant B	None	None	SC/FPC	FPC/CC	FPC	ALL	FPC
	25%	-0.25	CC	SC	SC/FPC	SC	SL/HB
		-0.50	LAV	LAV	SL/HB	LAV	SC
		-1.00	LAV	LAV	HB	LAV	HB
	50%	-0.25	FPC/CC	SC	SC	SC	SC
		-0.50	LAV	LAV	LAV/HB	LAV	LAV/HB
		-1.00	LAV	LAV	LAV	LAV	LAV
Item Estimate a	None	None	ALL	FPC/CC	FPC/CC	ALL	FPC/CC
	25%	-0.25	FPC/CC	FPC/CC	FPC/CC	ALL	FPC/CC
		-0.50	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
		-1.00	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
	50%	-0.25	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
		-0.50	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
		-1.00	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
Item Estimate b	None	None	ALL	ALL	FPC/CC	ALL	FPC/CC
	25%	-0.25	ALL	ALL	FPC	ALL	FPC
		-0.50	LAV	LAV/FPC	FPC	LAV/FPC	FPC
		-1.00	LAV	LAV	LAV	LAV	LAV
	50%	-0.25	ALL	FPC/CC	FPC	FPC	FPC
		-0.50	LAV	FPC	FPC	LAV/FPC	FPC
		-1.00	LAV	LAV	LAV	LAV	LAV
Item Estimate c	None	None	ALL	ALL	ALL	ALL	ALL
	25%	-0.25	ALL	ALL	ALL	ALL	ALL
		-0.50	ALL	ALL	ALL	ALL	ALL
		-1.00	ALL	ALL	ALL	ALL	ALL
	50%	-0.25	ALL	ALL	ALL	ALL	ALL
		-0.50	ALL	ALL	ALL	ALL	ALL
		-1.00	ALL	ALL	ALL	ALL	ALL

SC = Separate Calibration (SL/HB/LAV); CC = Concurrent Calibration; FPC = Fixed Parameter Calibration; SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; ALL = All five linking methods

Table 14 continued

Summary of Simulation Results – 1,000 Examinees

	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0,1)	N(0.5,1)	N(1,1)	S(0.5,1)	S(1,1)
True Scores	None	None	SC	SC	FPC	SC	SC
	25%	-0.25	SC	SC/FPC	FPC	SC	SC/FPC
		-0.50	SC/FPC	FPC	FPC	FPC	FPC
		-1.00	FPC	FPC	FPC	FPC	FPC
	50%	-0.25	SC	FPC	FPC	SC/FPC	SC/FPC
		-0.50	SL/FPC	FPC	FPC	FPC	FPC
		-1.00	FPC/CC	FPC	FPC	FPC	FPC
Observed Scores	None	None	SC/FPC	FPC/CC	FPC	ALL	FPC
	25%	-0.25	CC	SC	SC/FPC	SC	SL/HB
		-0.50	LAV	LAV	SL/HB	LAV	SC
		-1.00	LAV	LAV	HB	LAV	HB
	50%	-0.25	FPC/CC	SC	SC	SC	SC
		-0.50	LAV	LAV	LAV/HB	LAV	LAV/HB
		-1.00	LAV	LAV	LAV	LAV	LAV
Accuracy	None	None	ALL	FPC/CC	FPC/CC	ALL	FPC/CC
	25%	-0.25	FPC/CC	FPC/CC	FPC/CC	ALL	FPC/CC
		-0.50	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
		-1.00	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
	50%	-0.25	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
		-0.50	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
		-1.00	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC/CC
Consist-ency	None	None	ALL	ALL	FPC/CC	ALL	FPC/CC
	25%	-0.25	ALL	ALL	FPC	ALL	FPC
		-0.50	LAV	LAV/FPC	FPC	LAV/FPC	FPC
		-1.00	LAV	LAV	LAV	LAV	LAV
	50%	-0.25	ALL	FPC/CC	FPC	FPC	FPC
		-0.50	LAV	FPC	FPC	LAV/FPC	FPC
		-1.00	LAV	LAV	LAV	LAV	LAV

SC = Separate Calibration (SL/HB/LAV); CC = Concurrent Calibration; FPC = Fixed Parameter Calibration; SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; ALL = All five linking methods

Table 15

Summary of Simulation Results – 3,000 Examinees

	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0,1)	N(0.5,1)	N(1,1)	S(0.5,1)	S(1,1)
Linking Constant A	None	None	SC	SC/FPC	FPC	SC/FPC	SC/FPC
	25%	-0.25	SC/FPC	FPC	FPC/CC	FPC	SC/FPC
		-0.50	FPC	FPC/CC	FPC/CC	FPC/CC	SL/FPC
		-1.00	CC	CC	CC	CC	FPC/CC
	50%	-0.25	FPC	FPC	FPC/CC	ALL	FPC
		-0.50	FPC/CC	FPC/CC	FPC/CC	FPC/CC	FPC
		-1.00	CC	CC	CC	CC	CC
Linking Constant B	None	None	ALL	ALL	FPC/CC	ALL	SC/FPC
	25%	-0.25	LAV/CC	SC	SC	LAV	SC
		-0.50	LAV	LAV	LAV/HB	LAV	LAV
		-1.00	LAV	LAV	LAV/HB	LAV	LAV
	50%	-0.25	ALL	LAV/HB	SC	LAV/HB	LAV/HB
		-0.50	LAV	LAV	LAV	LAV	LAV
		-1.00	LAV	LAV	LAV	LAV	LAV
Item Estimate a	None	None	ALL	FPC/CC	FPC/CC	ALL	ALL
	25%	-0.25	ALL	FPC/CC	FPC/CC	ALL	ALL
		-0.50	FPC/CC	FPC/CC	FPC/CC	ALL	ALL
		-1.00	CC	CC	CC	CC	CC
	50%	-0.25	FPC/CC	FPC/CC	FPC/CC	ALL	ALL
		-0.50	FPC/CC	CC	CC	CC	CC
		-1.00	CC	CC	CC	CC	CC
Item Estimate b	None	None	ALL	FPC/CC	FPC/CC	ALL	FPC/CC
	25%	-0.25	LAV	FPC/CC	FPC/CC	LAV/FPC	FPC/CC
		-0.50	LAV	LAV	LAV/FPC	LAV	LAV
		-1.00	LAV	LAV	LAV	LAV	LAV
	50%	-0.25	ALL	FPC/CC	FPC/CC	LAV/FPC	FPC/CC
		-0.50	LAV	LAV	FPC	LAV	LAV/FPC
		-1.00	LAV	LAV	LAV	LAV	LAV
Item Estimate c	None	None	ALL	ALL	ALL	ALL	ALL
	25%	-0.25	ALL	ALL	ALL	ALL	ALL
		-0.50	ALL	ALL	ALL	ALL	ALL
		-1.00	ALL	ALL	ALL	ALL	ALL
	50%	-0.25	ALL	ALL	ALL	ALL	ALL
		-0.50	ALL	ALL	ALL	ALL	ALL
		-1.00	ALL	ALL	ALL	ALL	ALL

SC = Separate Calibration (SL/HB/LAV); CC = Concurrent Calibration; FPC = Fixed Parameter Calibration; SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; ALL = All five linking methods

Table 15 continued

Summary of Simulation Results – 3,000 Examinees

	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0,1)	N(0.5,1)	N(1,1)	S(0.5,1)	S(1,1)
True Scores	None	None	FPC	FPC	FPC	FPC	SL/HB
	25%	-0.25	LAV	FPC	FPC	LAV	LAV
		-0.50	LAV	LAV	LAV	LAV	LAV
		-1.00	LAV	LAV	LAV	LAV	LAV
	50%	-0.25	HB	LAV/HB	FPC	LAV	LAV/HB
		-0.50	LAV	LAV	LAV	LAV	LAV
		-1.00	HB	LAV	LAV	LAV	LAV
Observed Scores	None	None	FPC	FPC	FPC	CC	HB
	25%	-0.25	LAV	FPC	FPC	LAV	LAV
		-0.50	LAV	LAV	FPC	LAV	LAV
		-1.00	LAV	LAV	LAV	LAV	LAV
	50%	-0.25	HB/CC	HB/LAV	FPC	LAV	HB/LAV
		-0.50	LAV	LAV	LAV	LAV	LAV
		-1.00	SL/HB	LAV	LAV	LAV	LAV
Accuracy	None	None	ALL	ALL	ALL	ALL	ALL
	25%	-0.25	ALL	ALL	ALL	ALL	ALL
		-0.50	ALL	ALL	ALL	ALL	ALL
		-1.00	ALL	ALL	ALL	ALL	ALL
	50%	-0.25	ALL	ALL	ALL	ALL	ALL
		-0.50	ALL	ALL	ALL	ALL	ALL
		-1.00	LAV	LAV	LAV	LAV	LAV
Consist-ency	None	None	ALL	ALL	ALL	ALL	ALL
	25%	-0.25	ALL	ALL	ALL	ALL	ALL
		-0.50	ALL	ALL	ALL	ALL	ALL
		-1.00	ALL	ALL	ALL	ALL	ALL
	50%	-0.25	ALL	ALL	ALL	ALL	ALL
		-0.50	ALL	ALL	ALL	ALL	ALL
		-1.00	CC/FPC	HB/LAV	LAV	HB/LAV	LAV

SC = Separate Calibration (SL/HB/LAV); CC = Concurrent Calibration; FPC = Fixed Parameter Calibration; SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; ALL = All five linking methods

Empirical Analysis

Data from two forms of a high-stakes certification program were administered. For this analysis, the new form was linked to the base form, and results from each linking method are presented. Descriptive statistics for both forms are provided in Table 16. The base form and the new form were both built to similar statistical specifications as summarized in Table 17. A list of all item parameters can be found in Appendix H.

Table 16

Descriptive Statistics for Test Forms

	Base Form	**New Form**
	Mean (standard deviation)	Mean (standard deviation)
Total Items	110	110
Common Items	66	66
Number of Examinees	1,990	1,979
Test Score	88.02 (8.80)	89.55 (9.11)

Drift Detection. In order to connect the findings from the simulation study with that of the empirical data, it was important to determine whether any common items exhibited IPD. Alpha was set to 0.0007 (.05/66 common items) with the Bonferroni correction in order not to inflate the Type I error rate. Out of 110 scored items on the test forms, there were a total of 66 common items shared between forms. Among the 66 common items, eight (12%) were flagged for drift using the backward likelihood ratio test in the mirt package. Five of the eight drifted items appeared to become easier over time, which should be expected considering that most of the reasons for drift result in items becoming easier. The average difference between the base form and new form

Table 17

Item Estimates for Test Forms

	Base Form		New Form	
	Mean	Standard Deviation	Mean	Standard Deviation
Common Items				
Discrimination	0.797	0.293	0.797	0.293
Difficulty	-1.908	1.656	-1.908	1.656
Pseudo-guessing	0.310	0.041	0.310	0.041
Unique Items				
Discrimination	0.649	0.274	0.750	0.217
Difficulty	-1.582	2.598	-2.012	2.161
Pseudo-guessing	0.299	0.056	0.300	0.030
All Items				
Discrimination	0.738	0.293	0.778	0.265
Difficulty	-1.778	2.079	-1.950	1.865
Pseudo-guessing	0.306	0.048	0.306	0.037

difficulty (using SL estimates) for the drifted items was -0.08, which is based upon the cancellation of positive and negative drifting values. At this proportion and magnitude of drift, results from the empirical analysis are most comparable to the baseline and lowest drift conditions (i.e., 25% drifted items, -0.25 magnitude) in the simulation. Yet, it remains difficult to generalize the findings from the simulation to the empirical analysis due to differences in several factors (e.g., number of common items, direction of drift).

Table 18 provides difficulty statistics using the SL method (for comparison purposes only) for each of the common items detected for drift.

Table 18

Drift Detection for Real Data

Common Item	Base Form Difficulty	New Form Difficulty	Difference
9	0.482	-0.157	-0.640
23	1.400	0.934	-0.465
30	-0.137	-0.717	-0.581
38	-0.591	-0.971	-0.380
40	-1.910	-1.670	0.240
50	-0.267	0.234	0.501
55	-1.027	-1.579	-0.551
56	0.318	0.949	0.631

Linking Constants. The first research question examined linking constants *A* and *B* for each of the linking methods. The HB, CC, and FPC methods produced values of *A* under 1.00, while the SL and LAV methods produced values of *A* over 1.00. However, the difference between all of the linking methods was minimal – FPC had the smallest value of *A* at 0.974 and SL had the highest value at 1.011. These findings were anticipated because the groups linked were expected to be of equivalent ability, therefore linking constant *A* would be close to 1.00 and *B* would be close to 0. As can be seen in Table 19, each linking method returns values close to what was expected.

Linked Item Parameter Estimates. The second research question examined the linked item parameter estimates for all 110 items on the new form. Figure 77 plots the new form linked item parameter difficulty values for the 66 common items. Table 20 displays the mean and standard deviation of the linked item parameter estimates for each linking method. As can be seen in Figure 77 and Table 20, each linking method produced similar estimates for each item parameter. Mean discrimination values ranged from a low

Table 19

Empirical Analysis of Linking Constants

	A	*B*
SL	1.011	-0.081
HB	0.986	-0.123
LAV	1.009	-0.113
CC	0.986	0.030
FPC	0.974	-0.008

of 0.776 by SL to 0.797 by HB. LAV had the smallest average difficulty value of -1.999 and CC had the largest average difficulty value of -1.898. The pseudo-guessing parameters were nearly identical between the linking methods.

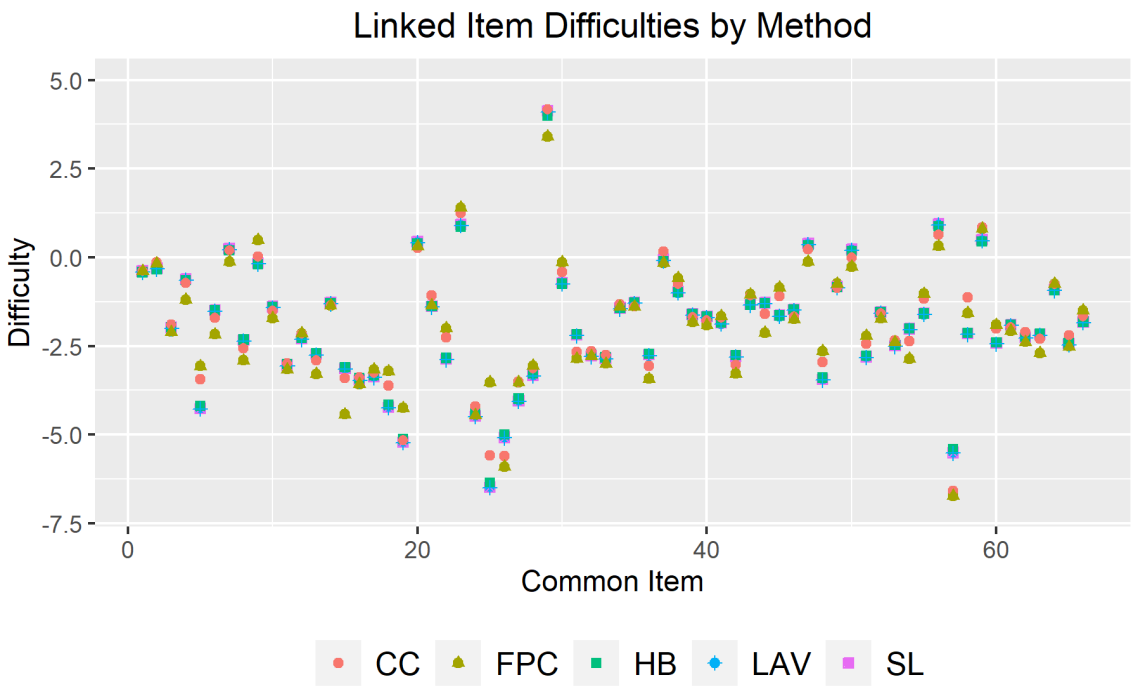

Figure 77. Linked Item Difficulty Values by Method.

Table 20

Empirical Analysis of Linked Item Parameter Estimates

	Discrimination		**Difficulty**		**Pseudo-Guessing**	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
SL	0.776	0.263	-1.972	1.916	0.303	0.037
HB	0.797	0.270	-1.967	1.868	0.303	0.037
LAV	0.778	0.264	-1.999	1.911	0.303	0.037
CC	0.788	0.273	-1.898	1.908	0.308	0.048
FPC	0.778	0.265	-1.950	1.865	0.306	0.037

Equated Scores. The third research question examined the equated scores obtained with IRT true score and observed score equating. Results are summarized in two ways. First, the mean and standard deviation of the new form (Form X) equated scores for each linking method are provided in Table 21. Second, difference plots that take the difference between the base form (Form Y) score equivalent and the new form (Form X) equated score are provided for IRT true score and observed equating in Figures 78 and 79, respectively. The PIE software (Hanson & Zeng, 1995) was used for IRT true score and observed score equating, which provided scores below the sum of the pseudo-guessing parameters for IRT true score equating via linear interpolation. Inspection of the mean equated scores (true and observed) for Form X indicated that all linking methods provided similar average scores and standard deviations. As can be seen in the difference plots, all linking methods followed the same trajectory of scores. However, a small spike in the IRT true score equating plot occurred at the sum of the pseudo-guessing parameters (approximately 33-34), where linear interpolation ended. The HB method exhibited the

largest differences between the base form (Form Y) score equivalent and the new form (Form X) score at the low end of the scale for both equated true and observed scores. At the higher end of the scale, for both equated true and observed scores, the HB and LAV methods exhibited the largest differences between the Form Y score equivalent and Form X. CC, FPC, and SL were nearly identical throughout most of the scale.

Table 21

Empirical Equating Results – Form X converted to Form Y Scale

	True Score		**Observed Score**	
	Mean	Standard Deviation	Mean	Standard Deviation
SL	54.16	31.26	53.94	31.31
HB	54.15	31.13	54.03	31.04
LAV	53.98	31.15	53.71	31.23
CC	54.08	31.36	53.89	31.42
FPC	54.12	31.34	53.96	31.36

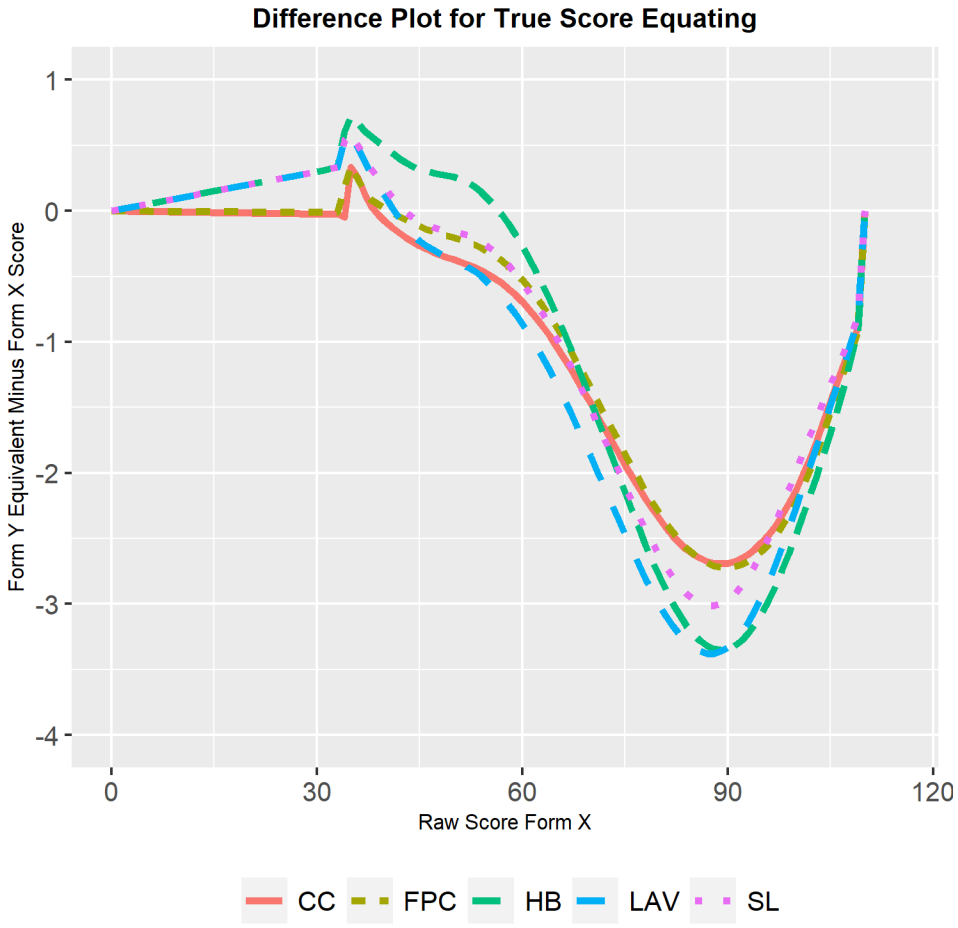

Figure 78. Empirical Analysis of IRT True Score Equating.

Figure 79. Empirical Analysis of IRT Observed Score Equating.

Classification. The fourth and fifth research questions examined the classification accuracy and consistency rates using the phi coefficient. As can be seen from Table 22, each linking method had similar accuracy and similar consistency rates. Classification consistency rates were lower than classification accuracy, which is similar to other studies examining classification rates with IRT (e.g., Lee, 2010; Lee et al., 2002; Wyse & Hao, 2012).

Table 22

Marginal Classification Accuracy and Consistency Rates

	Accuracy	***Consistency***
SL	0.899	0.859
HB	0.896	0.864
LAV	0.894	0.862
CC	0.909	0.856
FPC	0.904	0.859

Figures 80 and 81 plot the conditional classification accuracy and consistency rates, respectively. These plots display the probability of an examinee being classified as pass-pass or fail-fail based upon their expected sum score. The expected summed score was calculated by using the Lord and Wingersky recursion formula (1984) As can be seen, the conditional probability based on phi decreases as scores approach the cut score of 85. This occurs because error in ability estimation around the cut score may push an examinees' score past the cut in one instance but pull it below the cut in another instance.

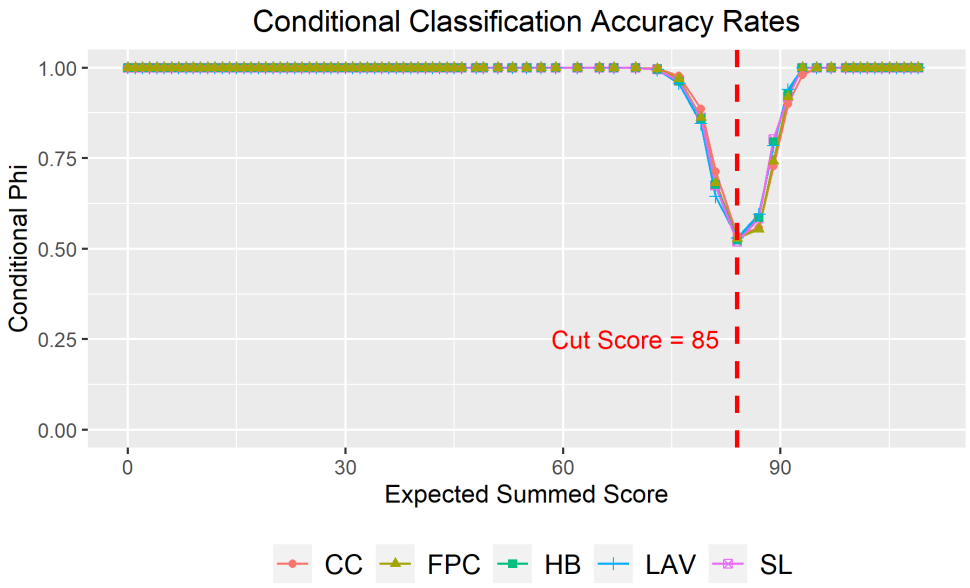

Figure 80. Conditional Classification Accuracy Rates.

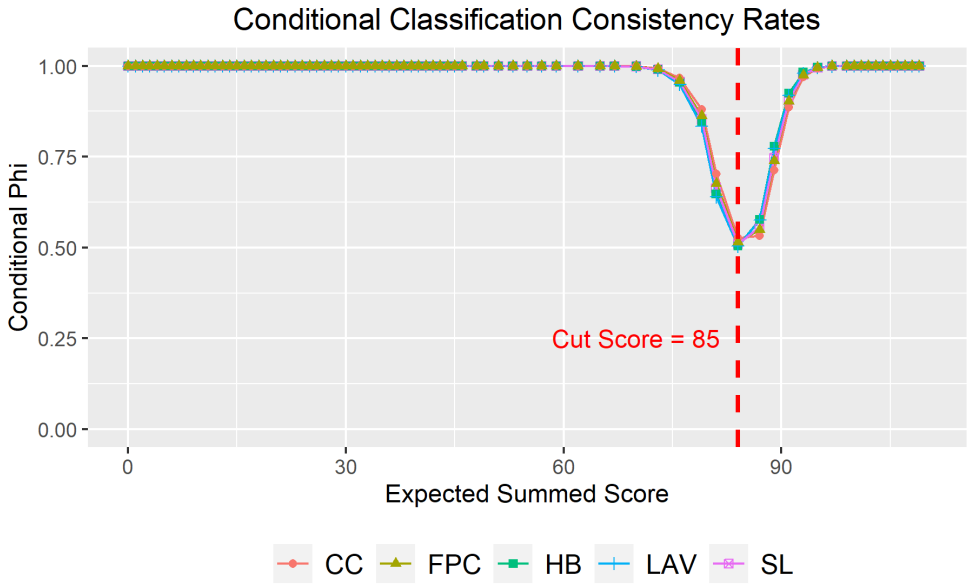

Figure 81. Conditional Classification Consistency Rates.

Validation Implications

The standard protocol for dealing with items that exhibit drift is to remove them from the anchor item set or to unscore them. While this helps to improve linking and equating outcomes, the results from this study and previous studies indicate that drift can go undetected, or false positives can occur (e.g., DeMars, 2004b; Donoghue & Isham, 1998). In order to ensure the accurate interpretation of test scores and their use, additional procedures are needed to help support the validity argument. Researchers have identified a number of reasons that could be responsible for causing items to drift over time. Yet, there are few resources for practitioners to consult when confronted with drift. The recommendations provided in Figure 82 are intended to help practitioners identify, address, and prevent drift from occurring.

Direction of Drift	Potential Reason for Drift	Procedure to Address
Positive (Harder)	Curriculum changes	Review with client (i.e., have certain subject areas been underemphasized recently?); administer curriculum survey
	Technological advances	Review with client; review test specifications (e.g., change in item/exam accessories)
	Item location changes	Review test specifications (i.e., is there consistency between and within test forms?)
	Unstable/poor calibration; improper modeling	Review calibration (e.g., proper model, sample size, populations, compare with previous form/exam calibration)
Negative (Easier)	Curriculum changes	Review with client (i.e., have certain subject areas been overemphasized recently?); administer curriculum survey
	Technological advances	Review with client; review test specifications (e.g., change in item format/item tools)
	Item location changes	Review test specifications (i.e., is there consistency between and within test forms?)
	Unstable/poor calibration; improper modeling	Review calibration (e.g., proper model, sample size, populations, compare with previous form/exam calibration)
	Motivation	Check item stats for field items promoted to scored items
	Item overexposure	Review how many forms an item is or has been active on
	Cheating	Review with client and testing center; check "brain dump" sites/ social media; check item response times
	Security breach	Check for breaches with testing center, testing company
	Test taking strategies; test savviness	Review items for contextual clues (e.g., length of response options); check test preparation courses
	Current news	Review with client (e.g., have new laws, medicine, technology changed the item)

Figure 82. Recommendations for Addressing Drift.

Figure 82 provides a framework for how practitioners may identify the reason for drift occurring, as well as the procedures that can be used to address drift and prevent it from reoccurring. The first column on the left specifies whether the drift is positive (becoming harder) or negative (becoming easier). Since drift will not be identified until a formal detection check has occurred, this serves as a natural starting point. Based upon the direction of drift, the second column lists the potential reasons for drift. The list of reasons is based upon commonly identified sources of drift in research and practice, although more may exist. Citations for each of the reasons can be found in Table 1. All the reasons listed for drift becoming harder are also reasons for why drift could become easier. Six additional reasons are listed, all of which are exclusive to drift becoming easier. The final column lists the procedures that can be used to verify which source of drift may have occurred. Unless there is reason to retain drifted items (e.g., content imbalance of anchor item set; subject matter experts advise against), the drifted items should be removed from the anchor item set. The procedures used to address each reason for drift will be expounded upon further here.

Curriculum changes may result in items becoming easier or harder. It is important to review with the client whether certain subject areas have received more attention (leading to items becoming easier) or less attention (items become harder). The reason for the shift in attention may be driven by changes in policy (Goldstein, 1983) or emphasis by teachers or textbooks (Bock et al., 1988). Where possible, it would be beneficial to administer a survey to gauge examinees' perceptions on how much their study materials

(e.g., classroom or textbook) covered material presented on an exam. Subject matter experts may also be consulted regarding the relevance of certain topical areas.

Advances in technology has also shifted the focus of what examinees have been expected to know and what is provided to them on the exam. Goldstein (1983) documents how providing calculators on exams has phased out the need for mental arithmetic. If the use of calculators or other accessories (e.g., highlighters, instructions, change in test time) has been added or removed while an item is active, it is important to evaluate the item for drift. If drift is found, the item should be unanchored and recalibrated to the item bank. Although these accessories do not change the item itself, any change to the actual item (e.g., wording, font, response options) should be treated as a different item altogether. A review of the test specifications should be enough to determine whether these changes have been made.

Changes in item location have the potential to cause items to drift. Although Sykes and Fitzpatrick (1992) found item location did not influence drift, Kingston and Dorans (1984) found that certain types of items (i.e., items with extensive instructions) could interact with practice effects to result in drift. That is, if all items on a test have the same set of instructions, changing the location of the item might not induce drift (unless randomly presented, test developers should still fix the location as much as possible to avoid enemy items). However, if different items have different sets of instructions, some of which are more convoluted, there is the potential for drift to occur when some examinees are familiar with the item type and others are not.

If the initial item estimation is not conducted according to recommended conditions, then results from the calibration may be more susceptible to error (Glas, 2000). For example, a minimum sample size of 1,000 has been recommended for use with the 3PL model (Hanson & Beguin, 2002). It is also important to consider seasonality effects (Wyse & Babcock, 2016), or changes in the populations (e.g., first-time test takers versus retest-takers) – where differing ability distributions can produce different item difficulty values. Depending upon the type of program, Wyse and Babcock (2016) found that certain programs (i.e., moderate sample size with or without seasonality effects) could conduct IRT calibration as early as eight months into the test development cycle, whereas others (i.e., small sample size with or without seasonality effects) would be better served waiting until the full exam cycle was complete. Even though testing companies have strict procedures for calibration, these practices may have been stretched during COVID-19, when testing centers were forced to close, leaving small sample sizes to be analyzed. Although drift is primarily considered to affect the new form being linked, it can operate rather insidiously if the item estimate is drifted for the bank scale. In this situation, the bank value is afflicted, and the subsequent pre-assembled forms will be easier or harder than intended. Because drift detection can produce false-positives or false-negatives (e.g., DeMars, 2004b; Donoghue & Isham, 1998), it is important to check the *b*-value against previous forms and bank scales to check for drift.

Motivation is a reason why items could become easier over time. It should be checked if items that have been recently promoted from experimental to scored become easier over time (Glas, 2000). If examinees have a sense of which items are field-test and

which are scored, their engagement is likely to vary. Field-test items are likely to be treated less seriously than scored items, therefore the item might appear more difficult than it really is. Once that item becomes scored, examinees would have more incentive to take the item seriously, which may result in the item becoming easier. Practitioners should check the item before and after item promotion.

Item overexposure occurs when the same item has been placed on too many forms leading the item to be recognized by test takers thereby losing its confidentiality (Jurich et al., 2012). It is important for test developers not to overuse items when new forms are assembled, despite desirable psychometric properties. Inspection of retest-takers and the number of forms the item is present on may provide an indication of whether drift may have occurred.

Cheating, or a breach in security, can occur in a number of different ways (Jurich et al., 2012). It should first be determined whether there was widespread cheating, or if cheating was relegated to just a few examinees. Checking with the testing center and the client might provide insight into the source of the cheating (e.g., within the test center, social media). Widespread cheating might be detected based upon differences in difficulty values, as well as overall test scores. Combing through social media and brain dump sites (Smith, 2004) might help to reveal if exam information is being discussed online. If cheating is whittled down to several examinees, it is important to know if the examinees were seated next to each other, and whether their response options were similar. It is also vital to ensure that test materials are handled securely and confidentially

during all meetings (job analysis, standard setting), otherwise, sensitive information can be leaked without the knowledge of any test personnel.

Although no test-takers should be penalized for being test-savvy, items should be reviewed to determine whether there are any obvious contextual clues that are causing the item to perform differently. Messick (1989) suggested that certain clues (e.g., length or response options) may unintentionally give away the correct answer to test-takers. Items can be screened for these problems during item writing, item review, and calibration as a field-test item. Test preparation courses may also give examinees test-taking tips that lead to drift (e.g., methods to attack items with "all of the above" response options).

The last reason listed is the potential for current news and media to provide attention to certain topics. O'Neill et al. (2013) provide an example of how a question about HIV was an arcane immunology topic in 1986 but became a current event in 1992 after an outbreak of cases. Thus, the extra attention provided to the topic would help to increase awareness of HIV and make the item easier over time. Subject areas such as law, medicine, or technology are more likely to change rapidly based upon new laws or discoveries that become available.

Failure to address drift may lead to negative consequences regardless of the direction of drift. For instance, examinees that encounter items that drift harder are unfairly penalized and may fail as a result. In turn, the examinee will have to invest more time and financial resources when restudying. The examinee might decide not to restudy and consider other school or career resources. Alternatively, examinees will benefit from items that drift easier, giving them an unfair advantage, which may help them pass.

Although this seems positive for the examinee, whether it be student or employee, it has the potential to harm the examinee and the organization sponsoring them. For example, a student that gains admission to a university or advanced placement in a class may end up struggling and need remedial help. On the other hand, an employee such as a doctor or lawyer might be at greater risk for harming a patient, client, or organization since he or she does not meet the minimal competence criteria.

Addressing Drift Using Kane's Argument-Based Approach. Chapter 2 discussed how drift affects the validity argument for the use of test scores and their interpretation according to Kane's (2006, 2013) argument-based approach. This section provides practitioners with an example of creating an IUA and validity argument when confronted with drift. The example illustrates how to construct an argument for the use of test scores on a hypothetical licensure exam using Kane's framework. The statement of the intended interpretation for test scores on this hypothetical licensure exam is that an examinee's score on the exam meets the performance standard required for professional practice. As a result, the examinee can practice as a licensed professional. This statement can be seen below in Figure 83.

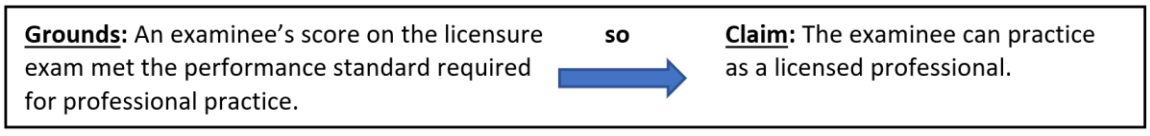

Figure 83. Intended Interpretation of Test Scores for a Licensure Exam.

This claim is then evaluated using Kane's IUA and validity argument, whose framework is illustrated in Figure 84 and has been adapted from Chapelle et al. (2010). The first column lists the four inferences (from Figure 1) required to claim that test scores from the licensure exam are useful for making decisions about whether examinees display at least minimal competence for an entry-level position. The second column lists the warrant that explains the information required to validate each inference. The warrant is composed by assumptions (third column) and backing (fourth column) that lists the requirements to be met, and the empirical analysis conducted to verify the requirements, respectively. It is important to note that the assumptions and backing listed in this example only apply to what IPD has the possibility of influencing. For example, one assumption that is not listed below for the scoring inference is – that rubrics for scoring essays are appropriate for demonstrating varying levels of proficiency regarding the construct of interest. The backing would be that the rubric for scoring essays was reviewed by experts. The actual argument would be much longer than what is listed below and would also include alternative hypotheses or rebuttals that threaten the validity argument. Although this example might be considered overly simplistic or unrealistic, it is meant to serve as a template for how a practitioner might choose to deal with different types of drift.

Inference	Warrant Licensing the Inference	Assumptions Underlying Inferences	Backing
Scoring/ Evaluation	Scoring keys/rubrics, testing conditions, statistical characteristics, and scaling-linking-equating procedures are appropriate, accurate, and/or free of bias	1. The statistical characteristics of items, response options, and test forms are appropriate for criterion-referenced decisions	1. DIF analysis revealed no items. IPD analysis revealed several items that were removed from the anchor set. Additional analyses revealed items were overexposed due to being on multiple test forms which led to un-scoring several items.
		2. The appropriate linking and equating procedures for test scores are used	2. Pre-equating with FPC is used. However, the newest test form had several items that drifted. These items required re-estimation in the bank scale. The new form also required a re-estimation of the cut score based upon the new parameter estimates.
Generalization	Observed scores are estimates of expected scores over the relevant parallel versions of tasks, occasions, and raters	1. Configuration of tasks is appropriate for intended interpretations	1. Evaluation of the test specifications revealed that the drifted items were placed towards the end of the base form and at the beginning of the new form. These items were removed from the anchor set and unscored (as stated in scoring inference).
		2. The appropriate linking and equating procedures for test scores are used	2. Populations to be linked and equated may have differing abilities (May = new graduates; July = more retesters) – studies indicate FPC is most robust to differences in ability; results of linking and equating (e.g., test characteristic curves; standard error estimates; test information) indicate robust properties
Extrapolation	The construct of knowledge and skills as assessed by the exam accounts for the role of performance in practice	1. Performance on the exam is related to criteria indicative of performance in practice	1. Re-tester scores measure their ability + familiarity with exam; first-time test taker scores measure just ability. Drifted items handled in scoring inference.
Utilization/ Decision	Estimates of the quality of performance in the domain of practice are useful for making pass/fail decisions about the minimum competence required for entry-level practice	1. The meaning of test scores is interpretable by stakeholders and useful for making decisions	1. Drifted items have been removed from anchor set and/or unscored.
		2. Evaluation of positive and negative consequences resulting from decision	2. Long-term follow-up; surveys

Note: For this to be a full argument, the table should also include challenges (alternative hypotheses) to each of the inferences that could threaten our validity argument. Each inference would also contain more assumptions and backing beyond the scope of just drift, linking, and equating.
Format adapted from Chappelle et al. (2010).

Figure 84. Example of Addressing IPD using Kane's Validity Argument.

Starting with the scoring inference – which requires that testing conditions, procedures, and scoring are accurate – contains two assumptions that could be affected by drift. The first is that the statistical characteristics of items and forms are appropriate. The backing requires that both DIF and IPD analyses be performed. For this example, IPD detection revealed that several items had been affected, several of which were removed from the anchor set – which is standard operation procedure (*Standards*, p. 98). However, several items also had to be unscored because of evidence of item overexposure. Un-scoring items due to overexposure would be a drastic measure to take, especially because examinees may have been better prepared for these items, or they correctly answered the items due to chance. But if the test specifications reveal that particular items are being

routinely used, then this may be the appropriate action, especially when considering the evidence from generalization. Before moving to generalization, the second assumption states that the appropriate linking and equating procedures are used. The backing for this assumption entails that a cut score had already been established for the new form based upon pre-equating with FPC. Had no drift been detected, no action would be required. However, the items that drifted require re-estimation for the item bank (*Standards*, p. 103). Additionally, the un-scoring of items precipitated the need for the cut score to be re-estimated (assuming a withholding of scores). These procedures help to ensure that practitioners took the necessary measures to ensure that scores for all examinees are fairly treated.

The generalization inference suggests that the observed scores from the test are what we would expect if the examinee took the test at another occasion using a different form, different raters, or different items. The first assumption is that the tasks/items are appropriate for intended interpretations. To substantiate this assumption, the practitioner found that not only was the item used on many test forms, but that the location of the item changed between the most recent forms (*Standards*, p. 85-86). The prior form placed the items at the end of the exam, while the new form placed the same items at the beginning of the exam, which may have enabled a recall effect. The second assumption also required that the appropriate linking and equating procedures were used (*Standards*, p. 97-98). Given that the new form testing window was July, and the prior form was administered in May, an inspection of the test takers may have revealed that the populations had very different ability levels. The May population had a higher proportion

of first-time graduates, while the July population had more retest takers. These reasons combined help to substantiate why several items unanchored and unscored.

In the extrapolation inference, the knowledge being measured should be an accurate reflection of the performance required in practice instead of construct-irrelevant sources. These sources should be investigated by the test developer and minimized where possible (*Standards*, p. 90). Because drift was present, and performance differences were found between re-testers and first time graduates, it could be assumed that re-testers had prior knowledge of the items, indicating that their performance was not just a reflection of their ability, but of their familiarity with the exam. This provided the re-testers with an unfair advantage over first time graduates.

In the utilization inference, the score estimates should reflect the examinees' ability level accurately to make an informed decision about whether minimal competence was met. The first assumption holds that the test scores are accurate, interpretable, and suitable for making decisions – this was substantiated by removing and un-scoring several items. The second assumption suggests the need for an evaluation of unintended consequences from the examinees (*Standards*, p. 30-31). This could facilitate the need for long-term follow up through surveys, interviews, or other measures. These measures would look to determine if the examinees who passed had any subsequent violations, code of misconduct, or other infractions as a result of their performance. A correlational study could determine whether there is a relation between test score and future misconduct. Within the context of licensure, which tests an examinees competence at the

time of the test, these longitudinal follow-ups are unlikely to occur, but they still represent a possibility.

In summary, this example may not depict a realistic scenario because drift is hard to detect, and even harder to determine the reason for. However, this is an illustration of procedures that can be used if drift is present. Furthermore, it provides an example of how the validity argument for test scores can be strengthened or hindered by drift.

Addressing Drift Using the Standards. Although Kane's approach to validation has been widely praised, particularly in language testing, the *Standards* remains the most renowned resource for guidance on testing. This section examines the criteria most relevant to handling drift when constructing a validity argument according to the *Standards'* five sources of validity evidence. Using the same example and intended interpretation (Figure 83), the validity argument according to the *Standards* can be found in Figure 85.

Test Content	Response Processes	Internal Structure	External Relations	Consequences
Standard 4.2: Test specifications should define the desired psychometric properties and the ordering of test items and sections. Standard 4.8: Expert judges were used to review items that may be inappropriate, confusing, offensive, or that may have changed/possess a different meaning. Standard 4.10: Test developers evaluated the psychometric properties of items during test development for DIF, IPD. Standard 4.24: Test developers have reviewed any test content, language, or curriculum that may have been outdated.	Standard 1.12: When statements about the processes employed by observers or scorers are part of the argument for validity, similar information (theoretical or empirical evidence) should be provided. This might include analysis of eye movements or response times if there is suspicion of cheating. Standard 6.6: Efforts should be made to ensure the integrity of test scores. Testing programs may use technologies during scoring to detect possible irregularities, such as computer analyses of erasure patterns, similar answer patterns for multiple test takers, plagiarism from online sources, or unusual item parameter shifts.	Standard 5.6: Testing programs that maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported. Standard 5.15: Characteristics of the anchor test and its similarity to the forms should be presented. Standard 5.19: When tests are created by rearranging items, evidence should indicate no distortions of scale scores. Context effects could occur, including speeded tests, long tests where fatigue may be a factor, adaptive tests, and tests developed from calibrated item pools. Options for gathering evidence include operational recalibrations of item parameter estimates initially derived using pretest data.	Standard 4.13: The test developer should investigate sources of irrelevant variance. A variety of methods may be used including analyses of correlations with measure of other relevant (convergent evidence) and irrelevant constructs (discriminant evidence). Standard 5.23: Cut scores defining categories should be informed by sound empirical data concerning the relation of test performance to the relevant criteria. Suitable criterion groups are often unavailable in credentialing; however, this information could be embedded into the standard setting process (e.g., not competent, minimally competent, competent).	Standard 1.25: When unintended consequences result from test use, it should be investigated if the test is sensitive to other factors outside of the construct (e.g., drift), or if the test fails to represent the intended construct. Standard 11.16: The level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for credential-worthy performance in the occupation (i.e., criterion-referenced). When there are alternate forms of a test, the cut score should refer to the same level of performance for all forms (e.g., stability of cut score).

Note: This list is not exhaustive – many more propositions could be made. Only the ones that are directly relevant to examining IPD for our claim that the observed score is reflective of the required knowledge and skills for licensure.

Figure 85. Example of Addressing IPD using the Standards.

As can be seen in Figure 85, each of the five sources of evidence occupy a column. Underneath each source of evidence are the requisite criteria that can be used to support the validity argument. Before unpacking each standard, it is important to note that this example only includes the standards that may be applicable to IPD. Additional claims must be made to defend the use of test scores for a specific purpose (e.g., the test content is representative of the domain of practice).

Four standards were listed under test content that have relevance to IPD. Standard 4.2 speaks to the test specifications, particularly ensuring that forms are built to the same statistical specifications and that items are presented in the same order as other forms.

Placing items in different locations on different forms could create context effects, especially if there are multiple item types. Expert judges can be used to review items as part of sensitivity review boards (Standard 4.8). Of primary relevance to drift is whether certain language may have changed over time. Standard 4.10 places responsibility on test developers for assembling or developing forms that contain items without DIF and IPD. Standard 4.24 suggests that test developers need to consider whether the content of the examination reflects the current domain of practice or whether the curriculum is outdated.

As it pertains to response processes, Standard 1.12 discusses procedures that can be taken when there is suspicion of cheating. Although cheating can be difficult to prove, an analysis of eye movements could reveal whether an examinee is looking over at another exam booklet or computer screen. Quick response times might also indicate that an item was not actually read, but that an examinee suspected of cheating looked at someone else's answers. Standard 6.6 expands upon 1.12 by justifying the use of technology to detect cheating through the use of similar erasure or answer patterns. Unusual item parameter shifts are a direct reference to drift and alludes to the fact that a drastic change in item difficulty could indicate that an item has been compromised.

Internal structure is the source of evidence for which IPD and DIF apply to the most. This is because detection of IPD and DIF occurs during scoring. Some of the particular standards that apply include Standard 5.6 – that the stability of the scale be maintained over time. Item banks should be recalibrated every time there is a new job analysis, but they may require more frequent updating. This would be true especially for medical or law fields, where certain items may behave differently due to new discoveries

or laws. Standard 5.15 speaks to drift within the context of linking and equating. The most commonly used data collection design is the CINEG, and the common items comprised in both forms should be statistically similar between each other, and as a microcosm of their own respective forms. Standard 5.19 is similar to Standard 4.2 in that it speaks specifically to context effects and the importance of item location, mindfulness of fatigue for longer tests, and consideration of adaptive tests (e.g., item overexposure rules, ensuring content balance). Standard 5.19 also speaks to ensuring that items promoted from the field-test stage exhibit the same properties as scored items.

A couple of standards can be placed under external relations. Standard 4.13 says that the test developer may consider correlational analyses to provide convergent evidence for constructs that are similar and discriminant evidence from dissimilar constructs. Standard 5.23 attests to the importance of ensuring that cut scores have clearly defined categories, possibly through empirical data relating test performance to a particular criterion. The standard also suggests that this information can be hard to obtain in credentialing, because these tests are not meant to be predictors of future performance. However, this information can be built into the standard setting process by ensuring that experts understand the difference between not competent, minimally competent, and competent. This understanding will help standard setting participants come up with a cut score that reflects minimal competence.

The last source of evidence concerns consequences arising from decisions made from test scores. Standard 1.25 touches upon investigating whether the score derived from the test reflects the construct of interest, or whether there is variance from another

source that is contributing to this score. One example would be if an exam contained only a few items that included a picture, re-testers may begin to get these items correct simply by memorizing them, as opposed to reflecting their knowledge of the exam. Standard 11.16 speaks to the nature of credentialing exams, that an examinee should pass if he or she meets the performance standard set by the standard setting. It also implies that even though the cut score might change depending upon form, that the equating procedure used will ensure that the performance standard remains the same for all examinees. This cannot be accomplished if there are items that exhibit IPD or DIF.

Extending Drift to Different Testing Contexts. This study has focused on the impact of drift for fixed form tests with dichotomously scored items. However, many other modalities of testing (e.g., computer-adaptive, multistage, web-based) with polytomously scored items are becoming more prevalent. Little research has been conducted on drift in these testing applications, with even less attention on validity. This section discusses how drift might affect the validity argument under different testing circumstances using a couple different examples.

When constructing a validity argument, the first step is always to make a claim for the intended use and interpretation of test scores. Once this claim has been made, the evidence needed to support the claim can be laid out. Readers are encouraged to consult the *Standards* five sources of validity evidence, Kane's argument-based approach (2006, 2013), or another validation framework that can be used to organize the evidence required to support their claim.

The use of passage-based or testlet-type items, a group of items with the same content, are most often observed in computer adaptive or multistage testing but may exist in other modalities. When presented in the context of IRT, practitioners should check that the assumption of local independence has not been violated. A violation of local independence could be attributed to an examinee having prior knowledge of a particular subject area; hence their responses would be reflective of their familiarity with the subject instead of their actual ability. One way to handle this issue is to use relevant subgroups in validity, reliability, and other studies when constructing the test (*Standards*, p. 64). Commentary from Standard 3.3 states that expert and sensitivity reviews can guard against construct-irrelevant context that may be more familiar to some than others (p. 64). Testlets, and other types of polytomously-scored items may also be used as anchor items for linking and equating purposes. It is important that these items be screened for drift, as inclusion of these items can have deleterious effects on linking and equating outcomes (Li, 2012).

Computer adaptive (CAT) or computer adaptive multistage testing (CA-MST) are two types of test modalities that have been increasingly used over the years. These tests are known for being more efficient in arriving at a test-taker's ability, but they also have unique challenges to consider. CAT or CA-MST's must ensure that examinees receive items from a different number of content areas, while also ensuring that the same items are not overexposed, which can lead to drift. This is reflected in Standard 4.3, which states that evidence should be documented for administration, scoring, and reporting rules in computer-adaptive and multistage-adaptive exams, including procedures for selecting

items or sets of items and controlling item exposure (*Standards*, p. 86). Procedures for selecting items are based upon the algorithms or item exposure controls used in the adaptive test. These algorithms and controls will work the best when there is a large item bank, or multiple item banks upon which to select items or item panels from (Luecht, 2014). When the algorithms are used to score complex examinee responses, theoretical and empirical rationale should be provided for responses at each score level (Standard 4.19, *Standards*, p. 91). For example, in a certification test, most of the items should be targeted at the performance standard so that a defensible decision (pass/fail) can be made about the examinee. However, in an educational placement test, where multiple cut scores are used to designate examinee performance, it is important for items to cover the entire range of the scale. Psychometrically defensible decisions need to be made for all examinees, ranging from "basic" to "advanced." Similar to Standard 4.19, Standard 5.16 also adds that the scores have comparable meaning over different sets of test items. Unlike fixed form tests, where items can be reviewed before or after administration, the same luxury does not apply to CATs because they are scored live. Instead, much of the effort to review items needs to be conducted to the item bank before it goes live (Luecht & Nungester, 1998). Therefore, additional evidence for the validity argument is required with CATs during the test design and development stage, with additional safeguards for automated algorithms and item exposure rates.

CHAPTER V

DISCUSSION

The current study examined the impact of IPD on five IRT linking methods. There were three major aims of this research: (1) to discuss implications of the impact of IPD in simulated and empirical datasets; (2) to identify which IRT linking method was most robust under different conditions of drift; and (3) to strengthen the validity argument made for the use of test scores by providing recommendations to practitioners confronted with IPD.

Study Findings and Conclusions

Several conclusions can be drawn from the results. First, the findings here are consistent with other studies, that have identified the magnitude of drifted items to have more of an effect on linking and equating outcomes than the proportion of drifted items (e.g., Kopp & Jones, 2020; Li, 2012; Risk, 2016). Studies examining the effect of drift using the Rasch model identified 0.50 logits as problematic (e.g., Draba, 1977; Kopp & Jones, 2020; Wright & Douglas; 1976). However, studies examining drift using the 3PL model have not specified a minimum threshold of drift magnitude that is problematic, though the recovery of parameter estimates and equated scores are negatively affected as drift increases (Hu et al., 2008; Jurich et al., 2012; Vukmirovic et al., 2003). Results from this study suggest that with an adequate sample size (3,000 per form), and a minimal

amount of drifted items (25% common items), 0.25 appears to be the threshold for which equated scores could begin to exceed the DTM threshold. This is due to the snowball effect that drift has on the linking constants, which affects the item parameter estimates, and then the equated scores. An increase in the magnitude of drift, or the proportion of drifted items (at the 0.50 magnitude), is subject to more severe consequences of one equated score point or more. However, the conditions of this simulation were unidirectional to represent a worst-case scenario. If confronted with multidirectional drift (as in the empirical analysis), the effects of drift on item parameter estimates may be washed out by the positive and negative values of drift. Furthermore, using a weighted RMSE may have provided more favorable equating results that would increase the threshold for which drift would exceed the DTM threshold.

Second, several studies have suggested that ability distributions have little effect on linking and equating outcomes when drift is present (e.g., He et al., 2015; Li, 2012; Witt et al., 2003). However, findings from this study indicate that ability distributions do have an impact on outcomes, consistent with other studies (e.g., Hu et al., 2008; Jurich et al., 2012). In some instances, the increase in ability distributions led to more profound effects in RMSE (i.e., linking constant *A*, linked difficulty estimates), and in other instances, the increase in ability distributions (mainly skewed) led to attenuated effects (i.e., linking constant *B*, equated scores). When the effects of drift were alleviated, a ceiling effect may have occurred, whereby the influence of drift did not benefit the examinees that already exhibited higher abilities (Jurich et al. 2012). In particular, FPC often produced the smallest values of RMSE of any linking methods under the N(1,1)

distribution. This was the most interesting finding, as most studies have found FPC to perform worse under differing ability distributions (e.g., Hu et al., 2008; Kim, 2006; Wollack et al., 2006). However, Keller and Keller (2011) found FPC was better at handling skewed ability distributions and was most suitable for changes in ability. Li et al. (1997) found that FPC produced more stable parameter estimates than SL under normal, negatively-skewed, and positively-skewed ability distributions.

Third, there was no single linking method that universally performed better than the others. But, the LAV method was the most consistent at returning the smallest RMSE values across linking constant *B*, item estimate *b*, and classification, especially for the highest magnitudes of drift (50% drifted items, -0.50 and -1.00 magnitudes of drift). On the other hand, the LAV method was most susceptible to linking constant *A* and item estimate *a*. Studies from the authors (i.e., He & Cui, 2020; He et al., 2015) found LAV to recover both linking constants and IRT equated true scores similar to or better than SL under the presence of one to three drifted items. This study was the first to compare the LAV to linking methods beyond SL, and with greater amounts of drifted items and magnitude. The LAV performed exceptionally well despite slightly elevated levels of SE, as values of RMSE and bias often remained unchanged. This may have been due to the weight function used to handle outliers. Items that exhibit drift are weighted less during the linking process thereby alleviating the impact of drift. Since drift was only manipulated in the difficulty parameter, the LAV may have accurately recovered the difficulty parameter at the expense of the discrimination parameter. More studies should

be conducted to evaluate its performance, and extend to usage with other models (e.g., Rasch, 2PL, GRM) and conditions.

Fourth, studies comparing SL to HB with and without drift have reported no difference between the two (e.g., Hanson & Beguin, 2002; Jurich et al., 2012; Keller & Keller, 2011; Kim & Kolen, 2007; Lee & Ban, 2010; Li et al., 2012; Sukin & Keller, 2008). These methods were not heavily reported on during this investigation because they were neither exceptionally good nor bad. In most conditions, the performance of the two methods was very similar, consistent with the aforementioned studies. Results from each research question are discussed in more detail in the following paragraphs.

Drift Detection. The focus of this study was not in examining drift detection methods, yet it was still important to determine whether the drift that was simulated was flagged and to what extent. Results from the likelihood ratio test indicated that when no drift was present, drift was detected no more than what would be expected based upon chance alone, which is consistent with several studies (e.g., DeMars, 2004b; Donoghue & Isham, 1998). The percentage of correct detections increased as the level of drift magnitude intensified, regardless of ability distribution, which was to be expected. The power to detect drift increased as sample size increased, which was also to be expected. However, drift detection slightly decreased as the ability of examinees increased. This is probably because examinees with high abilities were already expected to answer questions correct, regardless of whether or not the item drifted. But, most examinees (high and low ability) were able to answer these items correctly due to drift, which would lead to less discriminatory power and less detection power.

Results from the empirical analysis revealed eight (12%) of the 66 common items to drift. Five of these common items (63%) drifted easier, which is to be expected given that most occurrences of drift result in items becoming easier. This level of drift was most similar to the no drift or small drift conditions (25% drifted item, -0.25 magnitude). Best practice would be to review these items with the client and determine whether these items can be unanchored from the common item set. Further investigation should attempt to reveal reasons for why these items may have drifted.

Research Question 1: Linking Constants. Drift had a differential effect on the linking methods, such that the separate calibration methods underestimated linking constant *A*, while CC and FPC overestimated *A*. As drift increased in percentage of items and magnitude, greater values of bias and RMSE were observed. These findings for SL are consistent with Han (2008) and mostly consistent with Jurich et al. (2012), who found that SL underestimated linking constant *A* and was recovered less accurately (via bias and RMSE) as drift magnitude and the percentage of cheaters increased. Unlike this study, Li (2012) found that the RMSE of linking constant *A* did not change as the number of drifted items and magnitude of drift increased. Since *a*-drift was not manipulated, the findings from this study suggest that *b*-drift has the potential to affect linking constant *A*. This could occur due to a couple of reasons. One is that as the difficulty parameter changes due to drift, the variance increases and causes linking constant *A* to increase as well (Han, 2008). On the other hand, linking constant *A* could decrease because more examinees are answering items correctly, which diminishes the discriminatory power (discrimination estimate) of the anchor items on the new form (Jurich et al., 2012).

All linking methods tended to overestimate linking constant *B*, and values of bias and RMSE increased as drift magnitude increased, which is consistent with other studies examining unidirectional drift (Han, 2008; Jurich et al. 2012; Li, 2012). However, as both drift and ability increased, RMSE values decreased because there were fewer examinees that could benefit from the drift. That is, drift did not change the probability that examinees would get an item correct because they were already likely to answer the item correctly.

Overall, FPC and CC most accurately recovered linking constant *A*, while the LAV method most accurately recovered linking constant *B*, particularly at higher magnitudes of drift. Studies examining the LAV method (i.e., He & Cui, 2020; He et al., 2015) have found the LAV to perform similar to, or better than, the SL method in the presence of drift for both linking constants and IRT true score equating. By extension of this dissertation, the LAV method more accurately recovered linking constant *B* than the SL, HB, CC, and FPC methods. However, caution should be taken, as the linking constants derived from CC and FPC were based upon estimates from the new group ability distribution, which takes the performance on all items into account. The linking constants from the separate calibration methods were extracted only from the anchor items. Thus, the comparison was not identical between all linking methods.

Research Question 2: Linked Item Parameter Estimates. As mentioned by Han (2008), item parameter estimates can be directly affected from drift, or indirectly affected through the linking constants. It is no surprise then, that the findings for the recovery of the item estimates are consistent with those found for the linking constants.

Bias and RMSE values for both discrimination and difficulty increased as the percentage of drifted common items and drift magnitude increased, which is consistent with several studies (e.g., Han, 2008; Kopp & Jones, 2020; Li, 2012; Risk, 2016). CC and FPC most accurately recovered *a* and LAV most accurately recovered *b*, although FPC appeared to be most robust at the highest ability distributions – N(1,1) and S(1,1). To date, no studies have compared the performance of the LAV to FPC or CC, let alone the recovery of item estimates. So, there is no basis for comparison. However, Keller and Keller (2015) found that ability was recovered more accurately by CC and FPC compared to SL. The authors concluded that CC produced more stable results than SL. Chen (2013) found that FPC and SL performed better than CC in the recovery of theta, but CC was comparable when drifted items were removed from linking.

The finding that FPC was most robust to differing ability distributions is unexpected considering that studies (e.g., Hu et al., 2008; Kim, 2006; Wollack et al., 2006) have found FPC to be more sensitive to changes in ability distributions when recovering ability or item parameter estimates. Other studies have reported ability differences to have no effect on linking and equating (e.g., He et al., 2015; Li, 2012; Witt et al., 2003) although these studies did not observe ability differences greater than 0.60. However, Kim (2006) reported similar b-ARMSE values between the N(0, 1) and N(1, 1) distributions. Studies by Keller and Keller (2011) and Li et al. (1997) also provided support for the robustness of FPC under differing ability distributions.

Research Question 3. Equated Scores. Although conclusions can be drawn about which linking method most accurately recovered equated scores, every linking

method exceeded or nearly exceeded the DTM threshold of 0.5 for RMSE with the exception of the baseline condition – no drift, N(0,1) for the 3,000 sample size. These values increased as the proportion of drifted items and drift magnitude increased, which is consistent with other studies (e.g., Hu et al., 2008; Jurich et al., 2012; Kopp & Jones, 2020; Li, 2012; Risk, 2016). For practical purposes, these results would indicate that equated scores are subject to differences of one score point or more when ability distributions greatly differ and when drift is present. However, the recovery of equated scores would have improved if a weighted RMSE were used, which would increase the threshold for which drift affects equated scores. Nevertheless, values of bias were lowest for the LAV for nearly all equated scores except under N(1,1), where FPC typically yielded the smallest bias. These findings follow the same pattern as the results for the linking constants and item estimates.

Few studies have examined the LAV since it is a relatively new linking method introduced by He et al. (2015). Their findings (i.e., He & Cui, 2020; He et al., 2015) have indicated that the LAV produced RMSE and bias values smaller than SL for the recovery of linking constants and IRT true score equating in the presence of drift. Hu et al. (2008) found that CC and FPC recovered IRT true scores better than SL in the presence of drift when groups were equivalent; no linking method stood out in the presence of drift when groups were non-equivalent. Jurich et al. (2012) found no difference in the recovery of IRT true scores in the presence of drift for SL, HB, or FPC. Arce-Ferrer and Bulut (2017) found that SL produced equated cut scores with more precision than CC. The disparity in

findings illustrates how drift has a profound effect on equated scores and must be appropriately detected and removed.

Research Question 4. Classification Accuracy. Unlike equated scores, which were highly influenced by drift, all linking methods exhibited similar classification accuracy rates that were low in bias, SE, and RMSE. Only under the most extreme condition of drift (50% drifted items, -1.00 magnitude) did the LAV appear to retain a classification accuracy rate closer to the true classification rate. Sukin & Keller (2008) found no difference between the SL and HB methods for classification accuracy. Chen (2013) reported correct classification rates when drifted items were kept in the linking process for SL, CC, and FPC.

The accuracy rate was very high for all linking methods probably because most examinees were well over the cut score as a result of a higher ability and lower difficulty of the items. Had a positively skewed, or mean ability distribution lower than average (below 0), been introduced, then the accuracy rate may have declined because we would expect more examinees to be at the cut score. It is important to note that although the recovery of accuracy rates was robust, it does not imply that drift does not have an effect on classification. It simply means that drift pushed a lower-abled examinee to pass or that drift pulled a higher-abled examinee to fall below the cut score and fail on subsequent administrations of the test form.

Research Question 5. Classification Consistency. Similar to accuracy, classification consistency rates were recovered well for all linking methods. At the most extreme condition of drift, the LAV and HB methods yielded similar RMSE values that

were slightly better than the other linking methods. Keller & Keller (2015) found CC produced more accurate consistency rates than FPC and SL. Results from this study suggested that the performance of these three methods were comparable at both sample sizes and for all conditions. The one exception was at the highest magnitude of drift, where CC exhibited slightly higher RMSE values than either FPC or SL.

Empirical Analysis. Although the performance of each linking method could not be compared to true values, the linking methods yielded mostly similar linking constants, item parameter estimates, equated scores, and classification rates. As it pertains to drift detection, 12% of common items were flagged for drift using the Bonferonni correction. Without this adjustment, the percentage of items exhibiting drift would have tripled. Linking can be an iterative process that requires multiple runs when screening items. There is the possibility that some items can be erroneously flagged based upon chance alone. Thus, practitioners should be cautious when detecting drift, as drift can also go undetected or provide false positives (DeMars, 2004b; Donoghue & Isham, 1998).

Implications for Validity and Validation

This study has provided practitioners with recommendations for best practices when confronted with drifted items. Drifted items should always be removed unless there is reason not to (e.g., due to content imbalance or subject matter expert request). Additional efforts to identify the reason for drift should be investigated, as the reason might help to prevent future reoccurrences (e.g., changes in curriculum may inform teachers how much time to devote to each subject) and strengthen the validity argument

being made. Steps to construct a strong validity argument have been provided in the context of the *Standards'* five sources of evidence and Kane's argument-based approach.

Limitations and Directions for Future Research

This section discusses the limitations and directions for future studies. It should first be acknowledged that when linking outcomes become so undermined as to misconstrue results, choice of linking method is of little concern. While certain linking methods can alleviate some estimation error, such a distortion of results indicates larger problems with the test that may be traced back to the design and development of the exam. The best way to obtain accurate linking results and equated scores is by ensuring that all stages of test development are carried out thoroughly, securely, and with the utmost fidelity to the testing process.

Second, the conditions chosen for this study reflect conditions that might occur in the context of licensure and certification. However, most certification tests do not implement the 3PL model; rather, small candidate volumes are observed which restricts testing programs to using classical item statistics or the Rasch model. Similarly, the use of FPC is often implemented by testing programs as an effective and efficient way to maintain the item bank and pre-assemble forms. This study sought to identify a drift threshold for when linking and equating outcomes may become compromised, as no 3PL studies had suggested a threshold. It was also important to explore the effectiveness of different linking methods by stretching them to more extreme levels of ability, drift, and estimation. It would be desirable to learn the limitations of new methods by

understanding their performance in the context of drift, which is a salient, yet understudied phenomenon.

Third, this study focused purely upon dichotomously scored fixed-form tests. Although these tests are regularly observed in practice, newer technology enhanced item types and testing modalities are becoming more prevalent. While future research should focus upon different item types and test formats, the purpose of this research was to provide some clarity as to which linking method is most robust to drift since previous studies have not consistently identified one linking method.

Fourth, the ability distributions selected for this study were chosen to reflect those that might be found in the context of licensure – higher or negatively skewed distributions. These examinees are often more abled, but they do not reflect ability distributions from educational contexts, which often fill the entire range of the scale or could potentially be positively skewed. By placing more examinees at the lower ends of the scale, there would have been more variability in the findings for classification accuracy and consistency. Most of the examinees from the simulation already possessed high abilities, and the candidates from the empirical analysis were presented with very easy items.

Fifth, data from the empirical analysis was originally calibrated using the Rasch model available in Winsteps. Unlike Winsteps, which centers the mean of the items to 0, flexMIRT was used to reestimate the data using a 3PL model. Although the 3PL model was fitted to the data, it would not have been advisable to use the 3PL model with this program because the sample size of roughly 2,000 candidates per form was accumulated

over the course of several years. In order to use the 3PL model for this client, this sample size would need to be obtained within a year or faster so that experimental items could be linked to the item bank and used to assemble future forms that are published every year or sooner. There were also several items that had difficulty values lower than -3 units, which would almost never be administered as scored items.

Sixth, linking constants for CC and FPC were extracted using the performance of examinees on all items of the new form. In reality, the linking constants are extracted using only the common items, which was done for the separate calibration methods. These linking constants are then used to transform the new form parameter estimates for all items onto the scale of the base form. However, applying the linking constants from CC and FPC to transform the new form would not make sense because the new form estimates are already on the same scale of the base form. That is because the linking constants are A=1 and B=0. While it was important to try and compare linking constants from CC and FPC to the separate calibration methods, the comparison between the linking methods was not one to one.

Finally, this study considered the impact of drift when items were not removed for exhibiting drift. This was examined intentionally because drift can go undetected or operate in ways unbeknownst to practitioners. However, the lack of purification is a limitation because items should be removed for drift. Thus, future studies should examine the impact of drift for these linking methods when items are removed from the anchor item set.

REFERENCES

Accreditation Council for Graduate Medical Education. (2018). Program requirements for graduate medical education in internal medicine. Retrieved from http://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/140_internal_medicine_2017-07-01.pdf?ver=2017-06-30-083345-723.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arce-Ferrer, A. J., & Bulut, O. (2017). Investigating separate and concurrent approaches for item parameter drift in 3pl item response theory equating. *International Journal of Testing, 17*(1), 1-22.

Azzalini, A. (2020). The R package 'sn': The skew-normal and related distributions such as the skew-t (version 1.6-2).

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement, 36*(7), 565-580.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*(2), 147-162.

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, *68*(1), 1-22.

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

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L' Année Psychologie*, 11, 191-336.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *Bmj*, *310*(6973), 170.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*(4), 275-285.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791-799.

Breitbach, A. P., Downey, D. L., & Frager, A. J. (2013). The relationship between candidate psychological factors and first-attempt pass rate on the board of certification examination. *Athletic Training Education Journal*, *8*(1-2), 10-16.

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.

Cabin, R. J., & Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: When and how are the questions. *Bulletin of the Ecological Society of America*, *81*(3), 246-248.

Cai, L. (2017). flexMIRT® version 3.51. Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational measurement: Issues and practice*, *29*(1), 3-13.

Chen, Q. (2013). Remove or keep: Linking items showing item parameter drift (Doctoral dissertation). *Retrieved from ProQuest Dissertations and Theses database*. (*UMI No. 3560290*).

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

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

DeMars, C. E. (2004a). *Item parameter drift: The impact of curricular area*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

DeMars, C. E. (2004b). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*(3), 265-300.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item

parameter drift. *Applied Psychological Measurement*, *22*(1), 33-51.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. *Technical issues related to the introduction of the new SAT and PSAT/NMSQT*, 91-122.

Draba, R. (1977). *The identification and interpretation of item bias*. Research Memorandum No. 25, Statistical Laboratory, Department of Education, University of Chicago

Gaertner, M. N., & Briggs, D. C. (2009). *Detecting and addressing item parameter drift in IRT test equating contexts*. Unpublished manuscript. University of Colorado at Boulder.

Glas, G. A. (2000). Item calibration and parameter drift. In *Computerized adaptive testing: Theory and practice* (pp. 183-199). Springer.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 20(4), 369-377.

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

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27.

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

Han, K. T. (2008). Impact of item parameter drift on test equating and proficiency estimates (Doctoral dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3325324).

Han, K. T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (R-11-02). Graduate Management Admission Council Research Report.

Han, K. T., Wells, C. S., & Sireci, S. G. (2012). The impact of multidirectional item parameter drift on IRT scaling coefficients and proficiency estimates. *Applied Measurement in Education*, *25*(2), 97-117.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3-24.

Hanson, B., & Zeng, L. (1995). PIE: A computer program for IRT equating (Windows Console Version, Revised by Z. Cui, May 20, 2004) [Manual]. Unpublished manuscript, College of Education. University of Iowa, Iowa City

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

He, Y., & Cui, Z. (2020). Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Applied Psychological Measurement*, *44*(4), 296-310.

He, Y., Cui, Z., & Osterlind, S. J. (2015). New robust scale transformation methods in the

presence of outlying common items. *Applied Psychological Measurement*, *39*(8), 613-626.

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement*, *32*(4), 311-333.

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

Huggins-Manley, A. C. (2017). Psychometric consequences of subpopulation item parameter drift. *Educational and Psychological Measurement*, *77*(1), 143-164.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6*(3), 249-260.

Huynh, H, & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research and Evaluation*, *15*(2), 1-8.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education*, *71*(3), 229-250.

Jones, L. V. (1960). Some invariant finding under the method of successive intervals. In

H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications* (pp. 7-20). John Wiley & Sons.

Jurich, D. P., DeMars, C. E., & Goodman, J. T. (2012). Investigating the impact of compromised anchor items on IRT equating under the nonequivalent anchor test design. *Applied Psychological Measurement*, *36*(4), 291-308.

Kabasakal, K. Al, & Kelecioglu, H. (2015). Effect of differential item functioning on test equating. *Educational Sciences: Theory and Practice*, *15*(5), 1229-1246.

Kane, M. T. (1982). The validity of licensure examinations. *American Psychologist*, *37*(8), 911.

Kane, M. T. (2006). Validation. *Educational Measurement*, *4*(2), 17-64.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kang, T., & Petersen, N. S. (2011). Linking item parameters to a base scale. *Asia Pacific Education Review*, *13*(2), 311-321.

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. *Educational and Psychological Measurement*, *71*(2), 362-379.

Keller, L. A., & Keller, R. R. (2015). The effect of changing content on IRT scaling methods. *Applied Measurement in Education*, *28*(2), 99-114.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of

multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212-228.

Kim, K. Y. (2019). A comparison of the separate and concurrent calibration methods for the full-information bifactor model. *Applied Psychological Measurement*, *43*(7), 512-526.

Kim, K. Y., & Lee, W. C. (2017). The impact of three factors on the recovery of item parameters for the three-parameter logistic model. *Applied Measurement in Education*, *30*(3), 228-242.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355-381.

Kim, S., & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*(4), 371-397.

Kim, S., & Lee, W. C. (2004). *IRT scale linking methods for mixed-format tests*. ACT.

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

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

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for

IRT equating and adaptive testing. *Applied Psychological Measurement, 8*(2), 147-154.

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

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Kopp, J. P., & Jones, A. T. (2020). Impact of item parameter drift on rasch scale stability in small samples over multiple administrations. *Applied Measurement in Education*, *33*(1), 24-33.

Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1-17.

Lee, W. C., & Ban, J. C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education*, *23*(1), 23-48.

Lee, W. C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*(4), 412-432.

Li, D. Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Measurement*, *49*(2), 167-189.

Li, Y. (2012). *Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating*. (Research Report Series No. RR-12-09). Educational Testing Service.

Li, Y. H., Griffith, W. D., & Tam, H. P. (1997). *Equating multiple tests via an IRT linking design: Utilizing a single set of anchor items with fixed common item parameters during the calibration process*. Paper presented at the annual meeting of the Psychometric Society, Knoxville, TN.

Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linn, R. L. (1990). Has item response theory increased the validity of achievement test scores? *Applied Measurement in Education*, *3*(2), 115-141.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*(4), 453-461.

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

Luecht, R. M. (2014). Computer-adaptive testing. *Encyclopedia of Statistics in Behavioral Science*. Wiley.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive

sequential testing. *Journal of Educational Measurement, 35*(3), 229-249.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software, 48*(8), 1-31.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *ETS Research Bulletin Series, 1977*(1), i-41.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). Macmillan.

O'Neill, T., Peabody, M., Tan, R. J. B., & Du, Y. (2013). How much item drift is too much. *Rasch Measurement Transactions*, *27*(3), 1423-1424.

Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement*, *25*(4), 373-383.

Okrainec, A., Soper, N. J., Swanstrom, L. L., & Fried, G. M. (2011). Trends and results of the first 5 years of fundamentals of laparoscopic surgery (FLS) certification testing. *Surgical Endoscopy*, *25*(4), 1192-1198.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*(2), 223-241.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156.

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/.

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

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(1), 19-31.

Risk, N. M. (2016). The impact of item parameter drift in computer adaptive testing (CAT). *Journal of Applied Measurement*, *17*(1), 54-78.

Rupp, A. A., & Zumbo, B. D. (2003a). Bias coefficients for lack of invariance in unidimensional IRT models. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Rupp, A. A., & Zumbo, B. D. (2003b). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *The Alberta Journal of Educational Research*, *49*(3), 264-276.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63-84.

Shea, J. A., Norcini, J. J., Day, S. C., & Benson, J. A. (1991). A longitudinal description

of patterns of certification in internal medicine and the subspecialties. *Journal of General Internal Medicine*, *6*(6), 553-557.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Smith, R. W. (2004). The impact of braindump sites on item exposure and item parameter drift. Paper presented at the annual meeting of the American Education Research Association, San Diego, CA.

Stahl, J., & Muckle, T. (2007). Investigating drift displacement in rasch item calibrations. *Rasch Measurement Transactions*, *21*, 3.

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

Stone, C. A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education*, *4*(2), 125-141.

Sukin, T., & Keller, L. (2008). The effect of deleting anchor on the classification of examinees. *NERA Conference Proceedings*, 19, 2-14.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), *New Horizons in Testing* (pp. 13-30). Academic Press.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT b values. *Journal of Educational Measurement*, *29*(3), 201-211.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247-260.

Thurstone, L. L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology*, *18*(8), 505-524.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Uysal, İ., & Kilmen, S. (2016). Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences*, *8*(2), 1-11.

Veerkamp, W., & Glas, C. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, *25*(4), 373-389.

Vukmirovic, Z., Hu, H., & Turner, J. C. (2003). *The effects of outliers on IRT equating with fixed common item parameters*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

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

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*(3), 347-364.

Witt, E. A., Stahl, J. A., Bergstrom, B. A., & Muckle, T. (2003). *Impact of item drift with*

non-normal distributions. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wollack, J. A., Sung, H. J., & Kang, T. (2005). *Longitudinal effects of item parameter drift*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, CA.

Wollack, J. A., Sung, H. J., & Kang, T. (2006). *The impact of compounding item parameter drift on ability estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wright, B. D. & Douglas, G. A. (1976). *Rasch item analysis by hand*. Research Memorandum No. 21, Statistical Laboratory, Department of Education, University of Chicago.

Wyse, A. E., & Babcock, B. (2016). How does calibration timing and seasonality affect item parameter estimates. *Educational and Psychological Measurement*, *76*(3), 508-527.

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement*, *36*(7), 602-624.

Yang, W. L. (2000). *The effects of content homogeneity and equating method on the accuracy of common-item test equating*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Yurtçu, M., & Güzeller, C. O. (2018). Investigation of equating error in tests with

differential item functioning. *International Journal of Assessment Tools in Education*, *5*(1), 50-57.

APPENDIX A

GENERATING ITEM PARAMETERS

Base Form				New Form			
Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
1	0.913	0.721	0.340	1	0.913	0.721	0.340
2	0.853	0.700	0.340	2	0.853	0.700	0.340
3	1.203	-0.575	0.315	3	1.203	-0.575	0.315
4	1.208	-0.154	0.059	4	1.208	-0.154	0.059
5	1.091	-1.733	0.247	5	1.091	-1.733	0.247
6	1.133	-1.722	0.347	6	1.133	-1.722	0.347
7	0.931	1.077	0.246	7	0.931	1.077	0.246
8	0.507	0.324	0.286	8	0.507	0.324	0.286
9	1.492	-0.906	0.264	9	1.492	-0.906	0.264
10	1.489	0.440	0.256	10	1.489	0.440	0.256
11	0.917	-0.202	0.305	11	0.917	-0.202	0.305
12	1.489	-0.013	0.060	12	1.489	-0.013	0.060
13	0.963	1.365	0.055	13	0.963	1.365	0.055
14	0.838	-0.592	0.227	14	0.838	-0.592	0.227
15	1.111	-0.394	0.208	15	1.111	-0.394	0.208
16	0.551	1.538	0.199	16	0.551	1.538	0.199
17	0.767	-0.198	0.345	17	0.767	-0.198	0.345
18	0.723	0.593	0.120	18	0.723	0.593	0.120
19	1.118	-0.329	0.295	19	1.118	-0.329	0.295
20	0.947	0.442	0.154	20	0.947	0.442	0.154
21	1.269	-0.502	0.171	21	1.454	1.371	0.179
22	0.985	0.132	0.203	22	1.077	-0.565	0.168
23	0.596	-0.079	0.207	23	1.027	0.363	0.093
24	0.501	0.887	0.348	24	0.964	0.633	0.134
25	1.213	0.117	0.179	25	0.642	0.404	0.219
26	0.953	0.319	0.349	26	1.184	-0.106	0.331
27	1.065	-0.582	0.286	27	0.935	1.512	0.158
28	1.245	0.715	0.205	28	0.945	-0.095	0.303
29	1.498	-0.825	0.201	29	1.280	2.018	0.267
30	0.969	-0.360	0.323	30	1.247	-0.063	0.275
31	0.833	0.090	0.129	31	1.418	1.305	0.327
32	1.428	0.096	0.102	32	0.857	2.287	0.051
33	0.732	-0.202	0.170	33	1.195	-1.389	0.098
34	0.653	0.740	0.212	34	1.417	-0.279	0.170
35	0.841	0.123	0.123	35	0.667	-0.133	0.253
36	1.494	-0.029	0.163	36	0.742	0.636	0.194
37	0.750	-0.389	0.224	37	0.660	-0.284	0.210
38	1.124	0.511	0.113	38	0.562	-2.656	0.145
39	0.646	-0.914	0.290	39	1.024	-2.440	0.294
40	0.648	2.310	0.242	40	1.196	1.320	0.138

Base Form				New Form			
Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
41	0.900	-0.438	0.271	41	1.360	-0.307	0.173
42	1.409	0.764	0.182	42	1.313	-1.781	0.077
43	0.859	0.262	0.224	43	0.699	-0.172	0.290
44	1.253	0.773	0.127	44	1.495	1.215	0.158
45	0.563	-0.814	0.189	45	0.800	1.895	0.062
46	0.880	-0.438	0.101	46	1.032	-0.430	0.062
47	0.767	-0.720	0.234	47	0.873	-0.257	0.336
48	0.889	0.231	0.337	48	0.963	-1.763	0.162
49	1.372	-1.158	0.193	49	1.056	0.460	0.292
50	0.968	0.247	0.275	50	1.036	-0.640	0.323
51	1.052	-0.091	0.056	51	0.992	0.455	0.182
52	1.076	1.757	0.101	52	1.032	0.705	0.223
53	0.816	-0.138	0.242	53	0.854	1.035	0.072
54	0.571	-0.111	0.099	54	0.849	-0.609	0.099
55	0.901	-0.690	0.156	55	0.502	0.505	0.272
56	1.039	-0.222	0.106	56	0.885	-1.717	0.193
57	1.305	0.183	0.319	57	0.846	-0.784	0.256
58	0.923	0.417	0.121	58	1.491	-0.851	0.335
59	0.909	1.065	0.345	59	0.591	-2.414	0.199
60	1.485	0.970	0.056	60	1.041	0.036	0.191
61	0.768	-0.102	0.082	61	0.552	0.206	0.218
62	1.127	1.403	0.123	62	0.559	-0.361	0.246
63	0.825	-1.777	0.267	63	1.037	0.758	0.134
64	1.125	0.623	0.060	64	0.701	-0.727	0.344
65	0.536	-0.522	0.215	65	0.999	-1.368	0.243
66	0.844	1.322	0.255	66	0.872	0.433	0.225
67	0.916	-0.363	0.140	67	0.816	-0.811	0.235
68	1.302	1.319	0.167	68	0.503	1.444	0.328
69	0.859	0.044	0.270	69	0.633	-0.431	0.167
70	1.089	-1.879	0.339	70	1.054	0.656	0.136
71	0.875	-0.447	0.278	71	1.170	0.322	0.077
72	0.745	-1.739	0.225	72	0.852	-0.784	0.147
73	1.207	0.179	0.189	73	1.000	1.576	0.277
74	0.862	1.897	0.157	74	1.337	0.643	0.081
75	1.404	-2.272	0.165	75	1.432	0.090	0.263
76	1.133	0.980	0.112	76	0.671	0.277	0.340
77	0.955	-1.399	0.092	77	0.965	0.679	0.110
78	1.137	1.825	0.167	78	1.360	0.090	0.083
79	0.988	1.381	0.130	79	0.859	-2.993	0.067
80	1.137	-0.839	0.261	80	0.984	0.285	0.299

Base Form				New Form			
Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
81	0.877	-0.262	0.172	81	0.974	-0.367	0.224
82	0.509	-0.069	0.130	82	0.734	0.185	0.191
83	1.047	-0.379	0.170	83	0.867	0.582	0.160
84	1.198	2.582	0.109	84	0.991	1.400	0.134
85	0.705	0.130	0.299	85	0.876	-0.727	0.230
86	0.666	-0.713	0.208	86	1.334	1.303	0.296
87	0.869	0.638	0.169	87	0.856	0.336	0.079
88	0.845	0.202	0.222	88	0.870	1.039	0.339
89	1.126	-0.070	0.341	89	1.209	0.921	0.101
90	1.040	-0.092	0.245	90	0.683	0.721	0.076
91	1.310	0.449	0.150	91	0.988	-1.043	0.308
92	1.496	-1.064	0.149	92	0.535	-0.090	0.207
93	0.995	-1.162	0.333	93	1.350	0.624	0.247
94	0.993	1.649	0.164	94	0.918	-0.954	0.119
95	1.075	-2.062	0.219	95	0.860	-0.543	0.266
96	0.899	0.013	0.203	96	0.629	0.581	0.197
97	0.966	-1.088	0.092	97	0.998	0.768	0.340
98	0.970	0.271	0.122	98	0.760	0.464	0.322
99	1.079	1.008	0.265	99	0.840	-0.886	0.215
100	1.042	-2.074	0.139	100	1.386	-1.100	0.073

Note: The first 20 highlighted rows are common items

APPENDIX B

BIAS, SE, RMSE VALUES FOR LINKING CONSTANTS

Bias for Linking Constant A - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.002	-0.034	-0.070	-0.011	-0.025
	25%	-0.25	-0.012	-0.048	-0.084	-0.015	-0.040
		-0.50	-0.030	-0.062	-0.097	-0.036	-0.049
		-1.00	-0.061	-0.103	-0.138	-0.069	-0.088
	50%	-0.25	-0.024	-0.052	-0.095	-0.029	-0.039
		-0.50	-0.039	-0.077	-0.127	-0.044	-0.073
		-1.00	-0.102	-0.144	-0.192	-0.105	-0.136
HB	None	None	0.006	-0.027	-0.061	-0.004	-0.015
	25%	-0.25	-0.016	-0.049	-0.077	-0.015	-0.035
		-0.50	-0.047	-0.074	-0.103	-0.047	-0.058
		-1.00	-0.115	-0.146	-0.176	-0.118	-0.130
	50%	-0.25	-0.032	-0.057	-0.094	-0.034	-0.040
		-0.50	-0.068	-0.099	-0.141	-0.067	-0.093
		-1.00	-0.187	-0.220	-0.258	-0.181	-0.206
LAV	None	None	0.006	-0.024	-0.056	-0.006	-0.013
	25%	-0.25	-0.013	-0.046	-0.071	-0.013	-0.033
		-0.50	-0.035	-0.061	-0.090	-0.036	-0.051
		-1.00	-0.032	-0.078	-0.111	-0.061	-0.081
	50%	-0.25	-0.032	-0.052	-0.085	-0.037	-0.036
		-0.50	-0.075	-0.103	-0.142	-0.075	-0.098
		-1.00	-0.228	-0.252	-0.261	-0.227	-0.240
CC	None	None	0.128	0.106	0.097	0.096	0.110
	25%	-0.25	0.114	0.094	0.099	0.095	0.102
		-0.50	0.094	0.087	0.092	0.081	0.103
		-1.00	0.061	0.060	0.080	0.059	0.089
	50%	-0.25	0.100	0.094	0.090	0.084	0.108
		-0.50	0.085	0.083	0.085	0.078	0.093
		-1.00	0.032	0.056	0.084	0.052	0.088
FPC	None	None	0.077	0.054	0.044	0.047	0.052
	25%	-0.25	0.062	0.042	0.042	0.045	0.042
		-0.50	0.042	0.034	0.033	0.026	0.040
		-1.00	0.014	0.006	0.016	0.002	0.019
	50%	-0.25	0.046	0.040	0.033	0.032	0.047
		-0.50	0.033	0.026	0.020	0.023	0.027
		-1.00	-0.019	-0.012	-0.001	-0.021	0.001

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Linking Constant A - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.054	0.049	0.046	0.047	0.048
	25%	-0.25	0.047	0.045	0.049	0.046	0.050
		-0.50	0.050	0.041	0.046	0.047	0.047
		-1.00	0.043	0.046	0.046	0.040	0.049
	50%	-0.25	0.045	0.043	0.044	0.046	0.053
		-0.50	0.044	0.045	0.048	0.044	0.051
		-1.00	0.042	0.044	0.042	0.040	0.048
HB	None	None	0.052	0.046	0.045	0.046	0.046
	25%	-0.25	0.045	0.042	0.048	0.044	0.047
		-0.50	0.046	0.038	0.046	0.043	0.047
		-1.00	0.040	0.043	0.044	0.037	0.046
	50%	-0.25	0.042	0.043	0.042	0.043	0.053
		-0.50	0.041	0.043	0.047	0.042	0.050
		-1.00	0.037	0.041	0.042	0.036	0.045
LAV	None	None	0.055	0.052	0.048	0.052	0.052
	25%	-0.25	0.051	0.049	0.056	0.050	0.050
		-0.50	0.053	0.052	0.054	0.053	0.052
		-1.00	0.053	0.056	0.060	0.052	0.055
	50%	-0.25	0.046	0.048	0.049	0.045	0.056
		-0.50	0.050	0.053	0.054	0.048	0.059
		-1.00	0.056	0.055	0.056	0.048	0.049
CC	None	None	0.049	0.041	0.040	0.038	0.039
	25%	-0.25	0.044	0.040	0.038	0.038	0.039
		-0.50	0.043	0.038	0.037	0.039	0.041
		-1.00	0.041	0.039	0.036	0.034	0.038
	50%	-0.25	0.039	0.037	0.039	0.039	0.040
		-0.50	0.040	0.038	0.036	0.036	0.043
		-1.00	0.037	0.038	0.034	0.033	0.040
FPC	None	None	0.051	0.044	0.044	0.041	0.043
	25%	-0.25	0.046	0.042	0.044	0.041	0.045
		-0.50	0.044	0.041	0.043	0.041	0.046
		-1.00	0.042	0.042	0.040	0.036	0.041
	50%	-0.25	0.041	0.039	0.042	0.041	0.045
		-0.50	0.041	0.043	0.039	0.039	0.045
		-1.00	0.037	0.039	0.039	0.035	0.041

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Linking Constant A - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.054	0.059	0.083	0.048	0.054
	25%	-0.25	0.049	0.065	0.097	0.048	0.064
		-0.50	0.058	0.074	0.107	0.059	0.068
		-1.00	0.075	0.112	0.145	0.080	0.101
	50%	-0.25	0.051	0.067	0.105	0.054	0.066
		-0.50	0.058	0.089	0.135	0.062	0.090
		-1.00	0.111	0.151	0.196	0.113	0.144
HB	None	None	0.052	0.054	0.076	0.047	0.049
	25%	-0.25	0.048	0.064	0.091	0.046	0.058
		-0.50	0.066	0.083	0.112	0.063	0.075
		-1.00	0.122	0.153	0.181	0.123	0.138
	50%	-0.25	0.053	0.071	0.103	0.055	0.067
		-0.50	0.079	0.108	0.149	0.079	0.105
		-1.00	0.190	0.224	0.261	0.185	0.211
LAV	None	None	0.056	0.058	0.074	0.053	0.054
	25%	-0.25	0.052	0.067	0.091	0.051	0.060
		-0.50	0.063	0.080	0.105	0.064	0.073
		-1.00	0.062	0.096	0.126	0.081	0.098
	50%	-0.25	0.056	0.071	0.098	0.058	0.067
		-0.50	0.090	0.116	0.152	0.090	0.114
		-1.00	0.234	0.258	0.267	0.232	0.245
CC	None	None	0.137	0.114	0.105	0.103	0.117
	25%	-0.25	0.122	0.102	0.106	0.102	0.109
		-0.50	0.103	0.095	0.099	0.090	0.111
		-1.00	0.074	0.071	0.088	0.068	0.097
	50%	-0.25	0.107	0.100	0.098	0.093	0.115
		-0.50	0.094	0.092	0.093	0.086	0.103
		-1.00	0.048	0.068	0.091	0.061	0.096
FPC	None	None	0.092	0.070	0.063	0.062	0.068
	25%	-0.25	0.078	0.060	0.061	0.061	0.062
		-0.50	0.061	0.053	0.054	0.049	0.061
		-1.00	0.044	0.042	0.043	0.036	0.046
	50%	-0.25	0.062	0.056	0.054	0.053	0.065
		-0.50	0.053	0.050	0.044	0.045	0.053
		-1.00	0.042	0.040	0.039	0.041	0.041

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Linking Constant B - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.006	-0.032	-0.089	-0.018	-0.074
	25%	-0.25	0.067	0.023	-0.045	0.034	-0.014
		-0.50	0.125	0.078	0.002	0.097	0.031
		-1.00	0.229	0.169	0.072	0.180	0.100
	50%	-0.25	0.126	0.072	0.001	0.086	0.031
		-0.50	0.234	0.173	0.091	0.192	0.125
		-1.00	0.442	0.348	0.231	0.373	0.262
HB	None	None	0.005	-0.034	-0.088	-0.021	-0.075
	25%	-0.25	0.064	0.017	-0.051	0.026	-0.024
		-0.50	0.117	0.060	-0.019	0.079	0.006
		-1.00	0.196	0.114	-0.003	0.121	0.024
	50%	-0.25	0.121	0.063	-0.009	0.077	0.017
		-0.50	0.222	0.149	0.058	0.165	0.089
		-1.00	0.395	0.266	0.122	0.292	0.148
LAV	None	None	0.007	-0.035	-0.084	-0.019	-0.075
	25%	-0.25	0.051	0.007	-0.054	0.020	-0.029
		-0.50	0.066	0.018	-0.053	0.037	-0.029
		-1.00	0.055	0.009	-0.071	0.017	-0.054
	50%	-0.25	0.120	0.061	-0.005	0.073	0.021
		-0.50	0.197	0.125	0.045	0.141	0.065
		-1.00	0.313	0.167	0.025	0.167	0.028
CC	None	None	-0.038	0.004	0.047	0.033	0.065
	25%	-0.25	0.023	0.062	0.099	0.087	0.127
		-0.50	0.085	0.125	0.158	0.158	0.188
		-1.00	0.200	0.246	0.276	0.269	0.303
	50%	-0.25	0.088	0.117	0.152	0.144	0.183
		-0.50	0.204	0.235	0.272	0.265	0.301
		-1.00	0.450	0.477	0.517	0.515	0.535
FPC	None	None	-0.018	-0.004	0.010	0.013	0.014
	25%	-0.25	0.044	0.053	0.056	0.066	0.073
		-0.50	0.103	0.109	0.106	0.128	0.123
		-1.00	0.206	0.203	0.186	0.211	0.201
	50%	-0.25	0.108	0.105	0.107	0.122	0.127
		-0.50	0.220	0.212	0.206	0.230	0.226
		-1.00	0.430	0.397	0.374	0.416	0.383

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Linking Constant B - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.048	0.052	0.049	0.049	0.057
	25%	-0.25	0.053	0.050	0.055	0.057	0.052
		-0.50	0.050	0.049	0.055	0.051	0.054
		-1.00	0.053	0.047	0.058	0.052	0.057
	50%	-0.25	0.047	0.053	0.057	0.052	0.055
		-0.50	0.050	0.053	0.056	0.056	0.062
		-1.00	0.048	0.054	0.051	0.051	0.059
HB	None	None	0.048	0.052	0.050	0.050	0.059
	25%	-0.25	0.050	0.049	0.055	0.056	0.054
		-0.50	0.050	0.049	0.057	0.053	0.056
		-1.00	0.052	0.047	0.058	0.052	0.057
	50%	-0.25	0.048	0.051	0.058	0.053	0.060
		-0.50	0.050	0.054	0.057	0.059	0.064
		-1.00	0.050	0.054	0.059	0.047	0.065
LAV	None	None	0.051	0.055	0.057	0.053	0.064
	25%	-0.25	0.054	0.057	0.059	0.065	0.061
		-0.50	0.054	0.056	0.065	0.057	0.058
		-1.00	0.058	0.052	0.061	0.057	0.059
	50%	-0.25	0.048	0.057	0.066	0.054	0.067
		-0.50	0.058	0.062	0.070	0.074	0.075
		-1.00	0.081	0.071	0.068	0.070	0.065
CC	None	None	0.050	0.053	0.052	0.049	0.061
	25%	-0.25	0.052	0.052	0.059	0.057	0.052
		-0.50	0.052	0.050	0.060	0.053	0.056
		-1.00	0.059	0.048	0.062	0.057	0.060
	50%	-0.25	0.048	0.054	0.057	0.053	0.055
		-0.50	0.053	0.057	0.058	0.059	0.062
		-1.00	0.057	0.060	0.056	0.055	0.067
FPC	None	None	0.049	0.053	0.055	0.050	0.062
	25%	-0.25	0.053	0.052	0.058	0.057	0.054
		-0.50	0.050	0.050	0.062	0.052	0.055
		-1.00	0.055	0.049	0.063	0.057	0.062
	50%	-0.25	0.048	0.053	0.059	0.052	0.058
		-0.50	0.052	0.057	0.059	0.059	0.065
		-1.00	0.055	0.059	0.059	0.055	0.065

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Linking Constant B - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.049	0.061	0.101	0.052	0.093
	25%	-0.25	0.085	0.055	0.071	0.066	0.054
		-0.50	0.134	0.092	0.055	0.110	0.062
		-1.00	0.236	0.176	0.092	0.187	0.116
	50%	-0.25	0.135	0.089	0.057	0.101	0.063
		-0.50	0.240	0.181	0.106	0.200	0.139
		-1.00	0.445	0.352	0.237	0.376	0.269
HB	None	None	0.049	0.062	0.101	0.055	0.095
	25%	-0.25	0.081	0.052	0.075	0.062	0.059
		-0.50	0.127	0.078	0.060	0.095	0.056
		-1.00	0.202	0.123	0.058	0.132	0.062
	50%	-0.25	0.130	0.081	0.059	0.093	0.062
		-0.50	0.228	0.159	0.081	0.175	0.109
		-1.00	0.399	0.271	0.135	0.295	0.162
LAV	None	None	0.052	0.065	0.101	0.056	0.098
	25%	-0.25	0.075	0.057	0.080	0.068	0.068
		-0.50	0.085	0.058	0.083	0.068	0.065
		-1.00	0.080	0.053	0.093	0.060	0.080
	50%	-0.25	0.130	0.084	0.066	0.091	0.070
		-0.50	0.206	0.140	0.084	0.159	0.100
		-1.00	0.324	0.182	0.072	0.181	0.071
CC	None	None	0.063	0.053	0.070	0.059	0.089
	25%	-0.25	0.057	0.081	0.115	0.104	0.138
		-0.50	0.100	0.135	0.169	0.167	0.196
		-1.00	0.208	0.250	0.283	0.275	0.309
	50%	-0.25	0.100	0.129	0.162	0.154	0.191
		-0.50	0.211	0.242	0.278	0.272	0.307
		-1.00	0.454	0.481	0.520	0.518	0.539
FPC	None	None	0.053	0.053	0.055	0.051	0.063
	25%	-0.25	0.068	0.075	0.081	0.087	0.091
		-0.50	0.115	0.120	0.123	0.138	0.135
		-1.00	0.213	0.208	0.197	0.219	0.210
	50%	-0.25	0.118	0.118	0.123	0.133	0.139
		-0.50	0.226	0.220	0.214	0.237	0.235
		-1.00	0.434	0.401	0.379	0.420	0.389

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Linking Constant A - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	-0.003	-0.026	-0.059	-0.002	-0.003
	25%	-0.25	-0.011	-0.036	-0.074	-0.011	-0.012
		-0.50	-0.026	-0.050	-0.082	-0.019	-0.020
		-1.00	-0.061	-0.091	-0.128	-0.048	-0.050
	50%	-0.25	-0.016	-0.044	-0.079	-0.013	-0.014
		-0.50	-0.039	-0.068	-0.107	-0.032	-0.030
		-1.00	-0.097	-0.128	-0.169	-0.076	-0.088
HB	None	None	0.000	-0.019	-0.049	0.002	0.004
	25%	-0.25	-0.015	-0.037	-0.070	-0.015	-0.012
		-0.50	-0.044	-0.065	-0.091	-0.035	-0.033
		-1.00	-0.115	-0.140	-0.167	-0.101	-0.097
	50%	-0.25	-0.025	-0.050	-0.078	-0.020	-0.018
		-0.50	-0.068	-0.092	-0.124	-0.059	-0.055
		-1.00	-0.181	-0.205	-0.239	-0.157	-0.163
LAV	None	None	0.000	-0.020	-0.044	-0.004	-0.001
	25%	-0.25	-0.010	-0.033	-0.065	-0.018	-0.016
		-0.50	-0.020	-0.040	-0.073	-0.027	-0.024
		-1.00	-0.023	-0.052	-0.085	-0.042	-0.033
	50%	-0.25	-0.029	-0.052	-0.076	-0.025	-0.022
		-0.50	-0.094	-0.115	-0.141	-0.092	-0.086
		-1.00	-0.245	-0.248	-0.271	-0.217	-0.220
CC	None	None	0.052	0.040	0.033	0.036	0.049
	25%	-0.25	0.042	0.032	0.026	0.027	0.045
		-0.50	0.024	0.020	0.024	0.019	0.043
		-1.00	-0.017	-0.015	0.003	-0.004	0.031
	50%	-0.25	0.036	0.026	0.026	0.026	0.047
		-0.50	0.011	0.011	0.018	0.010	0.042
		-1.00	-0.046	-0.020	0.006	-0.012	0.026
FPC	None	None	0.024	0.012	0.001	0.005	0.008
	25%	-0.25	0.014	0.003	-0.007	-0.007	0.001
		-0.50	-0.002	-0.008	-0.013	-0.016	-0.003
		-1.00	-0.036	-0.042	-0.038	-0.043	-0.022
	50%	-0.25	0.008	-0.004	-0.010	-0.009	0.002
		-0.50	-0.016	-0.021	-0.024	-0.029	-0.008
		-1.00	-0.068	-0.059	-0.056	-0.064	-0.045

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Linking Constant A - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.028	0.032	0.031	0.031	0.030
	25%	-0.25	0.027	0.029	0.030	0.028	0.031
		-0.50	0.028	0.032	0.030	0.031	0.033
		-1.00	0.026	0.028	0.033	0.031	0.030
	50%	-0.25	0.031	0.028	0.030	0.030	0.033
		-0.50	0.031	0.032	0.028	0.029	0.029
		-1.00	0.028	0.030	0.031	0.027	0.029
HB	None	None	0.027	0.030	0.030	0.030	0.028
	25%	-0.25	0.027	0.028	0.028	0.028	0.032
		-0.50	0.026	0.031	0.028	0.030	0.033
		-1.00	0.024	0.025	0.030	0.029	0.029
	50%	-0.25	0.029	0.026	0.027	0.028	0.032
		-0.50	0.029	0.030	0.027	0.028	0.028
		-1.00	0.026	0.027	0.031	0.025	0.025
LAV	None	None	0.028	0.031	0.032	0.030	0.031
	25%	-0.25	0.034	0.033	0.032	0.030	0.034
		-0.50	0.033	0.037	0.033	0.040	0.040
		-1.00	0.033	0.032	0.039	0.040	0.038
	50%	-0.25	0.036	0.029	0.028	0.036	0.037
		-0.50	0.038	0.036	0.032	0.034	0.033
		-1.00	0.036	0.050	0.039	0.035	0.034
CC	None	None	0.025	0.027	0.026	0.028	0.025
	25%	-0.25	0.026	0.025	0.022	0.025	0.027
		-0.50	0.025	0.028	0.024	0.028	0.028
		-1.00	0.024	0.024	0.024	0.028	0.026
	50%	-0.25	0.027	0.024	0.025	0.027	0.027
		-0.50	0.028	0.026	0.024	0.026	0.023
		-1.00	0.026	0.027	0.026	0.024	0.024
FPC	None	None	0.026	0.029	0.029	0.029	0.027
	25%	-0.25	0.026	0.026	0.025	0.025	0.029
		-0.50	0.026	0.028	0.026	0.029	0.030
		-1.00	0.024	0.025	0.024	0.028	0.028
	50%	-0.25	0.028	0.025	0.029	0.028	0.029
		-0.50	0.028	0.027	0.025	0.026	0.024
		-1.00	0.026	0.028	0.028	0.024	0.025

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Linking Constant A - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.029	0.041	0.067	0.031	0.030
	25%	-0.25	0.029	0.046	0.080	0.030	0.033
		-0.50	0.038	0.059	0.087	0.037	0.039
		-1.00	0.066	0.096	0.132	0.057	0.058
	50%	-0.25	0.035	0.052	0.085	0.032	0.036
		-0.50	0.050	0.075	0.110	0.043	0.042
		-1.00	0.101	0.131	0.172	0.080	0.093
HB	None	None	0.027	0.036	0.058	0.030	0.029
	25%	-0.25	0.031	0.047	0.076	0.031	0.034
		-0.50	0.051	0.072	0.095	0.046	0.047
		-1.00	0.118	0.143	0.170	0.105	0.101
	50%	-0.25	0.038	0.056	0.083	0.035	0.037
		-0.50	0.073	0.097	0.127	0.065	0.061
		-1.00	0.183	0.207	0.241	0.159	0.165
LAV	None	None	0.028	0.037	0.055	0.031	0.031
	25%	-0.25	0.036	0.047	0.073	0.035	0.038
		-0.50	0.039	0.055	0.080	0.048	0.047
		-1.00	0.040	0.061	0.094	0.058	0.050
	50%	-0.25	0.047	0.060	0.081	0.044	0.043
		-0.50	0.102	0.121	0.145	0.098	0.092
		-1.00	0.248	0.253	0.274	0.219	0.223
CC	None	None	0.058	0.049	0.042	0.046	0.055
	25%	-0.25	0.049	0.040	0.034	0.036	0.052
		-0.50	0.035	0.034	0.034	0.034	0.052
		-1.00	0.030	0.029	0.024	0.028	0.041
	50%	-0.25	0.045	0.035	0.036	0.037	0.055
		-0.50	0.030	0.029	0.030	0.028	0.048
		-1.00	0.053	0.033	0.027	0.026	0.035
FPC	None	None	0.036	0.031	0.029	0.029	0.028
	25%	-0.25	0.030	0.027	0.026	0.026	0.029
		-0.50	0.026	0.029	0.029	0.033	0.030
		-1.00	0.043	0.048	0.046	0.052	0.036
	50%	-0.25	0.029	0.025	0.031	0.030	0.029
		-0.50	0.032	0.034	0.035	0.039	0.025
		-1.00	0.073	0.066	0.063	0.069	0.051

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Linking Constant B - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.005	-0.019	-0.059	0.004	-0.012
	25%	-0.25	0.061	0.039	-0.007	0.063	0.039
		-0.50	0.120	0.085	0.041	0.113	0.083
		-1.00	0.223	0.176	0.105	0.207	0.154
	50%	-0.25	0.115	0.085	0.039	0.117	0.091
		-0.50	0.234	0.183	0.126	0.224	0.178
		-1.00	0.434	0.361	0.265	0.401	0.332
HB	None	None	0.003	-0.021	-0.056	0.001	-0.013
	25%	-0.25	0.057	0.033	-0.012	0.055	0.029
		-0.50	0.111	0.068	0.020	0.093	0.055
		-1.00	0.192	0.119	0.031	0.146	0.070
	50%	-0.25	0.111	0.077	0.032	0.106	0.078
		-0.50	0.222	0.160	0.096	0.196	0.140
		-1.00	0.388	0.280	0.153	0.312	0.212
LAV	None	None	0.004	-0.020	-0.054	0.003	-0.015
	25%	-0.25	0.035	0.016	-0.022	0.036	0.012
		-0.50	0.037	0.009	-0.026	0.032	0.005
		-1.00	0.030	0.004	-0.040	0.028	-0.014
	50%	-0.25	0.108	0.067	0.029	0.094	0.068
		-0.50	0.177	0.106	0.042	0.122	0.067
		-1.00	0.309	0.168	0.027	0.144	0.039
CC	None	None	-0.013	-0.001	0.011	0.023	0.042
	25%	-0.25	0.045	0.062	0.072	0.086	0.100
		-0.50	0.108	0.114	0.133	0.143	0.156
		-1.00	0.224	0.233	0.246	0.265	0.269
	50%	-0.25	0.105	0.113	0.127	0.144	0.159
		-0.50	0.234	0.230	0.245	0.266	0.271
		-1.00	0.472	0.476	0.485	0.509	0.517
FPC	None	None	-0.004	-0.005	-0.010	0.010	0.009
	25%	-0.25	0.053	0.055	0.047	0.070	0.062
		-0.50	0.114	0.102	0.096	0.119	0.107
		-1.00	0.216	0.195	0.171	0.211	0.181
	50%	-0.25	0.112	0.105	0.098	0.127	0.118
		-0.50	0.234	0.209	0.197	0.234	0.211
		-1.00	0.437	0.398	0.357	0.413	0.378

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Linking Constant B - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.029	0.033	0.034	0.036	0.042
	25%	-0.25	0.030	0.032	0.037	0.033	0.035
		-0.50	0.032	0.034	0.036	0.035	0.039
		-1.00	0.030	0.030	0.038	0.031	0.041
	50%	-0.25	0.030	0.031	0.035	0.032	0.040
		-0.50	0.030	0.031	0.038	0.031	0.040
		-1.00	0.031	0.032	0.041	0.033	0.043
HB	None	None	0.028	0.032	0.033	0.034	0.040
	25%	-0.25	0.029	0.032	0.036	0.032	0.034
		-0.50	0.030	0.033	0.034	0.033	0.037
		-1.00	0.028	0.029	0.036	0.029	0.039
	50%	-0.25	0.029	0.029	0.033	0.031	0.038
		-0.50	0.029	0.029	0.038	0.029	0.038
		-1.00	0.029	0.030	0.040	0.031	0.040
LAV	None	None	0.031	0.032	0.036	0.036	0.042
	25%	-0.25	0.033	0.037	0.040	0.035	0.041
		-0.50	0.035	0.036	0.038	0.036	0.041
		-1.00	0.030	0.031	0.040	0.031	0.039
	50%	-0.25	0.035	0.032	0.044	0.036	0.046
		-0.50	0.044	0.041	0.048	0.040	0.051
		-1.00	0.050	0.050	0.044	0.041	0.044
CC	None	None	0.029	0.034	0.035	0.034	0.038
	25%	-0.25	0.029	0.032	0.037	0.032	0.034
		-0.50	0.032	0.035	0.035	0.033	0.037
		-1.00	0.031	0.032	0.037	0.031	0.041
	50%	-0.25	0.029	0.030	0.035	0.031	0.037
		-0.50	0.031	0.032	0.038	0.030	0.037
		-1.00	0.033	0.036	0.040	0.035	0.042
FPC	None	None	0.029	0.033	0.036	0.034	0.038
	25%	-0.25	0.029	0.033	0.037	0.032	0.034
		-0.50	0.032	0.034	0.035	0.033	0.037
		-1.00	0.030	0.032	0.035	0.030	0.040
	50%	-0.25	0.029	0.030	0.036	0.031	0.036
		-0.50	0.031	0.031	0.037	0.029	0.037
		-1.00	0.032	0.034	0.040	0.032	0.041

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Linking Constant B - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.030	0.038	0.068	0.036	0.043
	25%	-0.25	0.068	0.051	0.038	0.071	0.052
		-0.50	0.124	0.091	0.054	0.118	0.091
		-1.00	0.225	0.178	0.111	0.210	0.159
	50%	-0.25	0.119	0.091	0.052	0.121	0.100
		-0.50	0.236	0.186	0.132	0.226	0.183
		-1.00	0.435	0.362	0.268	0.403	0.335
HB	None	None	0.028	0.038	0.065	0.034	0.042
	25%	-0.25	0.063	0.046	0.038	0.064	0.045
		-0.50	0.115	0.075	0.039	0.099	0.067
		-1.00	0.194	0.123	0.047	0.149	0.080
	50%	-0.25	0.114	0.082	0.045	0.110	0.087
		-0.50	0.224	0.162	0.103	0.198	0.145
		-1.00	0.389	0.281	0.158	0.314	0.216
LAV	None	None	0.031	0.038	0.065	0.036	0.045
	25%	-0.25	0.048	0.040	0.046	0.050	0.042
		-0.50	0.051	0.037	0.046	0.048	0.041
		-1.00	0.043	0.031	0.057	0.042	0.042
	50%	-0.25	0.113	0.075	0.053	0.100	0.082
		-0.50	0.182	0.114	0.064	0.128	0.084
		-1.00	0.313	0.175	0.052	0.150	0.059
CC	None	None	0.031	0.034	0.036	0.041	0.057
	25%	-0.25	0.054	0.070	0.081	0.092	0.105
		-0.50	0.113	0.120	0.137	0.147	0.161
		-1.00	0.226	0.236	0.249	0.267	0.273
	50%	-0.25	0.109	0.117	0.132	0.148	0.163
		-0.50	0.236	0.232	0.248	0.268	0.274
		-1.00	0.473	0.477	0.487	0.510	0.519
FPC	None	None	0.029	0.034	0.037	0.035	0.039
	25%	-0.25	0.061	0.064	0.060	0.077	0.071
		-0.50	0.118	0.107	0.102	0.124	0.113
		-1.00	0.218	0.197	0.175	0.213	0.185
	50%	-0.25	0.116	0.110	0.105	0.131	0.123
		-0.50	0.236	0.212	0.200	0.236	0.214
		-1.00	0.438	0.399	0.359	0.415	0.380

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

APPENDIX C

BIAS, SE, RMSE VALUES FOR UNIQUE ITEM ESTIMATES

Bias for Discrimination of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.078	0.082	0.097	0.071	0.067
	25%	-0.25	0.082	0.087	0.115	0.072	0.072
		-0.50	0.092	0.097	0.126	0.078	0.081
		-1.00	0.104	0.134	0.170	0.095	0.109
	50%	-0.25	0.084	0.090	0.117	0.075	0.074
		-0.50	0.091	0.114	0.159	0.082	0.096
		-1.00	0.141	0.187	0.251	0.125	0.164
HB	None	None	0.077	0.078	0.089	0.069	0.063
	25%	-0.25	0.083	0.088	0.108	0.072	0.070
		-0.50	0.101	0.107	0.132	0.082	0.088
		-1.00	0.150	0.186	0.221	0.138	0.153
	50%	-0.25	0.087	0.094	0.116	0.077	0.074
		-0.50	0.111	0.136	0.177	0.094	0.113
		-1.00	0.240	0.297	0.360	0.211	0.259
LAV	None	None	0.077	0.077	0.085	0.070	0.062
	25%	-0.25	0.082	0.086	0.104	0.072	0.069
		-0.50	0.094	0.096	0.119	0.078	0.083
		-1.00	0.089	0.111	0.138	0.091	0.103
	50%	-0.25	0.088	0.091	0.108	0.078	0.072
		-0.50	0.118	0.142	0.178	0.100	0.120
		-1.00	0.303	0.355	0.369	0.276	0.314
CC	None	None	0.079	0.070	0.066	0.081	0.074
	25%	-0.25	0.076	0.070	0.064	0.082	0.076
		-0.50	0.076	0.066	0.064	0.080	0.073
		-1.00	0.074	0.066	0.063	0.077	0.074
	50%	-0.25	0.077	0.069	0.067	0.081	0.076
		-0.50	0.072	0.066	0.065	0.083	0.075
		-1.00	0.079	0.064	0.069	0.083	0.078
FPC	None	None	0.073	0.063	0.054	0.070	0.059
	25%	-0.25	0.074	0.063	0.052	0.072	0.060
		-0.50	0.078	0.064	0.055	0.073	0.061
		-1.00	0.084	0.070	0.053	0.078	0.065
	50%	-0.25	0.075	0.064	0.053	0.072	0.059
		-0.50	0.076	0.065	0.055	0.075	0.062
		-1.00	0.093	0.074	0.061	0.087	0.069

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Discrimination of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.136	0.136	0.143	0.129	0.128
	25%	-0.25	0.138	0.137	0.146	0.128	0.135
		-0.50	0.142	0.141	0.148	0.130	0.134
		-1.00	0.146	0.152	0.155	0.132	0.141
	50%	-0.25	0.140	0.140	0.145	0.128	0.135
		-0.50	0.140	0.146	0.157	0.132	0.139
		-1.00	0.150	0.156	0.169	0.140	0.149
HB	None	None	0.135	0.135	0.141	0.128	0.126
	25%	-0.25	0.138	0.136	0.144	0.128	0.133
		-0.50	0.144	0.142	0.149	0.130	0.135
		-1.00	0.155	0.160	0.163	0.138	0.147
	50%	-0.25	0.140	0.141	0.145	0.128	0.135
		-0.50	0.143	0.149	0.159	0.135	0.142
		-1.00	0.165	0.172	0.186	0.152	0.164
LAV	None	None	0.136	0.137	0.141	0.130	0.128
	25%	-0.25	0.140	0.139	0.148	0.129	0.135
		-0.50	0.144	0.145	0.150	0.133	0.137
		-1.00	0.145	0.151	0.158	0.135	0.143
	50%	-0.25	0.142	0.143	0.146	0.129	0.137
		-0.50	0.149	0.155	0.163	0.141	0.147
		-1.00	0.187	0.193	0.199	0.170	0.178
CC	None	None	0.123	0.113	0.112	0.110	0.105
	25%	-0.25	0.124	0.113	0.110	0.109	0.108
		-0.50	0.124	0.114	0.110	0.109	0.107
		-1.00	0.126	0.116	0.109	0.108	0.105
	50%	-0.25	0.123	0.114	0.110	0.108	0.107
		-0.50	0.121	0.114	0.110	0.109	0.107
		-1.00	0.120	0.111	0.108	0.107	0.103
FPC	None	None	0.128	0.120	0.119	0.117	0.112
	25%	-0.25	0.129	0.119	0.117	0.116	0.116
		-0.50	0.130	0.120	0.118	0.116	0.115
		-1.00	0.132	0.123	0.118	0.115	0.115
	50%	-0.25	0.129	0.121	0.117	0.114	0.115
		-0.50	0.127	0.121	0.119	0.116	0.115
		-1.00	0.126	0.119	0.119	0.115	0.114

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Discrimination of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.164	0.167	0.182	0.154	0.149
	25%	-0.25	0.167	0.170	0.194	0.153	0.159
		-0.50	0.177	0.180	0.204	0.158	0.162
		-1.00	0.187	0.212	0.237	0.169	0.185
	50%	-0.25	0.171	0.175	0.196	0.154	0.159
		-0.50	0.174	0.194	0.231	0.162	0.175
		-1.00	0.214	0.252	0.308	0.195	0.228
HB	None	None	0.162	0.163	0.176	0.152	0.145
	25%	-0.25	0.168	0.170	0.189	0.153	0.156
		-0.50	0.183	0.186	0.208	0.161	0.167
		-1.00	0.225	0.254	0.280	0.201	0.218
	50%	-0.25	0.173	0.178	0.195	0.155	0.160
		-0.50	0.190	0.211	0.246	0.172	0.188
		-1.00	0.299	0.348	0.409	0.266	0.311
LAV	None	None	0.163	0.164	0.174	0.155	0.147
	25%	-0.25	0.169	0.171	0.189	0.155	0.158
		-0.50	0.179	0.183	0.200	0.161	0.166
		-1.00	0.176	0.196	0.218	0.170	0.182
	50%	-0.25	0.175	0.178	0.190	0.157	0.161
		-0.50	0.198	0.219	0.249	0.180	0.196
		-1.00	0.364	0.408	0.422	0.330	0.365
CC	None	None	0.156	0.141	0.138	0.150	0.140
	25%	-0.25	0.154	0.141	0.135	0.150	0.143
		-0.50	0.155	0.140	0.136	0.148	0.141
		-1.00	0.154	0.141	0.135	0.146	0.141
	50%	-0.25	0.154	0.142	0.137	0.147	0.143
		-0.50	0.150	0.140	0.137	0.149	0.143
		-1.00	0.153	0.138	0.138	0.149	0.142
FPC	None	None	0.156	0.142	0.137	0.147	0.136
	25%	-0.25	0.156	0.141	0.134	0.148	0.139
		-0.50	0.160	0.143	0.137	0.147	0.139
		-1.00	0.162	0.148	0.136	0.149	0.140
	50%	-0.25	0.157	0.143	0.135	0.145	0.139
		-0.50	0.155	0.144	0.137	0.148	0.140
		-1.00	0.165	0.148	0.140	0.154	0.142

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Difficulty of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.168	0.190	0.210	0.182	0.197
	25%	-0.25	0.203	0.225	0.253	0.210	0.233
		-0.50	0.242	0.268	0.300	0.259	0.278
		-1.00	0.335	0.359	0.387	0.347	0.361
	50%	-0.25	0.239	0.263	0.294	0.250	0.263
		-0.50	0.338	0.365	0.398	0.349	0.377
		-1.00	0.545	0.564	0.591	0.550	0.568
HB	None	None	0.168	0.187	0.207	0.180	0.193
	25%	-0.25	0.200	0.221	0.245	0.206	0.225
		-0.50	0.234	0.257	0.287	0.248	0.264
		-1.00	0.301	0.329	0.355	0.312	0.328
	50%	-0.25	0.234	0.257	0.285	0.245	0.253
		-0.50	0.324	0.349	0.380	0.331	0.357
		-1.00	0.488	0.506	0.537	0.494	0.509
LAV	None	None	0.169	0.186	0.207	0.181	0.193
	25%	-0.25	0.193	0.214	0.240	0.202	0.220
		-0.50	0.200	0.223	0.252	0.216	0.235
		-1.00	0.195	0.219	0.251	0.215	0.235
	50%	-0.25	0.234	0.255	0.281	0.243	0.253
		-0.50	0.301	0.328	0.368	0.311	0.341
		-1.00	0.412	0.436	0.455	0.397	0.432
CC	None	None	0.170	0.177	0.186	0.180	0.184
	25%	-0.25	0.204	0.208	0.212	0.204	0.210
		-0.50	0.239	0.244	0.248	0.243	0.242
		-1.00	0.322	0.327	0.326	0.334	0.323
	50%	-0.25	0.236	0.240	0.242	0.236	0.233
		-0.50	0.325	0.326	0.322	0.325	0.321
		-1.00	0.545	0.552	0.543	0.556	0.536
FPC	None	None	0.169	0.174	0.182	0.173	0.178
	25%	-0.25	0.204	0.206	0.207	0.198	0.203
		-0.50	0.238	0.237	0.238	0.233	0.231
		-1.00	0.316	0.303	0.293	0.302	0.287
	50%	-0.25	0.235	0.236	0.238	0.230	0.224
		-0.50	0.325	0.318	0.306	0.309	0.303
		-1.00	0.515	0.493	0.467	0.483	0.456

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Difficulty of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.186	0.183	0.192	0.183	0.189
	25%	-0.25	0.186	0.178	0.193	0.180	0.193
		-0.50	0.181	0.177	0.192	0.177	0.192
		-1.00	0.175	0.175	0.184	0.169	0.185
	50%	-0.25	0.182	0.183	0.190	0.179	0.191
		-0.50	0.177	0.176	0.189	0.176	0.188
		-1.00	0.168	0.166	0.174	0.165	0.172
HB	None	None	0.187	0.184	0.194	0.185	0.191
	25%	-0.25	0.185	0.177	0.194	0.180	0.194
		-0.50	0.178	0.175	0.191	0.175	0.191
		-1.00	0.166	0.167	0.177	0.161	0.177
	50%	-0.25	0.180	0.182	0.190	0.177	0.191
		-0.50	0.172	0.172	0.187	0.173	0.184
		-1.00	0.154	0.153	0.164	0.151	0.162
LAV	None	None	0.189	0.187	0.196	0.186	0.195
	25%	-0.25	0.188	0.182	0.198	0.184	0.198
		-0.50	0.183	0.183	0.197	0.181	0.194
		-1.00	0.184	0.183	0.193	0.175	0.189
	50%	-0.25	0.181	0.186	0.195	0.179	0.195
		-0.50	0.177	0.178	0.191	0.179	0.190
		-1.00	0.163	0.162	0.175	0.158	0.161
CC	None	None	0.198	0.190	0.196	0.186	0.189
	25%	-0.25	0.197	0.185	0.197	0.183	0.194
		-0.50	0.191	0.185	0.196	0.180	0.194
		-1.00	0.185	0.181	0.192	0.174	0.189
	50%	-0.25	0.192	0.189	0.194	0.181	0.190
		-0.50	0.187	0.183	0.195	0.179	0.190
		-1.00	0.178	0.177	0.191	0.172	0.182
FPC	None	None	0.190	0.182	0.189	0.180	0.182
	25%	-0.25	0.190	0.177	0.189	0.177	0.187
		-0.50	0.183	0.177	0.189	0.173	0.186
		-1.00	0.178	0.174	0.184	0.166	0.180
	50%	-0.25	0.184	0.181	0.187	0.174	0.183
		-0.50	0.180	0.175	0.186	0.172	0.182
		-1.00	0.170	0.168	0.181	0.163	0.172

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Difficulty of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.264	0.279	0.302	0.270	0.288
	25%	-0.25	0.291	0.304	0.339	0.293	0.322
		-0.50	0.323	0.341	0.375	0.331	0.357
		-1.00	0.396	0.415	0.446	0.401	0.422
	50%	-0.25	0.320	0.340	0.372	0.325	0.345
		-0.50	0.397	0.420	0.457	0.407	0.438
		-1.00	0.578	0.595	0.624	0.580	0.599
HB	None	None	0.264	0.278	0.301	0.269	0.285
	25%	-0.25	0.289	0.301	0.333	0.289	0.316
		-0.50	0.315	0.331	0.364	0.321	0.344
		-1.00	0.364	0.384	0.411	0.368	0.387
	50%	-0.25	0.314	0.335	0.365	0.319	0.337
		-0.50	0.383	0.405	0.440	0.391	0.419
		-1.00	0.524	0.542	0.572	0.527	0.544
LAV	None	None	0.267	0.279	0.303	0.271	0.288
	25%	-0.25	0.284	0.298	0.331	0.288	0.314
		-0.50	0.289	0.308	0.340	0.298	0.324
		-1.00	0.287	0.305	0.337	0.293	0.318
	50%	-0.25	0.315	0.335	0.366	0.318	0.340
		-0.50	0.367	0.391	0.432	0.378	0.408
		-1.00	0.461	0.479	0.499	0.443	0.473
CC	None	None	0.275	0.275	0.287	0.270	0.277
	25%	-0.25	0.298	0.293	0.309	0.289	0.302
		-0.50	0.323	0.324	0.337	0.321	0.330
		-1.00	0.388	0.394	0.404	0.394	0.395
	50%	-0.25	0.321	0.324	0.333	0.315	0.319
		-0.50	0.392	0.394	0.403	0.391	0.395
		-1.00	0.580	0.589	0.587	0.589	0.575
FPC	None	None	0.267	0.265	0.276	0.260	0.267
	25%	-0.25	0.292	0.285	0.296	0.278	0.289
		-0.50	0.318	0.313	0.322	0.306	0.313
		-1.00	0.380	0.367	0.366	0.360	0.356
	50%	-0.25	0.316	0.315	0.321	0.304	0.305
		-0.50	0.388	0.380	0.381	0.372	0.372
		-1.00	0.550	0.530	0.512	0.518	0.497

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Pseudo-Guessing of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.057	0.067	0.078	0.059	0.068
	25%	-0.25	0.057	0.067	0.077	0.059	0.067
		-0.50	0.058	0.068	0.078	0.059	0.068
		-1.00	0.058	0.068	0.079	0.059	0.068
	50%	-0.25	0.057	0.067	0.078	0.059	0.068
		-0.50	0.057	0.067	0.079	0.059	0.068
		-1.00	0.056	0.067	0.078	0.059	0.068
HB	None	None	0.057	0.067	0.078	0.059	0.068
	25%	-0.25	0.057	0.067	0.077	0.059	0.067
		-0.50	0.058	0.068	0.078	0.059	0.068
		-1.00	0.058	0.068	0.079	0.059	0.068
	50%	-0.25	0.057	0.067	0.078	0.059	0.068
		-0.50	0.057	0.067	0.079	0.059	0.068
		-1.00	0.056	0.067	0.078	0.059	0.068
LAV	None	None	0.057	0.067	0.078	0.059	0.068
	25%	-0.25	0.057	0.067	0.077	0.059	0.067
		-0.50	0.058	0.068	0.078	0.059	0.068
		-1.00	0.058	0.068	0.079	0.059	0.068
	50%	-0.25	0.057	0.067	0.078	0.059	0.068
		-0.50	0.057	0.067	0.079	0.059	0.068
		-1.00	0.056	0.067	0.078	0.059	0.068
CC	None	None	0.058	0.065	0.071	0.058	0.064
	25%	-0.25	0.058	0.064	0.070	0.058	0.063
		-0.50	0.058	0.064	0.070	0.057	0.063
		-1.00	0.057	0.064	0.070	0.057	0.062
	50%	-0.25	0.057	0.064	0.070	0.057	0.063
		-0.50	0.056	0.063	0.070	0.057	0.063
		-1.00	0.054	0.061	0.068	0.055	0.062
FPC	None	None	0.057	0.065	0.071	0.058	0.064
	25%	-0.25	0.057	0.064	0.070	0.057	0.063
		-0.50	0.057	0.064	0.070	0.057	0.063
		-1.00	0.056	0.063	0.069	0.057	0.062
	50%	-0.25	0.057	0.063	0.070	0.057	0.063
		-0.50	0.056	0.063	0.069	0.057	0.063
		-1.00	0.054	0.061	0.067	0.055	0.061

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Pseudo-Guessing of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.031	0.031	0.029	0.029	0.029
	25%	-0.25	0.030	0.030	0.029	0.029	0.029
		-0.50	0.030	0.031	0.029	0.029	0.030
		-1.00	0.030	0.031	0.029	0.029	0.029
	50%	-0.25	0.030	0.031	0.030	0.029	0.029
		-0.50	0.031	0.031	0.030	0.029	0.029
		-1.00	0.030	0.031	0.030	0.029	0.029
HB	None	None	0.031	0.031	0.029	0.029	0.029
	25%	-0.25	0.030	0.030	0.029	0.029	0.029
		-0.50	0.030	0.031	0.029	0.029	0.030
		-1.00	0.030	0.031	0.029	0.029	0.029
	50%	-0.25	0.030	0.031	0.030	0.029	0.029
		-0.50	0.031	0.031	0.030	0.029	0.029
		-1.00	0.030	0.031	0.030	0.029	0.029
LAV	None	None	0.031	0.031	0.029	0.029	0.029
	25%	-0.25	0.030	0.030	0.029	0.029	0.029
		-0.50	0.030	0.031	0.029	0.029	0.030
		-1.00	0.030	0.031	0.029	0.029	0.029
	50%	-0.25	0.030	0.031	0.030	0.029	0.029
		-0.50	0.031	0.031	0.030	0.029	0.029
		-1.00	0.030	0.031	0.030	0.029	0.029
CC	None	None	0.032	0.029	0.025	0.028	0.026
	25%	-0.25	0.031	0.028	0.025	0.028	0.026
		-0.50	0.030	0.029	0.025	0.027	0.026
		-1.00	0.030	0.028	0.024	0.027	0.025
	50%	-0.25	0.031	0.029	0.025	0.027	0.025
		-0.50	0.030	0.028	0.024	0.027	0.025
		-1.00	0.027	0.026	0.023	0.025	0.024
FPC	None	None	0.031	0.029	0.025	0.028	0.026
	25%	-0.25	0.030	0.028	0.025	0.028	0.026
		-0.50	0.030	0.028	0.025	0.027	0.026
		-1.00	0.029	0.028	0.024	0.027	0.025
	50%	-0.25	0.030	0.028	0.025	0.027	0.025
		-0.50	0.030	0.027	0.024	0.027	0.025
		-1.00	0.027	0.026	0.023	0.025	0.024

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Pseudo-Guessing of 80 Unique Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.069	0.078	0.087	0.069	0.077
	25%	-0.25	0.068	0.078	0.087	0.069	0.078
		-0.50	0.069	0.079	0.088	0.069	0.078
		-1.00	0.069	0.079	0.088	0.069	0.078
	50%	-0.25	0.068	0.078	0.088	0.069	0.078
		-0.50	0.068	0.078	0.088	0.069	0.078
		-1.00	0.068	0.078	0.088	0.069	0.078
HB	None	None	0.069	0.078	0.087	0.069	0.077
	25%	-0.25	0.068	0.078	0.087	0.069	0.078
		-0.50	0.069	0.079	0.088	0.069	0.078
		-1.00	0.069	0.079	0.088	0.069	0.078
	50%	-0.25	0.068	0.078	0.088	0.069	0.078
		-0.50	0.068	0.078	0.088	0.069	0.078
		-1.00	0.068	0.078	0.088	0.069	0.078
LAV	None	None	0.069	0.078	0.087	0.069	0.077
	25%	-0.25	0.068	0.078	0.087	0.069	0.078
		-0.50	0.069	0.079	0.088	0.069	0.078
		-1.00	0.069	0.079	0.088	0.069	0.078
	50%	-0.25	0.068	0.078	0.088	0.069	0.078
		-0.50	0.068	0.078	0.088	0.069	0.078
		-1.00	0.068	0.078	0.088	0.069	0.078
CC	None	None	0.070	0.075	0.079	0.067	0.072
	25%	-0.25	0.069	0.074	0.078	0.067	0.071
		-0.50	0.069	0.074	0.078	0.066	0.071
		-1.00	0.068	0.072	0.076	0.065	0.070
	50%	-0.25	0.069	0.073	0.078	0.066	0.071
		-0.50	0.067	0.072	0.077	0.066	0.071
		-1.00	0.064	0.069	0.074	0.064	0.069
FPC	None	None	0.069	0.074	0.078	0.067	0.072
	25%	-0.25	0.068	0.073	0.077	0.067	0.071
		-0.50	0.068	0.073	0.077	0.066	0.071
		-1.00	0.067	0.072	0.076	0.065	0.070
	50%	-0.25	0.068	0.072	0.077	0.066	0.071
		-0.50	0.066	0.071	0.076	0.065	0.070
		-1.00	0.063	0.069	0.074	0.063	0.068

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Discrimination of 80 Unique Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.058	0.068	0.096	0.056	0.049
	25%	-0.25	0.063	0.077	0.115	0.057	0.051
		-0.50	0.072	0.093	0.126	0.060	0.058
		-1.00	0.102	0.137	0.179	0.071	0.075
	50%	-0.25	0.065	0.085	0.119	0.059	0.056
		-0.50	0.085	0.113	0.153	0.064	0.063
		-1.00	0.139	0.182	0.239	0.087	0.109
HB	None	None	0.057	0.063	0.085	0.056	0.048
	25%	-0.25	0.065	0.078	0.110	0.058	0.051
		-0.50	0.087	0.109	0.136	0.065	0.065
		-1.00	0.162	0.200	0.233	0.109	0.119
	50%	-0.25	0.071	0.091	0.117	0.060	0.057
		-0.50	0.111	0.141	0.174	0.075	0.080
		-1.00	0.247	0.294	0.350	0.167	0.198
LAV	None	None	0.057	0.064	0.080	0.056	0.049
	25%	-0.25	0.063	0.075	0.104	0.058	0.053
		-0.50	0.068	0.084	0.116	0.063	0.060
		-1.00	0.070	0.093	0.127	0.068	0.063
	50%	-0.25	0.074	0.094	0.115	0.062	0.059
		-0.50	0.140	0.170	0.198	0.098	0.109
		-1.00	0.353	0.370	0.409	0.249	0.282
CC	None	None	0.048	0.043	0.039	0.064	0.054
	25%	-0.25	0.048	0.043	0.040	0.064	0.053
		-0.50	0.053	0.048	0.041	0.065	0.056
		-1.00	0.071	0.061	0.047	0.065	0.055
	50%	-0.25	0.050	0.047	0.039	0.066	0.055
		-0.50	0.059	0.051	0.041	0.065	0.056
		-1.00	0.085	0.061	0.045	0.070	0.058
FPC	None	None	0.054	0.050	0.048	0.060	0.054
	25%	-0.25	0.057	0.052	0.051	0.062	0.055
		-0.50	0.064	0.060	0.056	0.066	0.061
		-1.00	0.085	0.080	0.072	0.076	0.067
	50%	-0.25	0.060	0.057	0.051	0.064	0.059
		-0.50	0.072	0.066	0.060	0.070	0.061
		-1.00	0.104	0.091	0.084	0.087	0.078

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Discrimination of 80 Unique Items- 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.098	0.098	0.098	0.083	0.082
	25%	-0.25	0.100	0.099	0.101	0.082	0.084
		-0.50	0.102	0.099	0.100	0.085	0.085
		-1.00	0.105	0.104	0.109	0.087	0.087
	50%	-0.25	0.101	0.100	0.100	0.082	0.085
		-0.50	0.104	0.105	0.102	0.085	0.084
		-1.00	0.109	0.109	0.114	0.089	0.090
HB	None	None	0.097	0.096	0.097	0.082	0.080
	25%	-0.25	0.100	0.099	0.099	0.082	0.084
		-0.50	0.103	0.100	0.100	0.087	0.086
		-1.00	0.111	0.110	0.114	0.092	0.092
	50%	-0.25	0.101	0.099	0.099	0.082	0.085
		-0.50	0.107	0.107	0.104	0.087	0.086
		-1.00	0.121	0.119	0.126	0.097	0.097
LAV	None	None	0.098	0.097	0.097	0.083	0.082
	25%	-0.25	0.102	0.100	0.100	0.084	0.086
		-0.50	0.103	0.100	0.101	0.090	0.088
		-1.00	0.102	0.101	0.107	0.090	0.088
	50%	-0.25	0.104	0.102	0.099	0.086	0.087
		-0.50	0.114	0.114	0.108	0.093	0.092
		-1.00	0.139	0.145	0.141	0.110	0.112
CC	None	None	0.093	0.086	0.081	0.077	0.073
	25%	-0.25	0.095	0.087	0.081	0.076	0.074
		-0.50	0.095	0.085	0.080	0.078	0.074
		-1.00	0.096	0.088	0.082	0.078	0.074
	50%	-0.25	0.094	0.086	0.080	0.075	0.074
		-0.50	0.095	0.087	0.079	0.076	0.072
		-1.00	0.095	0.084	0.079	0.076	0.072
FPC	None	None	0.096	0.089	0.085	0.080	0.078
	25%	-0.25	0.098	0.090	0.085	0.079	0.079
		-0.50	0.098	0.088	0.084	0.082	0.079
		-1.00	0.098	0.091	0.087	0.082	0.079
	50%	-0.25	0.097	0.089	0.085	0.079	0.078
		-0.50	0.097	0.090	0.083	0.080	0.077
		-1.00	0.098	0.088	0.086	0.081	0.078

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Discrimination of 80 Unique Items- 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.118	0.125	0.143	0.105	0.099
	25%	-0.25	0.123	0.132	0.157	0.105	0.102
		-0.50	0.131	0.142	0.166	0.110	0.107
		-1.00	0.153	0.177	0.213	0.117	0.120
	50%	-0.25	0.126	0.138	0.160	0.106	0.105
		-0.50	0.141	0.159	0.187	0.110	0.109
		-1.00	0.183	0.216	0.267	0.130	0.146
HB	None	None	0.117	0.121	0.135	0.105	0.097
	25%	-0.25	0.125	0.133	0.153	0.106	0.102
		-0.50	0.142	0.154	0.173	0.114	0.112
		-1.00	0.202	0.231	0.261	0.148	0.155
	50%	-0.25	0.130	0.141	0.158	0.107	0.106
		-0.50	0.161	0.182	0.206	0.120	0.122
		-1.00	0.280	0.319	0.373	0.199	0.224
LAV	None	None	0.117	0.122	0.133	0.106	0.099
	25%	-0.25	0.125	0.132	0.150	0.107	0.105
		-0.50	0.130	0.137	0.159	0.115	0.111
		-1.00	0.130	0.144	0.171	0.118	0.113
	50%	-0.25	0.135	0.145	0.157	0.111	0.109
		-0.50	0.187	0.209	0.229	0.142	0.148
		-1.00	0.382	0.399	0.434	0.277	0.306
CC	None	None	0.109	0.100	0.094	0.107	0.098
	25%	-0.25	0.111	0.100	0.093	0.107	0.099
		-0.50	0.114	0.101	0.094	0.109	0.100
		-1.00	0.125	0.112	0.098	0.109	0.099
	50%	-0.25	0.110	0.101	0.093	0.108	0.100
		-0.50	0.117	0.104	0.093	0.108	0.099
		-1.00	0.134	0.109	0.095	0.111	0.100
FPC	None	None	0.114	0.106	0.101	0.106	0.100
	25%	-0.25	0.118	0.109	0.104	0.107	0.101
		-0.50	0.123	0.112	0.106	0.112	0.104
		-1.00	0.136	0.127	0.117	0.117	0.109
	50%	-0.25	0.118	0.111	0.103	0.108	0.103
		-0.50	0.127	0.118	0.107	0.112	0.104
		-1.00	0.149	0.132	0.125	0.124	0.115

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Difficulty of 80 Unique Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.134	0.159	0.195	0.139	0.158
	25%	-0.25	0.173	0.202	0.238	0.168	0.185
		-0.50	0.222	0.254	0.285	0.203	0.228
		-1.00	0.322	0.353	0.386	0.295	0.313
	50%	-0.25	0.216	0.250	0.285	0.203	0.228
		-0.50	0.327	0.359	0.391	0.301	0.319
		-1.00	0.525	0.552	0.581	0.503	0.518
HB	None	None	0.133	0.156	0.190	0.137	0.154
	25%	-0.25	0.169	0.197	0.231	0.164	0.179
		-0.50	0.212	0.244	0.274	0.194	0.218
		-1.00	0.290	0.320	0.354	0.262	0.281
	50%	-0.25	0.211	0.244	0.277	0.197	0.220
		-0.50	0.312	0.343	0.376	0.285	0.304
		-1.00	0.470	0.498	0.529	0.447	0.464
LAV	None	None	0.134	0.156	0.188	0.139	0.156
	25%	-0.25	0.153	0.183	0.219	0.155	0.171
		-0.50	0.153	0.187	0.222	0.155	0.180
		-1.00	0.153	0.185	0.226	0.157	0.174
	50%	-0.25	0.209	0.236	0.273	0.190	0.216
		-0.50	0.268	0.301	0.341	0.235	0.269
		-1.00	0.405	0.422	0.447	0.327	0.369
CC	None	None	0.132	0.143	0.162	0.135	0.147
	25%	-0.25	0.169	0.180	0.191	0.163	0.168
		-0.50	0.214	0.228	0.233	0.196	0.208
		-1.00	0.316	0.338	0.345	0.297	0.300
	50%	-0.25	0.211	0.225	0.231	0.196	0.205
		-0.50	0.323	0.336	0.336	0.291	0.292
		-1.00	0.541	0.566	0.569	0.531	0.525
FPC	None	None	0.134	0.145	0.164	0.136	0.153
	25%	-0.25	0.169	0.181	0.194	0.161	0.171
		-0.50	0.215	0.226	0.230	0.190	0.206
		-1.00	0.306	0.309	0.310	0.267	0.268
	50%	-0.25	0.213	0.227	0.232	0.193	0.207
		-0.50	0.319	0.325	0.323	0.279	0.281
		-1.00	0.503	0.504	0.496	0.460	0.451

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Difficulty of 80 Unique Items- 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.146	0.145	0.153	0.139	0.145
	25%	-0.25	0.142	0.148	0.153	0.135	0.144
		-0.50	0.141	0.143	0.149	0.135	0.141
		-1.00	0.136	0.139	0.145	0.130	0.137
	50%	-0.25	0.143	0.144	0.149	0.134	0.143
		-0.50	0.140	0.143	0.146	0.131	0.140
		-1.00	0.132	0.133	0.140	0.127	0.134
HB	None	None	0.146	0.146	0.154	0.138	0.145
	25%	-0.25	0.141	0.148	0.153	0.134	0.143
		-0.50	0.138	0.141	0.146	0.133	0.139
		-1.00	0.128	0.131	0.139	0.124	0.130
	50%	-0.25	0.141	0.143	0.148	0.133	0.142
		-0.50	0.136	0.139	0.144	0.128	0.136
		-1.00	0.121	0.122	0.129	0.116	0.123
LAV	None	None	0.147	0.146	0.156	0.138	0.146
	25%	-0.25	0.143	0.151	0.156	0.136	0.145
		-0.50	0.144	0.147	0.152	0.136	0.143
		-1.00	0.143	0.145	0.155	0.134	0.141
	50%	-0.25	0.143	0.144	0.151	0.135	0.145
		-0.50	0.138	0.141	0.145	0.128	0.136
		-1.00	0.121	0.133	0.134	0.116	0.123
CC	None	None	0.151	0.146	0.150	0.137	0.140
	25%	-0.25	0.146	0.148	0.150	0.133	0.139
		-0.50	0.144	0.143	0.146	0.133	0.137
		-1.00	0.137	0.137	0.143	0.128	0.134
	50%	-0.25	0.146	0.143	0.146	0.132	0.139
		-0.50	0.142	0.142	0.144	0.129	0.135
		-1.00	0.132	0.133	0.140	0.124	0.131
FPC	None	None	0.147	0.142	0.146	0.134	0.136
	25%	-0.25	0.142	0.145	0.146	0.130	0.135
		-0.50	0.141	0.139	0.141	0.129	0.132
		-1.00	0.135	0.134	0.138	0.124	0.128
	50%	-0.25	0.142	0.139	0.142	0.129	0.134
		-0.50	0.138	0.138	0.139	0.124	0.130
		-1.00	0.129	0.129	0.133	0.118	0.124

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Difficulty of 80 Unique Items- 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.213	0.230	0.263	0.206	0.226
	25%	-0.25	0.241	0.269	0.300	0.229	0.248
		-0.50	0.278	0.307	0.335	0.259	0.283
		-1.00	0.361	0.389	0.423	0.336	0.355
	50%	-0.25	0.274	0.305	0.336	0.259	0.284
		-0.50	0.366	0.395	0.427	0.342	0.361
		-1.00	0.546	0.572	0.601	0.523	0.540
HB	None	None	0.212	0.227	0.260	0.205	0.223
	25%	-0.25	0.238	0.265	0.294	0.225	0.243
		-0.50	0.270	0.298	0.324	0.250	0.273
		-1.00	0.332	0.359	0.392	0.305	0.323
	50%	-0.25	0.270	0.299	0.330	0.253	0.277
		-0.50	0.352	0.381	0.412	0.327	0.346
		-1.00	0.494	0.520	0.551	0.470	0.489
LAV	None	None	0.213	0.228	0.260	0.206	0.224
	25%	-0.25	0.227	0.255	0.286	0.218	0.238
		-0.50	0.227	0.255	0.284	0.219	0.243
		-1.00	0.227	0.252	0.290	0.220	0.238
	50%	-0.25	0.270	0.294	0.328	0.248	0.275
		-0.50	0.317	0.347	0.383	0.284	0.315
		-1.00	0.434	0.451	0.475	0.361	0.400
CC	None	None	0.212	0.215	0.231	0.202	0.213
	25%	-0.25	0.239	0.250	0.259	0.221	0.230
		-0.50	0.273	0.285	0.290	0.251	0.262
		-1.00	0.355	0.375	0.386	0.338	0.342
	50%	-0.25	0.272	0.284	0.290	0.250	0.261
		-0.50	0.363	0.375	0.378	0.332	0.336
		-1.00	0.561	0.585	0.590	0.550	0.546
FPC	None	None	0.212	0.216	0.231	0.201	0.215
	25%	-0.25	0.238	0.249	0.257	0.219	0.230
		-0.50	0.272	0.281	0.284	0.244	0.257
		-1.00	0.346	0.349	0.352	0.308	0.310
	50%	-0.25	0.271	0.282	0.287	0.246	0.259
		-0.50	0.358	0.364	0.364	0.318	0.322
		-1.00	0.525	0.525	0.518	0.482	0.476

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Pseudo-Guessing of 80 Unique Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.046	0.056	0.069	0.042	0.051
	25%	-0.25	0.046	0.056	0.068	0.043	0.052
		-0.50	0.045	0.055	0.068	0.042	0.051
		-1.00	0.045	0.056	0.069	0.043	0.050
	50%	-0.25	0.045	0.056	0.068	0.043	0.051
		-0.50	0.045	0.056	0.068	0.043	0.051
		-1.00	0.045	0.056	0.068	0.043	0.052
HB	None	None	0.046	0.056	0.069	0.042	0.051
	25%	-0.25	0.046	0.056	0.068	0.043	0.052
		-0.50	0.045	0.055	0.068	0.042	0.051
		-1.00	0.045	0.056	0.069	0.043	0.050
	50%	-0.25	0.045	0.056	0.068	0.043	0.051
		-0.50	0.045	0.056	0.068	0.043	0.051
		-1.00	0.045	0.056	0.068	0.043	0.052
LAV	None	None	0.046	0.056	0.069	0.042	0.051
	25%	-0.25	0.046	0.056	0.068	0.043	0.052
		-0.50	0.045	0.055	0.068	0.042	0.051
		-1.00	0.045	0.056	0.069	0.043	0.050
	50%	-0.25	0.045	0.056	0.068	0.043	0.051
		-0.50	0.045	0.056	0.068	0.043	0.051
		-1.00	0.045	0.056	0.068	0.043	0.052
CC	None	None	0.046	0.053	0.062	0.042	0.050
	25%	-0.25	0.046	0.053	0.060	0.042	0.050
		-0.50	0.045	0.052	0.060	0.042	0.049
		-1.00	0.044	0.051	0.060	0.042	0.048
	50%	-0.25	0.045	0.053	0.060	0.043	0.049
		-0.50	0.044	0.052	0.059	0.042	0.049
		-1.00	0.042	0.050	0.057	0.042	0.049
FPC	None	None	0.046	0.053	0.061	0.042	0.050
	25%	-0.25	0.045	0.053	0.060	0.042	0.050
		-0.50	0.045	0.052	0.060	0.042	0.049
		-1.00	0.044	0.051	0.060	0.042	0.048
	50%	-0.25	0.044	0.052	0.060	0.043	0.049
		-0.50	0.044	0.052	0.059	0.042	0.049
		-1.00	0.042	0.050	0.057	0.042	0.049

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Pseudo-Guessing of 80 Unique Items- 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.033	0.034	0.035	0.031	0.031
	25%	-0.25	0.033	0.035	0.035	0.030	0.032
		-0.50	0.033	0.034	0.035	0.031	0.031
		-1.00	0.033	0.034	0.035	0.031	0.031
	50%	-0.25	0.033	0.034	0.035	0.030	0.031
		-0.50	0.033	0.035	0.035	0.030	0.032
		-1.00	0.033	0.034	0.036	0.030	0.032
HB	None	None	0.033	0.034	0.035	0.031	0.031
	25%	-0.25	0.033	0.035	0.035	0.030	0.032
		-0.50	0.033	0.034	0.035	0.031	0.031
		-1.00	0.033	0.034	0.035	0.031	0.031
	50%	-0.25	0.033	0.034	0.035	0.030	0.031
		-0.50	0.033	0.035	0.035	0.030	0.032
		-1.00	0.033	0.034	0.036	0.030	0.032
LAV	None	None	0.033	0.034	0.035	0.031	0.031
	25%	-0.25	0.033	0.035	0.035	0.030	0.032
		-0.50	0.033	0.034	0.035	0.031	0.031
		-1.00	0.033	0.034	0.035	0.031	0.031
	50%	-0.25	0.033	0.034	0.035	0.030	0.031
		-0.50	0.033	0.035	0.035	0.030	0.032
		-1.00	0.033	0.034	0.036	0.030	0.032
CC	None	None	0.033	0.033	0.031	0.030	0.029
	25%	-0.25	0.033	0.033	0.031	0.029	0.029
		-0.50	0.033	0.032	0.031	0.029	0.029
		-1.00	0.032	0.032	0.030	0.029	0.028
	50%	-0.25	0.033	0.032	0.031	0.029	0.029
		-0.50	0.032	0.033	0.030	0.029	0.028
		-1.00	0.031	0.030	0.029	0.027	0.027
FPC	None	None	0.033	0.033	0.031	0.030	0.029
	25%	-0.25	0.033	0.033	0.031	0.029	0.029
		-0.50	0.033	0.032	0.031	0.029	0.029
		-1.00	0.032	0.032	0.030	0.029	0.028
	50%	-0.25	0.033	0.032	0.031	0.029	0.029
		-0.50	0.032	0.032	0.030	0.029	0.029
		-1.00	0.031	0.030	0.029	0.027	0.027

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Pseudo-Guessing of 80 Unique Items- 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.061	0.071	0.083	0.056	0.064
	25%	-0.25	0.061	0.072	0.082	0.056	0.065
		-0.50	0.061	0.071	0.082	0.056	0.064
		-1.00	0.061	0.072	0.083	0.057	0.063
	50%	-0.25	0.060	0.071	0.082	0.056	0.064
		-0.50	0.061	0.072	0.082	0.056	0.064
		-1.00	0.060	0.071	0.082	0.056	0.065
HB	None	None	0.061	0.071	0.083	0.056	0.064
	25%	-0.25	0.061	0.072	0.082	0.056	0.065
		-0.50	0.061	0.071	0.082	0.056	0.064
		-1.00	0.061	0.072	0.083	0.057	0.063
	50%	-0.25	0.060	0.071	0.082	0.056	0.064
		-0.50	0.061	0.072	0.082	0.056	0.064
		-1.00	0.060	0.071	0.082	0.056	0.065
LAV	None	None	0.061	0.071	0.083	0.056	0.064
	25%	-0.25	0.061	0.072	0.082	0.056	0.065
		-0.50	0.061	0.071	0.082	0.056	0.064
		-1.00	0.061	0.072	0.083	0.057	0.063
	50%	-0.25	0.060	0.071	0.082	0.056	0.064
		-0.50	0.061	0.072	0.082	0.056	0.064
		-1.00	0.060	0.071	0.082	0.056	0.065
CC	None	None	0.062	0.068	0.074	0.055	0.061
	25%	-0.25	0.061	0.067	0.073	0.055	0.061
		-0.50	0.060	0.066	0.072	0.055	0.060
		-1.00	0.059	0.065	0.071	0.055	0.059
	50%	-0.25	0.060	0.066	0.072	0.055	0.061
		-0.50	0.059	0.066	0.070	0.055	0.060
		-1.00	0.056	0.062	0.068	0.054	0.059
FPC	None	None	0.061	0.067	0.074	0.055	0.061
	25%	-0.25	0.061	0.067	0.072	0.055	0.061
		-0.50	0.060	0.066	0.071	0.055	0.060
		-1.00	0.059	0.065	0.071	0.055	0.059
	50%	-0.25	0.060	0.066	0.072	0.055	0.061
		-0.50	0.059	0.065	0.070	0.055	0.060
		-1.00	0.056	0.062	0.068	0.054	0.059

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

APPENDIX D

BIAS, SE, RMSE VALUES FOR ALL ITEM ESTIMATES

Bias for Discrimination of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.078	0.081	0.097	0.069	0.065
	25%	-0.25	0.081	0.086	0.113	0.070	0.069
		-0.50	0.088	0.095	0.123	0.074	0.077
		-1.00	0.100	0.130	0.167	0.089	0.107
	50%	-0.25	0.081	0.089	0.117	0.072	0.071
		-0.50	0.088	0.110	0.154	0.078	0.092
		-1.00	0.138	0.182	0.242	0.122	0.161
HB	None	None	0.077	0.077	0.089	0.067	0.062
	25%	-0.25	0.082	0.086	0.106	0.070	0.067
		-0.50	0.096	0.104	0.128	0.078	0.084
		-1.00	0.148	0.182	0.218	0.133	0.152
	50%	-0.25	0.084	0.093	0.116	0.073	0.072
		-0.50	0.108	0.131	0.172	0.091	0.110
		-1.00	0.238	0.290	0.347	0.210	0.254
LAV	None	None	0.077	0.076	0.086	0.068	0.061
	25%	-0.25	0.081	0.085	0.102	0.070	0.066
		-0.50	0.090	0.095	0.117	0.074	0.079
		-1.00	0.085	0.107	0.137	0.085	0.101
	50%	-0.25	0.084	0.090	0.107	0.074	0.070
		-0.50	0.115	0.137	0.173	0.097	0.116
		-1.00	0.303	0.348	0.356	0.277	0.310
CC	None	None	0.077	0.069	0.065	0.078	0.072
	25%	-0.25	0.075	0.069	0.063	0.078	0.072
		-0.50	0.074	0.064	0.064	0.076	0.072
		-1.00	0.070	0.068	0.070	0.079	0.080
	50%	-0.25	0.075	0.066	0.064	0.075	0.071
		-0.50	0.070	0.062	0.063	0.077	0.071
		-1.00	0.073	0.067	0.077	0.086	0.086
FPC	None	None	0.075	0.067	0.060	0.072	0.064
	25%	-0.25	0.075	0.067	0.058	0.074	0.064
		-0.50	0.078	0.068	0.060	0.075	0.065
		-1.00	0.083	0.072	0.059	0.079	0.068
	50%	-0.25	0.076	0.067	0.058	0.074	0.063
		-0.50	0.077	0.068	0.060	0.076	0.066
		-1.00	0.091	0.076	0.065	0.086	0.072

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Discrimination of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.136	0.137	0.142	0.129	0.128
	25%	-0.25	0.138	0.137	0.145	0.128	0.134
		-0.50	0.141	0.141	0.148	0.130	0.133
		-1.00	0.145	0.151	0.155	0.132	0.141
	50%	-0.25	0.139	0.140	0.145	0.128	0.134
		-0.50	0.140	0.145	0.155	0.131	0.139
		-1.00	0.150	0.156	0.168	0.139	0.149
HB	None	None	0.135	0.135	0.140	0.127	0.126
	25%	-0.25	0.138	0.136	0.144	0.127	0.132
		-0.50	0.142	0.142	0.149	0.130	0.135
		-1.00	0.154	0.159	0.163	0.139	0.147
	50%	-0.25	0.139	0.141	0.144	0.128	0.135
		-0.50	0.144	0.148	0.158	0.134	0.141
		-1.00	0.164	0.172	0.185	0.152	0.164
LAV	None	None	0.136	0.137	0.140	0.130	0.129
	25%	-0.25	0.140	0.139	0.147	0.129	0.134
		-0.50	0.143	0.145	0.150	0.133	0.137
		-1.00	0.144	0.150	0.158	0.136	0.143
	50%	-0.25	0.141	0.143	0.146	0.129	0.137
		-0.50	0.149	0.155	0.162	0.140	0.146
		-1.00	0.187	0.192	0.199	0.170	0.177
CC	None	None	0.119	0.111	0.108	0.107	0.101
	25%	-0.25	0.120	0.110	0.106	0.106	0.104
		-0.50	0.119	0.111	0.106	0.106	0.102
		-1.00	0.120	0.111	0.106	0.105	0.102
	50%	-0.25	0.119	0.111	0.105	0.105	0.103
		-0.50	0.117	0.110	0.106	0.105	0.103
		-1.00	0.116	0.107	0.104	0.103	0.100
FPC	None	None	0.128	0.121	0.121	0.119	0.116
	25%	-0.25	0.129	0.121	0.119	0.119	0.118
		-0.50	0.130	0.122	0.120	0.119	0.118
		-1.00	0.131	0.124	0.120	0.118	0.118
	50%	-0.25	0.129	0.122	0.120	0.117	0.118
		-0.50	0.127	0.123	0.121	0.119	0.118
		-1.00	0.127	0.121	0.121	0.118	0.117

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Discrimination of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.164	0.166	0.180	0.152	0.149
	25%	-0.25	0.166	0.169	0.193	0.152	0.157
		-0.50	0.174	0.179	0.202	0.156	0.161
		-1.00	0.184	0.209	0.237	0.167	0.184
	50%	-0.25	0.168	0.174	0.196	0.153	0.158
		-0.50	0.174	0.192	0.227	0.160	0.174
		-1.00	0.212	0.248	0.302	0.192	0.226
HB	None	None	0.162	0.163	0.175	0.150	0.145
	25%	-0.25	0.167	0.169	0.188	0.151	0.154
		-0.50	0.180	0.185	0.206	0.159	0.165
		-1.00	0.222	0.250	0.279	0.199	0.218
	50%	-0.25	0.170	0.177	0.194	0.153	0.159
		-0.50	0.189	0.208	0.241	0.170	0.186
		-1.00	0.297	0.343	0.400	0.265	0.308
LAV	None	None	0.163	0.164	0.173	0.153	0.147
	25%	-0.25	0.168	0.170	0.188	0.153	0.156
		-0.50	0.176	0.182	0.199	0.159	0.165
		-1.00	0.174	0.194	0.218	0.168	0.182
	50%	-0.25	0.172	0.177	0.190	0.155	0.159
		-0.50	0.197	0.216	0.245	0.178	0.194
		-1.00	0.362	0.402	0.413	0.330	0.361
CC	None	None	0.150	0.137	0.133	0.144	0.134
	25%	-0.25	0.149	0.137	0.130	0.143	0.136
		-0.50	0.148	0.136	0.132	0.142	0.136
		-1.00	0.147	0.138	0.136	0.143	0.141
	50%	-0.25	0.148	0.136	0.130	0.140	0.135
		-0.50	0.144	0.134	0.132	0.142	0.136
		-1.00	0.146	0.137	0.140	0.148	0.144
FPC	None	None	0.156	0.145	0.140	0.149	0.140
	25%	-0.25	0.156	0.144	0.138	0.149	0.143
		-0.50	0.159	0.145	0.140	0.149	0.142
		-1.00	0.161	0.150	0.140	0.150	0.143
	50%	-0.25	0.157	0.146	0.139	0.147	0.142
		-0.50	0.155	0.146	0.140	0.150	0.143
		-1.00	0.163	0.150	0.143	0.154	0.144

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Difficulty of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.160	0.179	0.199	0.174	0.186
	25%	-0.25	0.199	0.217	0.245	0.203	0.226
		-0.50	0.240	0.265	0.293	0.255	0.273
		-1.00	0.347	0.369	0.396	0.353	0.370
	50%	-0.25	0.233	0.253	0.280	0.241	0.254
		-0.50	0.329	0.353	0.385	0.337	0.362
		-1.00	0.536	0.551	0.576	0.537	0.554
HB	None	None	0.160	0.178	0.197	0.173	0.184
	25%	-0.25	0.196	0.214	0.239	0.199	0.219
		-0.50	0.231	0.255	0.281	0.245	0.260
		-1.00	0.311	0.337	0.362	0.318	0.336
	50%	-0.25	0.229	0.248	0.273	0.236	0.246
		-0.50	0.315	0.338	0.368	0.320	0.345
		-1.00	0.477	0.493	0.521	0.481	0.495
LAV	None	None	0.161	0.177	0.198	0.174	0.183
	25%	-0.25	0.189	0.208	0.235	0.196	0.215
		-0.50	0.202	0.226	0.251	0.217	0.234
		-1.00	0.226	0.246	0.277	0.237	0.257
	50%	-0.25	0.229	0.246	0.271	0.234	0.246
		-0.50	0.295	0.320	0.359	0.303	0.331
		-1.00	0.410	0.431	0.455	0.396	0.429
CC	None	None	0.161	0.167	0.173	0.170	0.172
	25%	-0.25	0.193	0.196	0.199	0.192	0.196
		-0.50	0.223	0.230	0.231	0.227	0.225
		-1.00	0.305	0.304	0.301	0.307	0.296
	50%	-0.25	0.220	0.224	0.224	0.220	0.217
		-0.50	0.298	0.298	0.295	0.298	0.293
		-1.00	0.496	0.497	0.482	0.494	0.475
FPC	None	None	0.160	0.165	0.170	0.163	0.168
	25%	-0.25	0.188	0.190	0.191	0.183	0.187
		-0.50	0.215	0.215	0.216	0.211	0.210
		-1.00	0.278	0.268	0.259	0.267	0.255
	50%	-0.25	0.213	0.214	0.215	0.209	0.204
		-0.50	0.285	0.279	0.270	0.272	0.268
		-1.00	0.437	0.419	0.399	0.411	0.390

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Difficulty of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.184	0.182	0.189	0.182	0.188
	25%	-0.25	0.184	0.177	0.193	0.179	0.191
		-0.50	0.180	0.178	0.193	0.177	0.190
		-1.00	0.176	0.177	0.188	0.170	0.187
	50%	-0.25	0.181	0.183	0.189	0.178	0.190
		-0.50	0.177	0.177	0.192	0.176	0.188
		-1.00	0.170	0.172	0.185	0.168	0.179
HB	None	None	0.185	0.183	0.191	0.183	0.189
	25%	-0.25	0.183	0.176	0.194	0.179	0.192
		-0.50	0.176	0.176	0.192	0.175	0.189
		-1.00	0.167	0.169	0.181	0.162	0.179
	50%	-0.25	0.179	0.182	0.190	0.176	0.190
		-0.50	0.172	0.173	0.189	0.173	0.185
		-1.00	0.155	0.159	0.173	0.154	0.168
LAV	None	None	0.187	0.186	0.194	0.184	0.193
	25%	-0.25	0.186	0.181	0.198	0.183	0.196
		-0.50	0.181	0.183	0.199	0.181	0.193
		-1.00	0.185	0.186	0.199	0.177	0.192
	50%	-0.25	0.181	0.186	0.195	0.178	0.195
		-0.50	0.177	0.179	0.194	0.179	0.190
		-1.00	0.167	0.168	0.186	0.162	0.168
CC	None	None	0.192	0.185	0.189	0.180	0.182
	25%	-0.25	0.191	0.181	0.190	0.178	0.188
		-0.50	0.186	0.181	0.190	0.176	0.187
		-1.00	0.180	0.176	0.187	0.171	0.184
	50%	-0.25	0.187	0.185	0.188	0.176	0.183
		-0.50	0.183	0.179	0.189	0.175	0.183
		-1.00	0.176	0.175	0.185	0.170	0.178
FPC	None	None	0.187	0.181	0.186	0.179	0.180
	25%	-0.25	0.187	0.176	0.186	0.177	0.184
		-0.50	0.181	0.177	0.186	0.174	0.183
		-1.00	0.177	0.174	0.182	0.168	0.179
	50%	-0.25	0.182	0.180	0.184	0.174	0.181
		-0.50	0.179	0.175	0.184	0.173	0.180
		-1.00	0.171	0.170	0.179	0.165	0.172

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Difficulty of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.257	0.272	0.294	0.263	0.280
	25%	-0.25	0.286	0.298	0.332	0.288	0.316
		-0.50	0.319	0.339	0.372	0.328	0.352
		-1.00	0.406	0.424	0.455	0.408	0.430
	50%	-0.25	0.313	0.332	0.360	0.317	0.337
		-0.50	0.388	0.409	0.447	0.397	0.426
		-1.00	0.570	0.586	0.615	0.570	0.590
HB	None	None	0.257	0.271	0.293	0.263	0.278
	25%	-0.25	0.284	0.295	0.327	0.284	0.310
		-0.50	0.310	0.329	0.361	0.318	0.341
		-1.00	0.373	0.393	0.420	0.374	0.396
	50%	-0.25	0.308	0.327	0.355	0.312	0.330
		-0.50	0.374	0.395	0.431	0.381	0.408
		-1.00	0.514	0.531	0.562	0.516	0.535
LAV	None	None	0.260	0.272	0.296	0.265	0.281
	25%	-0.25	0.280	0.293	0.327	0.284	0.309
		-0.50	0.290	0.310	0.341	0.300	0.323
		-1.00	0.315	0.330	0.362	0.316	0.340
	50%	-0.25	0.309	0.328	0.357	0.311	0.334
		-0.50	0.360	0.384	0.425	0.371	0.399
		-1.00	0.460	0.478	0.504	0.443	0.473
CC	None	None	0.263	0.264	0.273	0.259	0.263
	25%	-0.25	0.285	0.280	0.292	0.276	0.287
		-0.50	0.307	0.309	0.318	0.304	0.311
		-1.00	0.370	0.371	0.378	0.369	0.369
	50%	-0.25	0.304	0.307	0.313	0.298	0.301
		-0.50	0.367	0.368	0.375	0.364	0.367
		-1.00	0.535	0.538	0.532	0.535	0.520
FPC	None	None	0.259	0.258	0.266	0.253	0.259
	25%	-0.25	0.279	0.273	0.282	0.267	0.277
		-0.50	0.299	0.296	0.303	0.290	0.295
		-1.00	0.349	0.339	0.338	0.333	0.330
	50%	-0.25	0.298	0.297	0.302	0.288	0.289
		-0.50	0.356	0.349	0.350	0.343	0.343
		-1.00	0.485	0.469	0.455	0.460	0.443

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Pseudo-Guessing of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.054	0.064	0.074	0.057	0.064
	25%	-0.25	0.055	0.064	0.074	0.057	0.064
		-0.50	0.055	0.064	0.075	0.056	0.065
		-1.00	0.055	0.065	0.075	0.057	0.065
	50%	-0.25	0.055	0.063	0.074	0.056	0.065
		-0.50	0.054	0.064	0.076	0.057	0.065
		-1.00	0.054	0.065	0.075	0.057	0.066
HB	None	None	0.054	0.064	0.074	0.057	0.064
	25%	-0.25	0.055	0.064	0.074	0.057	0.064
		-0.50	0.055	0.064	0.075	0.056	0.065
		-1.00	0.055	0.065	0.075	0.057	0.065
	50%	-0.25	0.055	0.063	0.074	0.056	0.065
		-0.50	0.054	0.064	0.076	0.057	0.065
		-1.00	0.055	0.065	0.075	0.057	0.066
LAV	None	None	0.054	0.064	0.074	0.057	0.064
	25%	-0.25	0.055	0.064	0.074	0.057	0.064
		-0.50	0.055	0.064	0.075	0.056	0.065
		-1.00	0.055	0.065	0.075	0.057	0.065
	50%	-0.25	0.055	0.063	0.074	0.056	0.065
		-0.50	0.054	0.064	0.076	0.057	0.065
		-1.00	0.054	0.065	0.075	0.057	0.066
CC	None	None	0.053	0.059	0.064	0.054	0.058
	25%	-0.25	0.053	0.059	0.064	0.054	0.058
		-0.50	0.053	0.059	0.064	0.053	0.058
		-1.00	0.053	0.058	0.063	0.053	0.057
	50%	-0.25	0.053	0.058	0.064	0.053	0.058
		-0.50	0.052	0.058	0.064	0.053	0.058
		-1.00	0.051	0.056	0.061	0.051	0.056
FPC	None	None	0.054	0.060	0.065	0.055	0.059
	25%	-0.25	0.054	0.060	0.064	0.054	0.059
		-0.50	0.054	0.060	0.064	0.054	0.059
		-1.00	0.053	0.059	0.064	0.054	0.058
	50%	-0.25	0.054	0.059	0.064	0.054	0.059
		-0.50	0.053	0.058	0.064	0.054	0.059
		-1.00	0.052	0.057	0.062	0.053	0.058

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Pseudo-Guessing of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.031	0.031	0.029	0.030	0.029
	25%	-0.25	0.031	0.030	0.030	0.030	0.030
		-0.50	0.031	0.031	0.029	0.029	0.030
		-1.00	0.030	0.031	0.029	0.029	0.029
	50%	-0.25	0.031	0.031	0.030	0.030	0.029
		-0.50	0.031	0.031	0.029	0.029	0.029
		-1.00	0.030	0.031	0.029	0.029	0.029
HB	None	None	0.031	0.031	0.029	0.030	0.029
	25%	-0.25	0.031	0.030	0.030	0.030	0.030
		-0.50	0.031	0.031	0.029	0.030	0.030
		-1.00	0.030	0.031	0.029	0.029	0.029
	50%	-0.25	0.031	0.031	0.030	0.030	0.029
		-0.50	0.031	0.031	0.029	0.029	0.029
		-1.00	0.030	0.031	0.029	0.029	0.029
LAV	None	None	0.031	0.031	0.029	0.030	0.029
	25%	-0.25	0.031	0.030	0.030	0.030	0.030
		-0.50	0.031	0.031	0.029	0.029	0.030
		-1.00	0.030	0.031	0.029	0.029	0.029
	50%	-0.25	0.031	0.031	0.030	0.030	0.029
		-0.50	0.031	0.031	0.029	0.029	0.029
		-1.00	0.030	0.031	0.029	0.029	0.029
CC	None	None	0.033	0.031	0.027	0.029	0.028
	25%	-0.25	0.032	0.030	0.027	0.029	0.028
		-0.50	0.031	0.030	0.027	0.029	0.028
		-1.00	0.031	0.029	0.026	0.028	0.027
	50%	-0.25	0.032	0.030	0.027	0.029	0.027
		-0.50	0.031	0.029	0.026	0.028	0.027
		-1.00	0.029	0.028	0.026	0.027	0.026
FPC	None	None	0.032	0.030	0.027	0.029	0.027
	25%	-0.25	0.031	0.029	0.027	0.029	0.027
		-0.50	0.031	0.029	0.026	0.028	0.027
		-1.00	0.030	0.029	0.026	0.028	0.027
	50%	-0.25	0.031	0.029	0.027	0.028	0.027
		-0.50	0.030	0.028	0.026	0.028	0.027
		-1.00	0.028	0.027	0.025	0.027	0.026

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Pseudo-Guessing of all 100 Items - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.066	0.075	0.084	0.067	0.075
	25%	-0.25	0.066	0.075	0.084	0.067	0.075
		-0.50	0.067	0.076	0.085	0.067	0.076
		-1.00	0.067	0.076	0.085	0.067	0.075
	50%	-0.25	0.066	0.075	0.084	0.067	0.075
		-0.50	0.066	0.075	0.085	0.067	0.076
		-1.00	0.066	0.076	0.085	0.067	0.076
HB	None	None	0.066	0.075	0.084	0.067	0.075
	25%	-0.25	0.066	0.075	0.084	0.067	0.076
		-0.50	0.067	0.076	0.085	0.067	0.076
		-1.00	0.067	0.076	0.085	0.067	0.075
	50%	-0.25	0.066	0.075	0.084	0.067	0.075
		-0.50	0.066	0.075	0.085	0.067	0.076
		-1.00	0.066	0.076	0.085	0.067	0.076
LAV	None	None	0.066	0.075	0.084	0.067	0.075
	25%	-0.25	0.066	0.075	0.084	0.067	0.075
		-0.50	0.067	0.076	0.085	0.067	0.076
		-1.00	0.067	0.076	0.085	0.067	0.075
	50%	-0.25	0.066	0.075	0.084	0.067	0.075
		-0.50	0.066	0.075	0.085	0.067	0.076
		-1.00	0.066	0.076	0.085	0.067	0.076
CC	None	None	0.067	0.071	0.074	0.064	0.068
	25%	-0.25	0.066	0.070	0.073	0.064	0.068
		-0.50	0.066	0.070	0.073	0.063	0.068
		-1.00	0.065	0.069	0.072	0.063	0.067
	50%	-0.25	0.066	0.070	0.073	0.063	0.067
		-0.50	0.064	0.069	0.073	0.063	0.067
		-1.00	0.062	0.067	0.070	0.062	0.066
FPC	None	None	0.067	0.071	0.074	0.065	0.069
	25%	-0.25	0.066	0.070	0.073	0.065	0.068
		-0.50	0.066	0.070	0.073	0.064	0.068
		-1.00	0.065	0.069	0.072	0.064	0.067
	50%	-0.25	0.066	0.069	0.073	0.064	0.068
		-0.50	0.064	0.069	0.072	0.064	0.068
		-1.00	0.062	0.067	0.070	0.062	0.066

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Discrimination of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.056	0.067	0.094	0.054	0.048
	25%	-0.25	0.060	0.075	0.112	0.055	0.049
		-0.50	0.068	0.091	0.123	0.058	0.055
		-1.00	0.100	0.134	0.176	0.068	0.074
	50%	-0.25	0.062	0.083	0.117	0.056	0.054
		-0.50	0.082	0.110	0.149	0.060	0.061
		-1.00	0.137	0.180	0.232	0.086	0.110
HB	None	None	0.055	0.062	0.083	0.054	0.047
	25%	-0.25	0.062	0.076	0.108	0.056	0.049
		-0.50	0.084	0.107	0.133	0.063	0.063
		-1.00	0.161	0.196	0.230	0.108	0.120
	50%	-0.25	0.069	0.089	0.115	0.057	0.056
		-0.50	0.109	0.138	0.171	0.072	0.080
		-1.00	0.247	0.292	0.343	0.171	0.201
LAV	None	None	0.055	0.062	0.079	0.054	0.047
	25%	-0.25	0.060	0.073	0.102	0.056	0.050
		-0.50	0.064	0.082	0.113	0.060	0.057
		-1.00	0.067	0.090	0.123	0.065	0.062
	50%	-0.25	0.072	0.092	0.113	0.059	0.058
		-0.50	0.139	0.168	0.195	0.097	0.110
		-1.00	0.354	0.369	0.402	0.255	0.287
CC	None	None	0.045	0.041	0.038	0.059	0.050
	25%	-0.25	0.046	0.040	0.038	0.059	0.050
		-0.50	0.049	0.045	0.042	0.061	0.056
		-1.00	0.064	0.060	0.056	0.069	0.065
	50%	-0.25	0.047	0.042	0.037	0.060	0.051
		-0.50	0.053	0.048	0.044	0.062	0.057
		-1.00	0.078	0.069	0.063	0.080	0.076
FPC	None	None	0.053	0.049	0.048	0.057	0.053
	25%	-0.25	0.055	0.051	0.050	0.059	0.054
		-0.50	0.061	0.057	0.054	0.062	0.058
		-1.00	0.077	0.074	0.067	0.070	0.063
	50%	-0.25	0.057	0.055	0.050	0.061	0.057
		-0.50	0.067	0.062	0.057	0.066	0.059
		-1.00	0.093	0.082	0.077	0.079	0.072

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Discrimination of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.098	0.099	0.098	0.082	0.082
	25%	-0.25	0.100	0.099	0.101	0.083	0.084
		-0.50	0.102	0.098	0.101	0.085	0.084
		-1.00	0.104	0.105	0.109	0.087	0.087
	50%	-0.25	0.100	0.100	0.100	0.082	0.084
		-0.50	0.103	0.104	0.102	0.085	0.084
		-1.00	0.108	0.109	0.114	0.088	0.091
HB	None	None	0.098	0.097	0.097	0.082	0.081
	25%	-0.25	0.100	0.099	0.099	0.083	0.084
		-0.50	0.103	0.100	0.101	0.086	0.086
		-1.00	0.110	0.110	0.114	0.092	0.092
	50%	-0.25	0.100	0.100	0.099	0.083	0.084
		-0.50	0.106	0.106	0.104	0.087	0.087
		-1.00	0.120	0.119	0.126	0.097	0.098
LAV	None	None	0.098	0.098	0.097	0.083	0.082
	25%	-0.25	0.101	0.101	0.101	0.084	0.086
		-0.50	0.103	0.100	0.101	0.089	0.088
		-1.00	0.102	0.102	0.107	0.091	0.089
	50%	-0.25	0.104	0.102	0.099	0.086	0.087
		-0.50	0.114	0.113	0.109	0.093	0.093
		-1.00	0.138	0.145	0.141	0.111	0.113
CC	None	None	0.090	0.083	0.078	0.074	0.071
	25%	-0.25	0.091	0.084	0.078	0.074	0.072
		-0.50	0.092	0.083	0.078	0.076	0.072
		-1.00	0.092	0.085	0.079	0.076	0.072
	50%	-0.25	0.090	0.084	0.077	0.073	0.071
		-0.50	0.091	0.084	0.076	0.074	0.070
		-1.00	0.092	0.081	0.077	0.074	0.070
FPC	None	None	0.096	0.091	0.087	0.083	0.082
	25%	-0.25	0.098	0.091	0.087	0.083	0.082
		-0.50	0.098	0.090	0.087	0.085	0.082
		-1.00	0.098	0.092	0.089	0.085	0.083
	50%	-0.25	0.097	0.091	0.087	0.082	0.082
		-0.50	0.097	0.092	0.086	0.084	0.081
		-1.00	0.098	0.090	0.089	0.084	0.082

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Discrimination of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.118	0.126	0.142	0.104	0.099
	25%	-0.25	0.122	0.132	0.157	0.104	0.101
		-0.50	0.130	0.141	0.164	0.108	0.105
		-1.00	0.151	0.175	0.210	0.116	0.120
	50%	-0.25	0.125	0.137	0.159	0.104	0.104
		-0.50	0.139	0.157	0.185	0.108	0.109
		-1.00	0.181	0.214	0.261	0.129	0.147
HB	None	None	0.117	0.122	0.135	0.103	0.097
	25%	-0.25	0.124	0.132	0.152	0.104	0.101
		-0.50	0.140	0.152	0.172	0.112	0.111
		-1.00	0.201	0.228	0.259	0.148	0.156
	50%	-0.25	0.129	0.140	0.157	0.105	0.105
		-0.50	0.159	0.179	0.203	0.119	0.122
		-1.00	0.279	0.317	0.367	0.202	0.227
LAV	None	None	0.117	0.122	0.132	0.104	0.098
	25%	-0.25	0.124	0.131	0.149	0.106	0.103
		-0.50	0.128	0.136	0.157	0.113	0.110
		-1.00	0.128	0.142	0.169	0.116	0.113
	50%	-0.25	0.133	0.144	0.156	0.109	0.108
		-0.50	0.186	0.207	0.226	0.141	0.149
		-1.00	0.383	0.398	0.428	0.282	0.310
CC	None	None	0.104	0.096	0.090	0.101	0.093
	25%	-0.25	0.106	0.097	0.090	0.102	0.094
		-0.50	0.108	0.098	0.092	0.105	0.098
		-1.00	0.118	0.109	0.102	0.109	0.105
	50%	-0.25	0.105	0.097	0.089	0.102	0.095
		-0.50	0.110	0.100	0.092	0.104	0.098
		-1.00	0.127	0.112	0.106	0.116	0.111
FPC	None	None	0.114	0.108	0.104	0.108	0.103
	25%	-0.25	0.117	0.110	0.106	0.108	0.104
		-0.50	0.121	0.112	0.108	0.112	0.107
		-1.00	0.132	0.124	0.117	0.117	0.110
	50%	-0.25	0.118	0.111	0.105	0.109	0.105
		-0.50	0.125	0.117	0.109	0.113	0.106
		-1.00	0.142	0.128	0.123	0.122	0.115

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Difficulty of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.124	0.148	0.184	0.134	0.150
	25%	-0.25	0.165	0.192	0.226	0.163	0.178
		-0.50	0.218	0.249	0.277	0.204	0.226
		-1.00	0.331	0.359	0.390	0.304	0.324
	50%	-0.25	0.206	0.235	0.266	0.196	0.219
		-0.50	0.315	0.343	0.371	0.294	0.309
		-1.00	0.512	0.533	0.559	0.495	0.507
HB	None	None	0.123	0.146	0.179	0.132	0.147
	25%	-0.25	0.161	0.188	0.220	0.160	0.172
		-0.50	0.209	0.239	0.266	0.195	0.216
		-1.00	0.297	0.325	0.358	0.271	0.291
	50%	-0.25	0.201	0.230	0.260	0.191	0.212
		-0.50	0.300	0.328	0.356	0.278	0.295
		-1.00	0.456	0.478	0.506	0.438	0.452
LAV	None	None	0.124	0.146	0.178	0.133	0.148
	25%	-0.25	0.147	0.175	0.210	0.152	0.165
		-0.50	0.158	0.189	0.220	0.162	0.183
		-1.00	0.184	0.210	0.248	0.184	0.201
	50%	-0.25	0.199	0.223	0.256	0.185	0.209
		-0.50	0.262	0.291	0.326	0.236	0.264
		-1.00	0.395	0.411	0.435	0.334	0.368
CC	None	None	0.119	0.129	0.146	0.124	0.134
	25%	-0.25	0.153	0.162	0.172	0.149	0.152
		-0.50	0.195	0.207	0.209	0.179	0.187
		-1.00	0.294	0.308	0.309	0.269	0.267
	50%	-0.25	0.189	0.201	0.205	0.179	0.186
		-0.50	0.289	0.298	0.296	0.261	0.260
		-1.00	0.488	0.500	0.494	0.466	0.455
FPC	None	None	0.124	0.134	0.149	0.126	0.140
	25%	-0.25	0.153	0.162	0.173	0.147	0.154
		-0.50	0.189	0.198	0.201	0.170	0.182
		-1.00	0.263	0.265	0.266	0.231	0.232
	50%	-0.25	0.188	0.199	0.203	0.172	0.183
		-0.50	0.272	0.278	0.276	0.240	0.243
		-1.00	0.420	0.421	0.414	0.385	0.379

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Difficulty of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.146	0.147	0.153	0.138	0.145
	25%	-0.25	0.142	0.148	0.152	0.135	0.143
		-0.50	0.142	0.144	0.148	0.134	0.143
		-1.00	0.138	0.141	0.147	0.131	0.138
	50%	-0.25	0.142	0.146	0.150	0.134	0.142
		-0.50	0.140	0.144	0.148	0.133	0.140
		-1.00	0.134	0.137	0.145	0.130	0.137
HB	None	None	0.146	0.147	0.154	0.138	0.145
	25%	-0.25	0.141	0.147	0.152	0.135	0.143
		-0.50	0.139	0.142	0.146	0.132	0.140
		-1.00	0.129	0.134	0.140	0.124	0.131
	50%	-0.25	0.140	0.144	0.150	0.133	0.141
		-0.50	0.136	0.140	0.145	0.129	0.136
		-1.00	0.122	0.125	0.133	0.119	0.126
LAV	None	None	0.146	0.148	0.155	0.138	0.146
	25%	-0.25	0.144	0.150	0.155	0.136	0.145
		-0.50	0.145	0.148	0.152	0.136	0.144
		-1.00	0.145	0.148	0.157	0.135	0.142
	50%	-0.25	0.143	0.146	0.153	0.135	0.144
		-0.50	0.139	0.143	0.147	0.130	0.137
		-1.00	0.123	0.138	0.140	0.120	0.127
CC	None	None	0.147	0.144	0.146	0.136	0.138
	25%	-0.25	0.144	0.146	0.147	0.132	0.138
		-0.50	0.143	0.142	0.144	0.132	0.135
		-1.00	0.137	0.137	0.141	0.129	0.134
	50%	-0.25	0.143	0.142	0.145	0.131	0.136
		-0.50	0.140	0.141	0.142	0.129	0.134
		-1.00	0.133	0.134	0.140	0.127	0.132
FPC	None	None	0.146	0.143	0.146	0.136	0.138
	25%	-0.25	0.143	0.145	0.145	0.133	0.136
		-0.50	0.141	0.140	0.142	0.132	0.134
		-1.00	0.137	0.136	0.139	0.128	0.131
	50%	-0.25	0.142	0.140	0.142	0.132	0.136
		-0.50	0.139	0.139	0.140	0.128	0.132
		-1.00	0.132	0.132	0.136	0.123	0.128

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Difficulty of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.205	0.224	0.254	0.202	0.220
	25%	-0.25	0.233	0.260	0.289	0.225	0.244
		-0.50	0.275	0.302	0.328	0.259	0.282
		-1.00	0.369	0.395	0.427	0.346	0.365
	50%	-0.25	0.265	0.293	0.323	0.252	0.276
		-0.50	0.354	0.381	0.410	0.336	0.353
		-1.00	0.534	0.555	0.583	0.517	0.531
HB	None	None	0.205	0.222	0.251	0.200	0.217
	25%	-0.25	0.230	0.256	0.284	0.222	0.239
		-0.50	0.267	0.293	0.317	0.250	0.272
		-1.00	0.339	0.364	0.395	0.314	0.333
	50%	-0.25	0.261	0.288	0.316	0.247	0.270
		-0.50	0.340	0.367	0.396	0.321	0.339
		-1.00	0.480	0.503	0.532	0.463	0.479
LAV	None	None	0.206	0.222	0.252	0.201	0.219
	25%	-0.25	0.222	0.248	0.277	0.217	0.234
		-0.50	0.231	0.257	0.283	0.225	0.247
		-1.00	0.256	0.278	0.312	0.247	0.264
	50%	-0.25	0.261	0.284	0.315	0.243	0.268
		-0.50	0.311	0.338	0.370	0.285	0.311
		-1.00	0.425	0.444	0.468	0.369	0.402
CC	None	None	0.201	0.205	0.218	0.193	0.203
	25%	-0.25	0.225	0.234	0.242	0.211	0.218
		-0.50	0.257	0.267	0.271	0.237	0.245
		-1.00	0.336	0.350	0.355	0.315	0.315
	50%	-0.25	0.253	0.264	0.269	0.234	0.244
		-0.50	0.334	0.344	0.344	0.307	0.309
		-1.00	0.512	0.526	0.525	0.494	0.485
FPC	None	None	0.205	0.208	0.221	0.196	0.208
	25%	-0.25	0.226	0.235	0.241	0.211	0.219
		-0.50	0.253	0.260	0.263	0.231	0.241
		-1.00	0.312	0.315	0.317	0.282	0.284
	50%	-0.25	0.253	0.261	0.265	0.232	0.242
		-0.50	0.322	0.327	0.327	0.290	0.293
		-1.00	0.455	0.456	0.450	0.421	0.416

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Pseudo-Guessing of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.042	0.053	0.065	0.041	0.049
	25%	-0.25	0.042	0.052	0.065	0.041	0.049
		-0.50	0.042	0.053	0.064	0.041	0.049
		-1.00	0.043	0.053	0.066	0.042	0.049
	50%	-0.25	0.042	0.053	0.065	0.041	0.049
		-0.50	0.043	0.053	0.065	0.042	0.049
		-1.00	0.043	0.054	0.065	0.042	0.050
HB	None	None	0.042	0.053	0.065	0.041	0.049
	25%	-0.25	0.042	0.052	0.065	0.041	0.049
		-0.50	0.042	0.053	0.064	0.041	0.049
		-1.00	0.043	0.053	0.066	0.042	0.049
	50%	-0.25	0.042	0.053	0.065	0.041	0.049
		-0.50	0.043	0.053	0.065	0.042	0.049
		-1.00	0.043	0.054	0.065	0.042	0.050
LAV	None	None	0.042	0.053	0.065	0.041	0.049
	25%	-0.25	0.042	0.052	0.065	0.041	0.049
		-0.50	0.042	0.053	0.064	0.041	0.049
		-1.00	0.043	0.053	0.066	0.042	0.049
	50%	-0.25	0.042	0.053	0.065	0.041	0.049
		-0.50	0.043	0.053	0.065	0.042	0.049
		-1.00	0.043	0.054	0.065	0.042	0.050
CC	None	None	0.041	0.048	0.054	0.038	0.045
	25%	-0.25	0.041	0.047	0.053	0.039	0.045
		-0.50	0.040	0.046	0.053	0.038	0.044
		-1.00	0.040	0.046	0.053	0.039	0.044
	50%	-0.25	0.040	0.047	0.053	0.039	0.044
		-0.50	0.040	0.046	0.052	0.039	0.043
		-1.00	0.039	0.045	0.051	0.039	0.044
FPC	None	None	0.043	0.049	0.055	0.040	0.046
	25%	-0.25	0.042	0.048	0.054	0.040	0.046
		-0.50	0.042	0.047	0.054	0.040	0.045
		-1.00	0.041	0.047	0.054	0.040	0.045
	50%	-0.25	0.042	0.048	0.054	0.040	0.045
		-0.50	0.041	0.047	0.053	0.040	0.045
		-1.00	0.040	0.046	0.052	0.040	0.045

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Pseudo-Guessing of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.034	0.035	0.035	0.031	0.032
	25%	-0.25	0.033	0.035	0.036	0.031	0.032
		-0.50	0.034	0.035	0.035	0.031	0.032
		-1.00	0.034	0.035	0.036	0.031	0.031
	50%	-0.25	0.033	0.035	0.036	0.031	0.032
		-0.50	0.033	0.036	0.035	0.031	0.032
		-1.00	0.033	0.035	0.035	0.031	0.032
HB	None	None	0.034	0.035	0.035	0.031	0.032
	25%	-0.25	0.033	0.035	0.036	0.031	0.032
		-0.50	0.034	0.035	0.035	0.031	0.032
		-1.00	0.034	0.035	0.036	0.031	0.031
	50%	-0.25	0.033	0.035	0.036	0.031	0.032
		-0.50	0.033	0.036	0.035	0.031	0.032
		-1.00	0.033	0.035	0.035	0.031	0.032
LAV	None	None	0.034	0.035	0.035	0.031	0.032
	25%	-0.25	0.033	0.035	0.036	0.031	0.032
		-0.50	0.034	0.035	0.035	0.031	0.032
		-1.00	0.034	0.035	0.036	0.031	0.031
	50%	-0.25	0.033	0.035	0.036	0.031	0.032
		-0.50	0.033	0.036	0.035	0.031	0.032
		-1.00	0.033	0.035	0.035	0.031	0.032
CC	None	None	0.033	0.033	0.032	0.030	0.030
	25%	-0.25	0.033	0.034	0.032	0.030	0.030
		-0.50	0.033	0.033	0.032	0.030	0.030
		-1.00	0.033	0.033	0.031	0.030	0.029
	50%	-0.25	0.033	0.033	0.032	0.030	0.030
		-0.50	0.033	0.033	0.031	0.030	0.030
		-1.00	0.032	0.032	0.031	0.029	0.029
FPC	None	None	0.034	0.033	0.032	0.031	0.031
	25%	-0.25	0.033	0.034	0.032	0.030	0.031
		-0.50	0.033	0.033	0.032	0.031	0.030
		-1.00	0.033	0.033	0.031	0.030	0.030
	50%	-0.25	0.033	0.033	0.032	0.030	0.030
		-0.50	0.033	0.033	0.031	0.030	0.030
		-1.00	0.032	0.032	0.031	0.029	0.029

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Pseudo-Guessing of all 100 Items - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.059	0.069	0.079	0.055	0.063
	25%	-0.25	0.059	0.069	0.080	0.055	0.063
		-0.50	0.059	0.069	0.079	0.055	0.062
		-1.00	0.060	0.070	0.081	0.056	0.062
	50%	-0.25	0.059	0.069	0.080	0.055	0.062
		-0.50	0.059	0.070	0.080	0.055	0.062
		-1.00	0.059	0.070	0.080	0.055	0.063
HB	None	None	0.059	0.069	0.079	0.055	0.063
	25%	-0.25	0.059	0.069	0.080	0.055	0.063
		-0.50	0.059	0.069	0.079	0.055	0.062
		-1.00	0.060	0.070	0.081	0.056	0.062
	50%	-0.25	0.059	0.069	0.080	0.055	0.062
		-0.50	0.059	0.070	0.080	0.055	0.062
		-1.00	0.059	0.070	0.080	0.055	0.063
LAV	None	None	0.059	0.069	0.079	0.055	0.063
	25%	-0.25	0.059	0.069	0.080	0.055	0.063
		-0.50	0.059	0.069	0.079	0.055	0.062
		-1.00	0.060	0.070	0.081	0.056	0.062
	50%	-0.25	0.059	0.069	0.080	0.055	0.062
		-0.50	0.059	0.070	0.080	0.055	0.062
		-1.00	0.059	0.070	0.080	0.055	0.063
CC	None	None	0.058	0.063	0.068	0.053	0.058
	25%	-0.25	0.058	0.063	0.067	0.053	0.058
		-0.50	0.057	0.062	0.067	0.053	0.057
		-1.00	0.056	0.061	0.066	0.053	0.057
	50%	-0.25	0.057	0.062	0.067	0.053	0.057
		-0.50	0.056	0.062	0.066	0.053	0.057
		-1.00	0.055	0.060	0.064	0.052	0.057
FPC	None	None	0.059	0.064	0.069	0.054	0.059
	25%	-0.25	0.059	0.064	0.068	0.054	0.059
		-0.50	0.058	0.063	0.067	0.054	0.058
		-1.00	0.057	0.062	0.067	0.054	0.058
	50%	-0.25	0.058	0.063	0.068	0.054	0.059
		-0.50	0.057	0.063	0.066	0.054	0.058
		-1.00	0.055	0.060	0.065	0.053	0.058

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

APPENDIX E

BIAS, SE, RMSE VALUES FOR EQUATED SCORES

Bias for Equated True Scores - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.118	0.268	0.540	0.188	0.157
	25%	-0.25	0.551	0.503	0.686	0.605	0.439
		-0.50	1.000	0.898	0.969	1.149	0.873
		-1.00	1.867	1.698	1.670	1.959	1.665
	50%	-0.25	1.001	0.857	0.934	1.076	0.771
		-0.50	1.892	1.700	1.712	1.987	1.750
		-1.00	3.681	3.413	3.304	3.668	3.400
HB	None	None	0.096	0.251	0.554	0.212	0.241
	25%	-0.25	0.520	0.474	0.614	0.527	0.320
		-0.50	0.945	0.851	0.889	1.010	0.730
		-1.00	1.751	1.677	1.616	1.718	1.518
	50%	-0.25	0.947	0.810	0.864	0.999	0.646
		-0.50	1.792	1.640	1.614	1.821	1.598
		-1.00	3.528	3.382	3.307	3.426	3.237
LAV	None	None	0.096	0.255	0.550	0.198	0.252
	25%	-0.25	0.395	0.396	0.580	0.456	0.259
		-0.50	0.474	0.482	0.628	0.538	0.388
		-1.00	0.246	0.467	0.677	0.457	0.472
	50%	-0.25	0.946	0.767	0.821	0.973	0.652
		-0.50	1.572	1.476	1.522	1.610	1.449
		-1.00	3.270	3.111	2.726	2.973	2.895
CC	None	None	0.448	0.514	0.721	0.511	0.563
	25%	-0.25	0.538	0.535	0.652	0.581	0.514
		-0.50	0.978	0.861	0.819	1.175	0.887
		-1.00	1.960	1.879	1.645	2.199	1.890
	50%	-0.25	0.986	0.842	0.812	1.103	0.762
		-0.50	1.996	1.834	1.588	2.109	1.839
		-1.00	4.136	3.979	3.624	4.240	3.851
FPC	None	None	0.203	0.256	0.477	0.352	0.364
	25%	-0.25	0.517	0.313	0.315	0.517	0.273
		-0.50	0.968	0.765	0.519	1.046	0.727
		-1.00	1.836	1.551	1.218	1.833	1.461
	50%	-0.25	0.991	0.777	0.536	1.026	0.658
		-0.50	1.932	1.682	1.372	1.924	1.613
		-1.00	3.763	3.383	2.933	3.587	3.151

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Equated True Scores - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.549	0.529	0.533	0.552	0.569
	25%	-0.25	0.541	0.491	0.549	0.507	0.546
		-0.50	0.528	0.506	0.593	0.584	0.559
		-1.00	0.553	0.580	0.594	0.513	0.625
	50%	-0.25	0.509	0.571	0.572	0.534	0.564
		-0.50	0.512	0.568	0.620	0.556	0.603
		-1.00	0.537	0.556	0.621	0.541	0.614
HB	None	None	0.540	0.519	0.550	0.564	0.568
	25%	-0.25	0.532	0.483	0.560	0.512	0.541
		-0.50	0.532	0.513	0.609	0.567	0.577
		-1.00	0.569	0.611	0.612	0.533	0.616
	50%	-0.25	0.489	0.569	0.572	0.513	0.584
		-0.50	0.512	0.577	0.637	0.559	0.594
		-1.00	0.546	0.583	0.699	0.549	0.671
LAV	None	None	0.604	0.606	0.602	0.613	0.648
	25%	-0.25	0.645	0.623	0.677	0.669	0.679
		-0.50	0.664	0.710	0.752	0.735	0.700
		-1.00	0.700	0.740	0.762	0.732	0.789
	50%	-0.25	0.579	0.692	0.694	0.612	0.736
		-0.50	0.735	0.780	0.836	0.798	0.820
		-1.00	1.107	1.010	1.114	1.016	0.929
CC	None	None	0.459	0.441	0.456	0.454	0.468
	25%	-0.25	0.457	0.414	0.440	0.421	0.442
		-0.50	0.440	0.443	0.479	0.483	0.464
		-1.00	0.490	0.474	0.465	0.443	0.486
	50%	-0.25	0.425	0.476	0.466	0.424	0.436
		-0.50	0.434	0.471	0.468	0.459	0.465
		-1.00	0.470	0.457	0.509	0.441	0.483
FPC	None	None	0.477	0.463	0.492	0.471	0.480
	25%	-0.25	0.473	0.432	0.461	0.445	0.474
		-0.50	0.452	0.471	0.516	0.501	0.486
		-1.00	0.496	0.498	0.501	0.459	0.516
	50%	-0.25	0.444	0.497	0.496	0.448	0.469
		-0.50	0.457	0.499	0.493	0.489	0.500
		-1.00	0.484	0.471	0.536	0.462	0.518

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Equated True Scores - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.577	0.619	0.825	0.600	0.600
	25%	-0.25	0.797	0.736	0.922	0.795	0.727
		-0.50	1.181	1.079	1.189	1.308	1.078
		-1.00	1.991	1.855	1.829	2.052	1.841
	50%	-0.25	1.158	1.086	1.148	1.217	1.001
		-0.50	1.995	1.853	1.885	2.084	1.897
		-1.00	3.758	3.506	3.414	3.729	3.501
HB	None	None	0.560	0.603	0.852	0.614	0.622
	25%	-0.25	0.778	0.707	0.877	0.741	0.647
		-0.50	1.153	1.036	1.126	1.191	0.971
		-1.00	1.905	1.836	1.774	1.859	1.698
	50%	-0.25	1.108	1.047	1.089	1.145	0.918
		-0.50	1.919	1.804	1.798	1.945	1.760
		-1.00	3.633	3.483	3.444	3.522	3.379
LAV	None	None	0.623	0.684	0.892	0.657	0.702
	25%	-0.25	0.782	0.763	0.942	0.817	0.738
		-0.50	0.859	0.888	1.032	0.935	0.819
		-1.00	0.751	0.921	1.096	0.895	0.944
	50%	-0.25	1.160	1.089	1.133	1.173	1.030
		-0.50	1.818	1.753	1.806	1.871	1.735
		-1.00	3.507	3.319	2.999	3.206	3.093
CC	None	None	0.658	0.696	0.878	0.712	0.758
	25%	-0.25	0.754	0.709	0.809	0.781	0.733
		-0.50	1.088	1.026	1.005	1.277	1.042
		-1.00	2.029	1.950	1.756	2.245	1.958
	50%	-0.25	1.090	1.030	0.982	1.189	0.949
		-0.50	2.049	1.909	1.712	2.163	1.907
		-1.00	4.167	4.010	3.670	4.265	3.883
FPC	None	None	0.526	0.547	0.723	0.609	0.631
	25%	-0.25	0.709	0.565	0.573	0.701	0.577
		-0.50	1.080	0.916	0.782	1.161	0.878
		-1.00	1.913	1.644	1.341	1.893	1.554
	50%	-0.25	1.094	0.943	0.781	1.121	0.815
		-0.50	1.994	1.766	1.480	1.987	1.692
		-1.00	3.802	3.422	2.990	3.620	3.197

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Equated Observed Scores - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.142	0.236	0.559	0.224	0.129
	25%	-0.25	0.542	0.411	0.599	0.620	0.371
		-0.50	0.970	0.753	0.797	1.140	0.773
		-1.00	1.779	1.498	1.351	1.901	1.508
	50%	-0.25	0.964	0.718	0.775	1.056	0.657
		-0.50	1.777	1.493	1.386	1.877	1.565
		-1.00	3.438	3.077	2.782	3.422	3.069
HB	None	None	0.114	0.242	0.603	0.218	0.227
	25%	-0.25	0.521	0.386	0.564	0.553	0.268
		-0.50	0.951	0.720	0.730	1.040	0.661
		-1.00	1.747	1.530	1.307	1.760	1.437
	50%	-0.25	0.931	0.682	0.724	1.002	0.550
		-0.50	1.742	1.475	1.305	1.783	1.466
		-1.00	3.419	3.148	2.886	3.324	3.030
LAV	None	None	0.116	0.255	0.611	0.219	0.243
	25%	-0.25	0.407	0.325	0.554	0.486	0.218
		-0.50	0.499	0.389	0.563	0.597	0.339
		-1.00	0.246	0.369	0.626	0.519	0.393
	50%	-0.25	0.931	0.641	0.707	0.985	0.551
		-0.50	1.555	1.332	1.227	1.610	1.342
		-1.00	3.201	2.911	2.350	2.943	2.747
CC	None	None	0.414	0.513	0.737	0.395	0.464
	25%	-0.25	0.417	0.458	0.604	0.507	0.379
		-0.50	0.836	0.695	0.675	1.069	0.747
		-1.00	1.778	1.648	1.357	2.024	1.667
	50%	-0.25	0.835	0.669	0.673	0.994	0.618
		-0.50	1.757	1.575	1.300	1.900	1.608
		-1.00	3.767	3.543	3.134	3.830	3.414
FPC	None	None	0.153	0.254	0.490	0.321	0.305
	25%	-0.25	0.453	0.251	0.286	0.522	0.276
		-0.50	0.892	0.673	0.416	1.038	0.704
		-1.00	1.717	1.423	1.066	1.781	1.392
	50%	-0.25	0.908	0.681	0.422	1.005	0.622
		-0.50	1.776	1.513	1.201	1.817	1.507
		-1.00	3.486	3.099	2.641	3.356	2.919

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Equated Observed Scores - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.500	0.499	0.509	0.525	0.544
	25%	-0.25	0.505	0.456	0.516	0.479	0.511
		-0.50	0.484	0.480	0.571	0.554	0.518
		-1.00	0.519	0.529	0.552	0.475	0.579
	50%	-0.25	0.477	0.532	0.536	0.495	0.533
		-0.50	0.481	0.529	0.583	0.523	0.563
		-1.00	0.482	0.511	0.571	0.483	0.559
HB	None	None	0.490	0.493	0.528	0.534	0.544
	25%	-0.25	0.498	0.454	0.530	0.486	0.511
		-0.50	0.491	0.485	0.586	0.535	0.536
		-1.00	0.531	0.563	0.566	0.491	0.576
	50%	-0.25	0.460	0.532	0.540	0.473	0.556
		-0.50	0.481	0.538	0.596	0.522	0.556
		-1.00	0.491	0.531	0.646	0.490	0.601
LAV	None	None	0.551	0.576	0.576	0.584	0.620
	25%	-0.25	0.600	0.588	0.649	0.630	0.641
		-0.50	0.617	0.678	0.730	0.702	0.657
		-1.00	0.661	0.699	0.725	0.694	0.758
	50%	-0.25	0.545	0.646	0.658	0.566	0.705
		-0.50	0.679	0.716	0.775	0.735	0.753
		-1.00	1.025	0.930	1.043	0.918	0.841
CC	None	None	0.430	0.427	0.441	0.435	0.454
	25%	-0.25	0.426	0.394	0.422	0.401	0.417
		-0.50	0.409	0.421	0.467	0.464	0.436
		-1.00	0.469	0.446	0.439	0.416	0.460
	50%	-0.25	0.397	0.450	0.442	0.397	0.419
		-0.50	0.413	0.447	0.450	0.438	0.449
		-1.00	0.428	0.426	0.491	0.405	0.442
FPC	None	None	0.444	0.443	0.474	0.456	0.467
	25%	-0.25	0.442	0.406	0.442	0.426	0.450
		-0.50	0.426	0.447	0.497	0.479	0.461
		-1.00	0.471	0.470	0.473	0.426	0.488
	50%	-0.25	0.421	0.471	0.465	0.416	0.452
		-0.50	0.432	0.476	0.475	0.461	0.478
		-1.00	0.440	0.437	0.511	0.420	0.475

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Equated Observed Scores - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.536	0.571	0.813	0.590	0.566
	25%	-0.25	0.760	0.647	0.826	0.787	0.661
		-0.50	1.120	0.950	1.031	1.278	0.970
		-1.00	1.883	1.642	1.524	1.975	1.665
	50%	-0.25	1.100	0.955	0.991	1.175	0.899
		-0.50	1.863	1.640	1.580	1.960	1.701
		-1.00	3.493	3.155	2.899	3.465	3.152
HB	None	None	0.516	0.570	0.862	0.588	0.593
	25%	-0.25	0.748	0.627	0.810	0.741	0.597
		-0.50	1.125	0.917	0.982	1.194	0.887
		-1.00	1.877	1.672	1.481	1.875	1.593
	50%	-0.25	1.070	0.925	0.950	1.122	0.835
		-0.50	1.849	1.626	1.508	1.883	1.611
		-1.00	3.502	3.230	3.013	3.401	3.144
LAV	None	None	0.575	0.651	0.906	0.636	0.670
	25%	-0.25	0.747	0.692	0.893	0.802	0.688
		-0.50	0.832	0.809	0.960	0.944	0.757
		-1.00	0.713	0.826	1.019	0.902	0.878
	50%	-0.25	1.117	0.969	1.014	1.151	0.947
		-0.50	1.761	1.581	1.527	1.825	1.594
		-1.00	3.404	3.091	2.623	3.138	2.912
CC	None	None	0.608	0.681	0.879	0.621	0.679
	25%	-0.25	0.655	0.633	0.755	0.693	0.626
		-0.50	0.951	0.882	0.881	1.172	0.902
		-1.00	1.847	1.722	1.481	2.068	1.737
	50%	-0.25	0.947	0.886	0.852	1.075	0.815
		-0.50	1.814	1.659	1.447	1.955	1.682
		-1.00	3.795	3.573	3.186	3.852	3.444
FPC	None	None	0.478	0.524	0.711	0.574	0.584
	25%	-0.25	0.642	0.509	0.537	0.687	0.549
		-0.50	0.997	0.826	0.707	1.143	0.843
		-1.00	1.787	1.509	1.189	1.833	1.478
	50%	-0.25	1.008	0.849	0.688	1.089	0.773
		-0.50	1.834	1.598	1.317	1.877	1.584
		-1.00	3.518	3.133	2.698	3.384	2.960

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Equated True Scores - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.048	0.240	0.480	0.417	0.182
	25%	-0.25	0.519	0.487	0.654	0.849	0.611
		-0.50	1.003	0.907	0.942	1.308	1.114
		-1.00	1.890	1.731	1.726	2.222	1.934
	50%	-0.25	0.975	0.874	0.939	1.292	1.105
		-0.50	1.920	1.737	1.710	2.266	2.002
		-1.00	3.714	3.469	3.328	4.035	3.719
HB	None	None	0.024	0.246	0.484	0.422	0.212
	25%	-0.25	0.479	0.449	0.601	0.780	0.512
		-0.50	0.939	0.858	0.874	1.173	0.952
		-1.00	1.806	1.709	1.665	1.867	1.605
	50%	-0.25	0.925	0.828	0.877	1.204	0.996
		-0.50	1.819	1.666	1.632	2.116	1.842
		-1.00	3.547	3.391	3.306	3.665	3.411
LAV	None	None	0.036	0.243	0.491	0.413	0.192
	25%	-0.25	0.261	0.318	0.532	0.590	0.357
		-0.50	0.223	0.308	0.525	0.473	0.325
		-1.00	0.207	0.380	0.575	0.413	0.285
	50%	-0.25	0.900	0.762	0.839	1.098	0.930
		-0.50	1.523	1.434	1.436	1.509	1.451
		-1.00	3.523	3.130	2.950	2.924	2.926
CC	None	None	0.227	0.305	0.448	0.430	0.350
	25%	-0.25	0.505	0.466	0.497	0.802	0.605
		-0.50	1.026	0.960	0.860	1.322	1.188
		-1.00	2.059	2.061	1.943	2.481	2.254
	50%	-0.25	1.014	0.924	0.855	1.278	1.135
		-0.50	2.085	1.985	1.829	2.383	2.179
		-1.00	4.282	4.255	4.025	4.650	4.382
FPC	None	None	0.078	0.156	0.277	0.381	0.294
	25%	-0.25	0.509	0.400	0.355	0.772	0.599
		-0.50	1.005	0.894	0.725	1.232	1.104
		-1.00	1.915	1.755	1.564	2.151	1.903
	50%	-0.25	1.010	0.899	0.757	1.244	1.112
		-0.50	2.020	1.861	1.678	2.252	2.047
		-1.00	3.888	3.679	3.392	4.071	3.778

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Equated True Scores - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.318	0.387	0.364	0.335	0.373
	25%	-0.25	0.325	0.348	0.370	0.369	0.359
		-0.50	0.362	0.347	0.358	0.360	0.415
		-1.00	0.353	0.370	0.404	0.346	0.370
	50%	-0.25	0.357	0.356	0.364	0.338	0.395
		-0.50	0.349	0.360	0.356	0.332	0.391
		-1.00	0.346	0.393	0.420	0.353	0.409
HB	None	None	0.313	0.369	0.359	0.322	0.358
	25%	-0.25	0.325	0.337	0.354	0.367	0.365
		-0.50	0.345	0.344	0.356	0.358	0.423
		-1.00	0.345	0.370	0.406	0.363	0.409
	50%	-0.25	0.350	0.343	0.353	0.335	0.390
		-0.50	0.342	0.355	0.369	0.335	0.398
		-1.00	0.358	0.411	0.457	0.372	0.416
LAV	None	None	0.348	0.401	0.381	0.350	0.412
	25%	-0.25	0.415	0.432	0.440	0.447	0.474
		-0.50	0.436	0.454	0.462	0.479	0.534
		-1.00	0.438	0.448	0.524	0.474	0.506
	50%	-0.25	0.465	0.453	0.441	0.469	0.533
		-0.50	0.603	0.603	0.541	0.554	0.605
		-1.00	0.729	0.898	0.840	0.709	0.700
CC	None	None	0.282	0.341	0.310	0.299	0.320
	25%	-0.25	0.298	0.303	0.307	0.333	0.317
		-0.50	0.320	0.294	0.302	0.325	0.369
		-1.00	0.320	0.341	0.319	0.308	0.331
	50%	-0.25	0.316	0.311	0.323	0.309	0.341
		-0.50	0.310	0.309	0.313	0.299	0.337
		-1.00	0.328	0.352	0.341	0.313	0.343
FPC	None	None	0.286	0.351	0.328	0.306	0.337
	25%	-0.25	0.302	0.313	0.322	0.341	0.340
		-0.50	0.326	0.301	0.313	0.340	0.388
		-1.00	0.324	0.354	0.330	0.324	0.348
	50%	-0.25	0.317	0.318	0.341	0.320	0.355
		-0.50	0.313	0.317	0.326	0.307	0.350
		-1.00	0.331	0.349	0.374	0.319	0.358

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Equated True Scores - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.323	0.479	0.646	0.567	0.428
	25%	-0.25	0.634	0.632	0.782	0.929	0.712
		-0.50	1.100	1.014	1.047	1.358	1.197
		-1.00	1.955	1.815	1.817	2.255	1.983
	50%	-0.25	1.066	0.987	1.043	1.337	1.179
		-0.50	1.975	1.818	1.787	2.294	2.048
		-1.00	3.749	3.522	3.392	4.056	3.755
HB	None	None	0.315	0.469	0.651	0.558	0.428
	25%	-0.25	0.603	0.592	0.727	0.868	0.631
		-0.50	1.042	0.965	0.978	1.231	1.057
		-1.00	1.877	1.780	1.748	1.941	1.701
	50%	-0.25	1.021	0.938	0.978	1.252	1.076
		-0.50	1.883	1.747	1.713	2.152	1.901
		-1.00	3.597	3.450	3.375	3.713	3.471
LAV	None	None	0.350	0.494	0.674	0.575	0.466
	25%	-0.25	0.504	0.559	0.729	0.765	0.600
		-0.50	0.500	0.569	0.742	0.726	0.651
		-1.00	0.497	0.620	0.843	0.687	0.596
	50%	-0.25	1.053	0.926	0.987	1.200	1.080
		-0.50	1.680	1.594	1.575	1.678	1.629
		-1.00	3.628	3.279	3.100	3.042	3.038
CC	None	None	0.369	0.469	0.566	0.549	0.496
	25%	-0.25	0.611	0.588	0.607	0.876	0.692
		-0.50	1.089	1.033	0.944	1.363	1.248
		-1.00	2.098	2.103	1.991	2.502	2.280
	50%	-0.25	1.076	1.012	0.950	1.316	1.190
		-0.50	2.115	2.021	1.880	2.404	2.208
		-1.00	4.300	4.274	4.047	4.663	4.397
FPC	None	None	0.299	0.395	0.462	0.515	0.472
	25%	-0.25	0.605	0.540	0.503	0.848	0.693
		-0.50	1.076	0.966	0.825	1.280	1.173
		-1.00	1.961	1.811	1.619	2.179	1.939
	50%	-0.25	1.074	0.980	0.867	1.286	1.169
		-0.50	2.054	1.902	1.730	2.276	2.079
		-1.00	3.910	3.703	3.422	4.087	3.799

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Equated Observed Scores - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.044	0.261	0.545	0.486	0.203
	25%	-0.25	0.477	0.423	0.620	0.905	0.605
		-0.50	0.932	0.758	0.810	1.319	1.071
		-1.00	1.764	1.494	1.409	2.162	1.824
	50%	-0.25	0.898	0.730	0.822	1.296	1.047
		-0.50	1.757	1.479	1.402	2.166	1.853
		-1.00	3.413	3.066	2.772	3.758	3.392
HB	None	None	0.018	0.286	0.580	0.473	0.199
	25%	-0.25	0.448	0.391	0.593	0.855	0.519
		-0.50	0.906	0.719	0.748	1.240	0.962
		-1.00	1.762	1.534	1.357	1.962	1.616
	50%	-0.25	0.869	0.692	0.778	1.237	0.961
		-0.50	1.721	1.453	1.333	2.101	1.768
		-1.00	3.381	3.105	2.853	3.601	3.248
LAV	None	None	0.029	0.281	0.599	0.487	0.196
	25%	-0.25	0.245	0.299	0.561	0.696	0.392
		-0.50	0.215	0.284	0.540	0.608	0.392
		-1.00	0.159	0.383	0.627	0.541	0.305
	50%	-0.25	0.854	0.636	0.755	1.155	0.914
		-0.50	1.485	1.270	1.167	1.637	1.476
		-1.00	3.392	2.887	2.556	2.980	2.869
CC	None	None	0.224	0.335	0.501	0.424	0.312
	25%	-0.25	0.417	0.396	0.466	0.809	0.588
		-0.50	0.918	0.813	0.708	1.287	1.124
		-1.00	1.924	1.848	1.663	2.359	2.094
	50%	-0.25	0.892	0.769	0.712	1.239	1.065
		-0.50	1.882	1.729	1.530	2.240	2.010
		-1.00	3.918	3.800	3.507	4.291	3.957
FPC	None	None	0.069	0.177	0.324	0.454	0.348
	25%	-0.25	0.455	0.332	0.313	0.849	0.660
		-0.50	0.931	0.788	0.600	1.272	1.132
		-1.00	1.807	1.606	1.388	2.135	1.869
	50%	-0.25	0.924	0.784	0.622	1.272	1.126
		-0.50	1.859	1.664	1.463	2.189	1.969
		-1.00	3.593	3.347	3.045	3.852	3.543

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Equated Observed Scores - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.306	0.384	0.353	0.321	0.361
	25%	-0.25	0.315	0.339	0.351	0.348	0.353
		-0.50	0.349	0.333	0.343	0.342	0.403
		-1.00	0.338	0.353	0.386	0.331	0.356
	50%	-0.25	0.346	0.345	0.352	0.318	0.384
		-0.50	0.338	0.336	0.349	0.312	0.378
		-1.00	0.326	0.372	0.396	0.329	0.380
HB	None	None	0.301	0.367	0.350	0.311	0.346
	25%	-0.25	0.314	0.329	0.339	0.345	0.357
		-0.50	0.336	0.327	0.344	0.342	0.408
		-1.00	0.323	0.360	0.390	0.350	0.388
	50%	-0.25	0.341	0.336	0.345	0.314	0.380
		-0.50	0.332	0.332	0.360	0.316	0.383
		-1.00	0.331	0.383	0.429	0.347	0.388
LAV	None	None	0.335	0.399	0.372	0.335	0.395
	25%	-0.25	0.391	0.415	0.419	0.420	0.452
		-0.50	0.410	0.434	0.444	0.454	0.511
		-1.00	0.408	0.440	0.505	0.458	0.494
	50%	-0.25	0.441	0.431	0.421	0.438	0.513
		-0.50	0.562	0.558	0.510	0.508	0.569
		-1.00	0.670	0.831	0.784	0.651	0.639
CC	None	None	0.276	0.345	0.304	0.294	0.309
	25%	-0.25	0.285	0.297	0.304	0.324	0.317
		-0.50	0.309	0.284	0.296	0.311	0.361
		-1.00	0.304	0.338	0.310	0.297	0.313
	50%	-0.25	0.307	0.306	0.319	0.291	0.332
		-0.50	0.305	0.299	0.296	0.283	0.324
		-1.00	0.309	0.332	0.336	0.299	0.325
FPC	None	None	0.279	0.353	0.322	0.298	0.334
	25%	-0.25	0.289	0.301	0.316	0.328	0.336
		-0.50	0.318	0.291	0.302	0.328	0.374
		-1.00	0.306	0.350	0.324	0.312	0.331
	50%	-0.25	0.311	0.312	0.330	0.301	0.342
		-0.50	0.305	0.309	0.313	0.292	0.336
		-1.00	0.311	0.329	0.362	0.301	0.339

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Equated Observed Scores - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.310	0.486	0.690	0.616	0.428
	25%	-0.25	0.591	0.574	0.740	0.973	0.704
		-0.50	1.023	0.881	0.917	1.364	1.150
		-1.00	1.821	1.584	1.517	2.190	1.866
	50%	-0.25	0.985	0.861	0.928	1.335	1.121
		-0.50	1.807	1.566	1.496	2.191	1.897
		-1.00	3.439	3.113	2.850	3.775	3.421
HB	None	None	0.302	0.490	0.722	0.593	0.409
	25%	-0.25	0.569	0.540	0.709	0.927	0.632
		-0.50	1.000	0.840	0.856	1.289	1.055
		-1.00	1.821	1.604	1.462	2.022	1.697
	50%	-0.25	0.960	0.821	0.883	1.278	1.039
		-0.50	1.779	1.533	1.433	2.129	1.819
		-1.00	3.421	3.152	2.920	3.638	3.298
LAV	None	None	0.336	0.512	0.754	0.628	0.452
	25%	-0.25	0.474	0.531	0.737	0.838	0.605
		-0.50	0.471	0.536	0.737	0.811	0.669
		-1.00	0.450	0.608	0.863	0.770	0.595
	50%	-0.25	0.992	0.815	0.899	1.240	1.053
		-0.50	1.623	1.422	1.323	1.771	1.628
		-1.00	3.480	3.019	2.705	3.079	2.963
CC	None	None	0.360	0.489	0.604	0.538	0.457
	25%	-0.25	0.539	0.529	0.576	0.877	0.675
		-0.50	0.984	0.897	0.811	1.325	1.184
		-1.00	1.956	1.892	1.715	2.379	2.119
	50%	-0.25	0.960	0.877	0.824	1.273	1.120
		-0.50	1.915	1.769	1.589	2.260	2.038
		-1.00	3.934	3.818	3.532	4.303	3.973
FPC	None	None	0.288	0.405	0.487	0.571	0.508
	25%	-0.25	0.552	0.483	0.467	0.915	0.744
		-0.50	0.999	0.862	0.716	1.315	1.194
		-1.00	1.844	1.660	1.444	2.159	1.900
	50%	-0.25	0.989	0.872	0.751	1.308	1.178
		-0.50	1.892	1.705	1.514	2.210	1.999
		-1.00	3.610	3.367	3.074	3.865	3.562

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

APPENDIX F

BIAS, SE, RMSE VALUES FOR CLASSIFICATION RATES

Bias for Classification Accuracy - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	-0.003	-0.003	-0.003	-0.005	-0.002
	25%	-0.25	-0.004	-0.005	-0.006	-0.007	-0.005
		-0.50	-0.006	-0.009	-0.009	-0.010	-0.008
		-1.00	-0.015	-0.019	-0.017	-0.019	-0.015
	50%	-0.25	-0.006	-0.009	-0.009	-0.010	-0.007
		-0.50	-0.016	-0.020	-0.019	-0.020	-0.017
		-1.00	-0.049	-0.054	-0.047	-0.052	-0.043
HB	None	None	-0.003	-0.003	-0.003	-0.005	-0.002
	25%	-0.25	-0.004	-0.005	-0.005	-0.006	-0.004
		-0.50	-0.005	-0.008	-0.007	-0.009	-0.006
		-1.00	-0.009	-0.013	-0.012	-0.013	-0.011
	50%	-0.25	-0.006	-0.008	-0.008	-0.009	-0.006
		-0.50	-0.013	-0.017	-0.016	-0.017	-0.014
		-1.00	-0.040	-0.045	-0.039	-0.044	-0.035
LAV	None	None	-0.003	-0.003	-0.003	-0.005	-0.002
	25%	-0.25	-0.003	-0.005	-0.005	-0.006	-0.004
		-0.50	-0.003	-0.004	-0.005	-0.006	-0.004
		-1.00	-0.003	-0.003	-0.003	-0.004	-0.002
	50%	-0.25	-0.006	-0.008	-0.008	-0.009	-0.006
		-0.50	-0.010	-0.015	-0.015	-0.015	-0.012
		-1.00	-0.025	-0.028	-0.023	-0.024	-0.021
CC	None	None	-0.007	-0.005	-0.003	-0.006	-0.002
	25%	-0.25	-0.007	-0.006	-0.004	-0.007	-0.003
		-0.50	-0.008	-0.008	-0.006	-0.010	-0.006
		-1.00	-0.014	-0.016	-0.012	-0.018	-0.012
	50%	-0.25	-0.008	-0.008	-0.006	-0.010	-0.005
		-0.50	-0.014	-0.016	-0.012	-0.018	-0.012
		-1.00	-0.045	-0.047	-0.035	-0.050	-0.036
FPC	None	None	-0.004	-0.003	-0.002	-0.005	-0.001
	25%	-0.25	-0.005	-0.005	-0.003	-0.006	-0.002
		-0.50	-0.007	-0.007	-0.005	-0.008	-0.005
		-1.00	-0.014	-0.014	-0.010	-0.015	-0.009
	50%	-0.25	-0.007	-0.007	-0.005	-0.009	-0.004
		-0.50	-0.015	-0.015	-0.011	-0.016	-0.011
		-1.00	-0.043	-0.041	-0.029	-0.041	-0.028

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Classification Accuracy - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.003	0.002	0.002	0.002	0.002
	25%	-0.25	0.003	0.003	0.003	0.003	0.002
		-0.50	0.003	0.004	0.004	0.004	0.004
		-1.00	0.006	0.007	0.006	0.006	0.005
	50%	-0.25	0.004	0.004	0.004	0.004	0.003
		-0.50	0.005	0.007	0.006	0.006	0.005
		-1.00	0.010	0.010	0.010	0.010	0.009
HB	None	None	0.003	0.002	0.002	0.002	0.002
	25%	-0.25	0.003	0.003	0.003	0.003	0.002
		-0.50	0.003	0.004	0.004	0.004	0.003
		-1.00	0.005	0.007	0.005	0.006	0.005
	50%	-0.25	0.004	0.004	0.004	0.004	0.003
		-0.50	0.005	0.006	0.006	0.006	0.005
		-1.00	0.010	0.010	0.010	0.009	0.009
LAV	None	None	0.003	0.002	0.002	0.002	0.002
	25%	-0.25	0.003	0.003	0.003	0.003	0.003
		-0.50	0.003	0.003	0.003	0.003	0.003
		-1.00	0.003	0.003	0.002	0.003	0.002
	50%	-0.25	0.004	0.004	0.004	0.004	0.003
		-0.50	0.007	0.007	0.008	0.007	0.006
		-1.00	0.016	0.015	0.012	0.014	0.010
CC	None	None	0.002	0.001	0.001	0.001	0.001
	25%	-0.25	0.002	0.002	0.002	0.002	0.001
		-0.50	0.002	0.003	0.002	0.003	0.002
		-1.00	0.004	0.005	0.004	0.005	0.004
	50%	-0.25	0.002	0.003	0.002	0.003	0.002
		-0.50	0.004	0.005	0.004	0.004	0.003
		-1.00	0.009	0.008	0.007	0.008	0.007
FPC	None	None	0.002	0.002	0.001	0.002	0.001
	25%	-0.25	0.003	0.002	0.002	0.002	0.002
		-0.50	0.003	0.003	0.003	0.003	0.002
		-1.00	0.005	0.005	0.004	0.005	0.003
	50%	-0.25	0.003	0.003	0.002	0.003	0.002
		-0.50	0.005	0.005	0.004	0.005	0.004
		-1.00	0.009	0.008	0.007	0.008	0.007

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Classification Accuracy - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.004	0.004	0.003	0.005	0.003
	25%	-0.25	0.005	0.006	0.007	0.007	0.005
		-0.50	0.007	0.010	0.010	0.011	0.008
		-1.00	0.016	0.021	0.018	0.020	0.016
	50%	-0.25	0.007	0.010	0.009	0.011	0.008
		-0.50	0.017	0.021	0.019	0.021	0.017
		-1.00	0.050	0.055	0.048	0.053	0.044
HB	None	None	0.004	0.004	0.003	0.005	0.002
	25%	-0.25	0.005	0.006	0.006	0.007	0.004
		-0.50	0.006	0.009	0.008	0.010	0.007
		-1.00	0.011	0.015	0.013	0.015	0.012
	50%	-0.25	0.007	0.009	0.009	0.010	0.007
		-0.50	0.014	0.019	0.017	0.018	0.015
		-1.00	0.041	0.046	0.040	0.045	0.036
LAV	None	None	0.004	0.004	0.004	0.006	0.003
	25%	-0.25	0.005	0.005	0.006	0.007	0.004
		-0.50	0.004	0.006	0.006	0.007	0.005
		-1.00	0.004	0.004	0.004	0.005	0.003
	50%	-0.25	0.007	0.009	0.009	0.010	0.007
		-0.50	0.012	0.016	0.017	0.017	0.014
		-1.00	0.029	0.032	0.026	0.028	0.023
CC	None	None	0.007	0.005	0.003	0.007	0.002
	25%	-0.25	0.007	0.006	0.004	0.008	0.004
		-0.50	0.008	0.009	0.006	0.010	0.006
		-1.00	0.015	0.017	0.013	0.019	0.013
	50%	-0.25	0.008	0.008	0.006	0.010	0.005
		-0.50	0.015	0.016	0.012	0.018	0.012
		-1.00	0.045	0.048	0.036	0.051	0.036
FPC	None	None	0.005	0.003	0.002	0.005	0.002
	25%	-0.25	0.006	0.005	0.004	0.006	0.003
		-0.50	0.007	0.008	0.006	0.009	0.005
		-1.00	0.014	0.015	0.010	0.015	0.010
	50%	-0.25	0.007	0.008	0.006	0.009	0.005
		-0.50	0.015	0.016	0.012	0.017	0.011
		-1.00	0.044	0.042	0.030	0.042	0.029

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Classification Consistency - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	-0.004	-0.005	-0.004	-0.004	-0.002
	25%	-0.25	-0.005	-0.007	-0.007	-0.006	-0.006
		-0.50	-0.004	-0.009	-0.010	-0.008	-0.009
		-1.00	-0.002	-0.011	-0.015	-0.011	-0.014
	50%	-0.25	-0.004	-0.010	-0.011	-0.009	-0.009
		-0.50	-0.003	-0.013	-0.016	-0.013	-0.015
		-1.00	0.004	-0.015	-0.025	-0.016	-0.024
HB	None	None	-0.005	-0.005	-0.004	-0.004	-0.002
	25%	-0.25	-0.004	-0.007	-0.007	-0.006	-0.005
		-0.50	-0.001	-0.007	-0.009	-0.007	-0.008
		-1.00	0.005	-0.005	-0.010	-0.004	-0.008
	50%	-0.25	-0.003	-0.009	-0.010	-0.008	-0.008
		-0.50	0.000	-0.010	-0.014	-0.010	-0.013
		-1.00	0.015	-0.005	-0.016	-0.006	-0.015
LAV	None	None	-0.005	-0.005	-0.004	-0.004	-0.002
	25%	-0.25	-0.004	-0.006	-0.007	-0.005	-0.005
		-0.50	-0.002	-0.005	-0.006	-0.004	-0.005
		-1.00	-0.003	-0.003	-0.004	-0.001	-0.002
	50%	-0.25	-0.003	-0.009	-0.010	-0.007	-0.008
		-0.50	0.001	-0.008	-0.013	-0.008	-0.011
		-1.00	0.020	0.003	-0.009	0.003	-0.007
CC	None	None	-0.012	-0.008	-0.003	-0.006	-0.002
	25%	-0.25	-0.012	-0.011	-0.006	-0.009	-0.006
		-0.50	-0.012	-0.014	-0.010	-0.012	-0.009
		-1.00	-0.011	-0.018	-0.018	-0.017	-0.017
	50%	-0.25	-0.012	-0.014	-0.010	-0.012	-0.009
		-0.50	-0.012	-0.019	-0.017	-0.018	-0.016
		-1.00	-0.008	-0.029	-0.033	-0.028	-0.031
FPC	None	None	-0.007	-0.005	-0.002	-0.003	0.000
	25%	-0.25	-0.007	-0.008	-0.005	-0.006	-0.004
		-0.50	-0.007	-0.010	-0.008	-0.008	-0.007
		-1.00	-0.006	-0.013	-0.013	-0.011	-0.011
	50%	-0.25	-0.007	-0.010	-0.008	-0.009	-0.007
		-0.50	-0.007	-0.015	-0.014	-0.013	-0.013
		-1.00	-0.002	-0.020	-0.024	-0.019	-0.023

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Classification Consistency - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.005	0.004	0.003	0.004	0.003
	25%	-0.25	0.005	0.004	0.003	0.004	0.003
		-0.50	0.005	0.004	0.004	0.004	0.003
		-1.00	0.005	0.004	0.004	0.004	0.004
	50%	-0.25	0.005	0.004	0.004	0.004	0.004
		-0.50	0.005	0.004	0.003	0.004	0.004
		-1.00	0.005	0.005	0.004	0.004	0.004
HB	None	None	0.005	0.004	0.003	0.004	0.003
	25%	-0.25	0.005	0.003	0.004	0.004	0.003
		-0.50	0.005	0.003	0.004	0.004	0.004
		-1.00	0.005	0.004	0.004	0.004	0.004
	50%	-0.25	0.004	0.004	0.004	0.004	0.004
		-0.50	0.005	0.004	0.004	0.004	0.004
		-1.00	0.004	0.004	0.004	0.004	0.004
LAV	None	None	0.006	0.004	0.004	0.004	0.004
	25%	-0.25	0.005	0.004	0.004	0.005	0.004
		-0.50	0.005	0.005	0.004	0.005	0.004
		-1.00	0.006	0.005	0.004	0.004	0.004
	50%	-0.25	0.005	0.005	0.004	0.004	0.004
		-0.50	0.006	0.005	0.005	0.006	0.005
		-1.00	0.007	0.005	0.005	0.004	0.004
CC	None	None	0.003	0.003	0.003	0.003	0.003
	25%	-0.25	0.003	0.003	0.003	0.003	0.003
		-0.50	0.003	0.003	0.003	0.003	0.003
		-1.00	0.003	0.003	0.004	0.003	0.003
	50%	-0.25	0.003	0.003	0.003	0.003	0.003
		-0.50	0.003	0.003	0.003	0.003	0.003
		-1.00	0.003	0.003	0.004	0.003	0.004
FPC	None	None	0.004	0.003	0.003	0.003	0.003
	25%	-0.25	0.004	0.003	0.003	0.004	0.003
		-0.50	0.004	0.003	0.004	0.004	0.003
		-1.00	0.004	0.004	0.004	0.003	0.004
	50%	-0.25	0.004	0.004	0.004	0.004	0.003
		-0.50	0.004	0.004	0.003	0.004	0.004
		-1.00	0.004	0.004	0.004	0.003	0.004

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Classification Consistency - 1,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.007	0.006	0.005	0.005	0.004
	25%	-0.25	0.007	0.008	0.008	0.007	0.007
		-0.50	0.006	0.010	0.011	0.009	0.010
		-1.00	0.005	0.012	0.015	0.012	0.014
	50%	-0.25	0.006	0.010	0.011	0.010	0.010
		-0.50	0.006	0.014	0.017	0.014	0.016
		-1.00	0.007	0.016	0.025	0.017	0.024
HB	None	None	0.007	0.006	0.005	0.006	0.004
	25%	-0.25	0.006	0.008	0.008	0.007	0.006
		-0.50	0.005	0.008	0.010	0.008	0.008
		-1.00	0.007	0.006	0.010	0.006	0.009
	50%	-0.25	0.005	0.009	0.010	0.009	0.009
		-0.50	0.005	0.011	0.014	0.011	0.013
		-1.00	0.015	0.006	0.016	0.007	0.016
LAV	None	None	0.007	0.006	0.006	0.006	0.004
	25%	-0.25	0.007	0.007	0.008	0.007	0.006
		-0.50	0.006	0.007	0.007	0.006	0.006
		-1.00	0.006	0.005	0.005	0.005	0.004
	50%	-0.25	0.006	0.010	0.011	0.008	0.009
		-0.50	0.006	0.010	0.014	0.010	0.012
		-1.00	0.021	0.006	0.010	0.006	0.008
CC	None	None	0.012	0.008	0.004	0.007	0.004
	25%	-0.25	0.013	0.011	0.007	0.010	0.006
		-0.50	0.012	0.014	0.011	0.013	0.010
		-1.00	0.011	0.019	0.018	0.018	0.017
	50%	-0.25	0.013	0.014	0.010	0.013	0.009
		-0.50	0.013	0.019	0.018	0.018	0.017
		-1.00	0.009	0.029	0.033	0.028	0.032
FPC	None	None	0.008	0.006	0.004	0.005	0.003
	25%	-0.25	0.008	0.008	0.006	0.007	0.005
		-0.50	0.008	0.011	0.009	0.009	0.007
		-1.00	0.007	0.014	0.013	0.012	0.012
	50%	-0.25	0.008	0.011	0.009	0.009	0.007
		-0.50	0.008	0.015	0.015	0.014	0.013
		-1.00	0.004	0.021	0.025	0.019	0.023

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Classification Accuracy - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	-0.001	-0.001	-0.002	-0.003	-0.001
	25%	-0.25	-0.002	-0.004	-0.004	-0.005	-0.003
		-0.50	-0.005	-0.008	-0.008	-0.008	-0.006
		-1.00	-0.014	-0.018	-0.017	-0.017	-0.013
	50%	-0.25	-0.005	-0.008	-0.008	-0.008	-0.006
		-0.50	-0.015	-0.019	-0.017	-0.018	-0.014
		-1.00	-0.047	-0.053	-0.045	-0.050	-0.040
HB	None	None	-0.001	-0.001	-0.002	-0.003	-0.001
	25%	-0.25	-0.002	-0.004	-0.004	-0.005	-0.003
		-0.50	-0.004	-0.006	-0.006	-0.006	-0.005
		-1.00	-0.008	-0.012	-0.012	-0.012	-0.009
	50%	-0.25	-0.004	-0.007	-0.007	-0.007	-0.005
		-0.50	-0.013	-0.016	-0.015	-0.016	-0.012
		-1.00	-0.039	-0.044	-0.038	-0.041	-0.032
LAV	None	None	-0.001	-0.001	-0.002	-0.003	-0.001
	25%	-0.25	-0.001	-0.003	-0.003	-0.004	-0.002
		-0.50	-0.001	-0.002	-0.003	-0.003	-0.001
		-1.00	0.000	-0.001	-0.001	-0.002	0.000
	50%	-0.25	-0.004	-0.006	-0.007	-0.006	-0.005
		-0.50	-0.007	-0.010	-0.011	-0.008	-0.008
		-1.00	-0.023	-0.026	-0.022	-0.015	-0.014
CC	None	None	-0.003	-0.002	-0.002	-0.004	-0.001
	25%	-0.25	-0.004	-0.004	-0.003	-0.005	-0.003
		-0.50	-0.006	-0.007	-0.006	-0.008	-0.005
		-1.00	-0.014	-0.019	-0.016	-0.020	-0.014
	50%	-0.25	-0.006	-0.007	-0.006	-0.008	-0.005
		-0.50	-0.015	-0.018	-0.015	-0.018	-0.014
		-1.00	-0.051	-0.058	-0.045	-0.058	-0.044
FPC	None	None	-0.002	-0.001	-0.001	-0.003	0.000
	25%	-0.25	-0.003	-0.004	-0.003	-0.004	-0.002
		-0.50	-0.005	-0.007	-0.006	-0.007	-0.005
		-1.00	-0.013	-0.015	-0.012	-0.015	-0.011
	50%	-0.25	-0.005	-0.007	-0.006	-0.007	-0.005
		-0.50	-0.015	-0.017	-0.014	-0.017	-0.012
		-1.00	-0.045	-0.047	-0.036	-0.046	-0.035

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Classification Accuracy - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.001	0.001	0.001	0.001	0.001
	25%	-0.25	0.002	0.002	0.002	0.002	0.002
		-0.50	0.002	0.003	0.002	0.002	0.002
		-1.00	0.004	0.004	0.003	0.004	0.003
	50%	-0.25	0.002	0.003	0.002	0.002	0.002
		-0.50	0.004	0.004	0.003	0.003	0.003
		-1.00	0.006	0.007	0.006	0.006	0.006
HB	None	None	0.001	0.001	0.001	0.001	0.001
	25%	-0.25	0.002	0.002	0.002	0.002	0.001
		-0.50	0.002	0.003	0.002	0.002	0.002
		-1.00	0.003	0.004	0.003	0.003	0.003
	50%	-0.25	0.002	0.002	0.002	0.002	0.002
		-0.50	0.004	0.004	0.003	0.003	0.003
		-1.00	0.006	0.007	0.006	0.005	0.005
LAV	None	None	0.001	0.001	0.001	0.001	0.001
	25%	-0.25	0.002	0.002	0.002	0.002	0.002
		-0.50	0.002	0.002	0.002	0.002	0.002
		-1.00	0.002	0.001	0.002	0.002	0.001
	50%	-0.25	0.003	0.003	0.003	0.002	0.002
		-0.50	0.005	0.006	0.004	0.004	0.004
		-1.00	0.010	0.013	0.010	0.008	0.007
CC	None	None	0.001	0.001	0.001	0.001	0.001
	25%	-0.25	0.001	0.001	0.001	0.001	0.001
		-0.50	0.002	0.002	0.002	0.002	0.002
		-1.00	0.003	0.004	0.003	0.004	0.003
	50%	-0.25	0.002	0.002	0.002	0.002	0.002
		-0.50	0.003	0.003	0.003	0.003	0.003
		-1.00	0.006	0.007	0.005	0.006	0.006
FPC	None	None	0.001	0.001	0.001	0.001	0.001
	25%	-0.25	0.002	0.002	0.002	0.002	0.001
		-0.50	0.002	0.002	0.002	0.002	0.002
		-1.00	0.004	0.003	0.003	0.003	0.003
	50%	-0.25	0.002	0.002	0.002	0.002	0.002
		-0.50	0.004	0.004	0.003	0.003	0.003
		-1.00	0.006	0.006	0.005	0.006	0.005

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Classification Accuracy - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.002	0.002	0.002	0.003	0.002
	25%	-0.25	0.003	0.005	0.005	0.005	0.004
		-0.50	0.006	0.008	0.008	0.008	0.006
		-1.00	0.015	0.018	0.017	0.018	0.014
	50%	-0.25	0.006	0.008	0.008	0.008	0.006
		-0.50	0.015	0.019	0.017	0.018	0.015
		-1.00	0.048	0.053	0.045	0.050	0.040
HB	None	None	0.002	0.002	0.002	0.003	0.001
	25%	-0.25	0.003	0.004	0.004	0.005	0.003
		-0.50	0.004	0.007	0.007	0.007	0.005
		-1.00	0.009	0.013	0.012	0.012	0.009
	50%	-0.25	0.005	0.007	0.007	0.007	0.006
		-0.50	0.013	0.017	0.015	0.016	0.013
		-1.00	0.039	0.045	0.038	0.041	0.033
LAV	None	None	0.002	0.002	0.002	0.003	0.001
	25%	-0.25	0.002	0.003	0.004	0.004	0.003
		-0.50	0.002	0.003	0.003	0.003	0.002
		-1.00	0.002	0.002	0.002	0.002	0.001
	50%	-0.25	0.005	0.007	0.007	0.007	0.006
		-0.50	0.008	0.012	0.012	0.009	0.009
		-1.00	0.025	0.029	0.024	0.016	0.016
CC	None	None	0.003	0.002	0.002	0.004	0.001
	25%	-0.25	0.004	0.004	0.004	0.005	0.003
		-0.50	0.006	0.008	0.006	0.008	0.006
		-1.00	0.015	0.019	0.016	0.020	0.015
	50%	-0.25	0.006	0.008	0.006	0.008	0.006
		-0.50	0.015	0.018	0.015	0.019	0.014
		-1.00	0.052	0.058	0.046	0.059	0.044
FPC	None	None	0.002	0.002	0.002	0.003	0.001
	25%	-0.25	0.003	0.004	0.004	0.005	0.003
		-0.50	0.005	0.007	0.006	0.007	0.005
		-1.00	0.014	0.016	0.013	0.016	0.011
	50%	-0.25	0.006	0.007	0.006	0.008	0.005
		-0.50	0.015	0.017	0.014	0.017	0.013
		-1.00	0.045	0.048	0.036	0.047	0.035

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

Bias for Classification Consistency - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	-0.002	-0.002	-0.003	-0.001	-0.001
	25%	-0.25	-0.002	-0.005	-0.006	-0.004	-0.004
		-0.50	-0.001	-0.007	-0.009	-0.006	-0.007
		-1.00	0.002	-0.009	-0.013	-0.009	-0.012
	50%	-0.25	-0.001	-0.007	-0.009	-0.006	-0.007
		-0.50	0.000	-0.011	-0.014	-0.011	-0.013
		-1.00	0.007	-0.013	-0.023	-0.015	-0.022
HB	None	None	-0.002	-0.002	-0.003	-0.001	-0.001
	25%	-0.25	-0.001	-0.004	-0.005	-0.003	-0.003
		-0.50	0.001	-0.005	-0.007	-0.004	-0.005
		-1.00	0.009	-0.002	-0.008	-0.002	-0.006
	50%	-0.25	0.000	-0.006	-0.008	-0.005	-0.007
		-0.50	0.003	-0.007	-0.012	-0.007	-0.011
		-1.00	0.017	-0.003	-0.014	-0.004	-0.014
LAV	None	None	-0.002	-0.002	-0.003	-0.001	-0.001
	25%	-0.25	-0.001	-0.004	-0.004	-0.002	-0.002
		-0.50	0.000	-0.002	-0.003	0.000	-0.001
		-1.00	0.000	-0.001	-0.002	0.002	0.001
	50%	-0.25	0.000	-0.006	-0.008	-0.004	-0.006
		-0.50	0.007	-0.003	-0.008	-0.001	-0.006
		-1.00	0.025	0.006	-0.005	0.008	-0.002
CC	None	None	-0.005	-0.004	-0.002	-0.002	-0.001
	25%	-0.25	-0.006	-0.007	-0.006	-0.005	-0.004
		-0.50	-0.005	-0.009	-0.009	-0.007	-0.008
		-1.00	-0.002	-0.013	-0.017	-0.012	-0.015
	50%	-0.25	-0.005	-0.010	-0.009	-0.007	-0.008
		-0.50	-0.004	-0.015	-0.017	-0.013	-0.015
		-1.00	0.002	-0.023	-0.032	-0.021	-0.030
FPC	None	None	-0.003	-0.002	-0.002	0.000	0.000
	25%	-0.25	-0.003	-0.005	-0.005	-0.003	-0.003
		-0.50	-0.002	-0.007	-0.008	-0.005	-0.006
		-1.00	0.001	-0.010	-0.013	-0.007	-0.010
	50%	-0.25	-0.002	-0.008	-0.008	-0.005	-0.006
		-0.50	-0.001	-0.012	-0.014	-0.009	-0.012
		-1.00	0.004	-0.017	-0.024	-0.014	-0.021

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

SE for Classification Consistency - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.003	0.003	0.002	0.003	0.002
	25%	-0.25	0.003	0.003	0.002	0.003	0.002
		-0.50	0.003	0.003	0.002	0.003	0.002
		-1.00	0.003	0.003	0.002	0.003	0.002
	50%	-0.25	0.003	0.003	0.002	0.002	0.002
		-0.50	0.003	0.003	0.002	0.003	0.003
		-1.00	0.003	0.003	0.003	0.003	0.003
HB	None	None	0.003	0.003	0.002	0.002	0.002
	25%	-0.25	0.003	0.003	0.002	0.002	0.002
		-0.50	0.003	0.003	0.002	0.003	0.002
		-1.00	0.003	0.002	0.002	0.003	0.002
	50%	-0.25	0.003	0.003	0.002	0.002	0.002
		-0.50	0.003	0.003	0.002	0.003	0.002
		-1.00	0.003	0.003	0.003	0.003	0.003
LAV	None	None	0.003	0.003	0.002	0.003	0.002
	25%	-0.25	0.004	0.003	0.003	0.003	0.003
		-0.50	0.004	0.003	0.002	0.003	0.003
		-1.00	0.004	0.003	0.002	0.003	0.002
	50%	-0.25	0.004	0.003	0.003	0.003	0.003
		-0.50	0.004	0.003	0.003	0.003	0.003
		-1.00	0.005	0.003	0.003	0.003	0.003
CC	None	None	0.002	0.002	0.002	0.002	0.002
	25%	-0.25	0.002	0.002	0.002	0.002	0.002
		-0.50	0.002	0.002	0.002	0.002	0.002
		-1.00	0.003	0.003	0.002	0.002	0.002
	50%	-0.25	0.003	0.002	0.002	0.002	0.002
		-0.50	0.003	0.003	0.002	0.002	0.002
		-1.00	0.003	0.003	0.003	0.003	0.003
FPC	None	None	0.003	0.002	0.002	0.002	0.002
	25%	-0.25	0.003	0.002	0.002	0.002	0.002
		-0.50	0.003	0.003	0.002	0.002	0.002
		-1.00	0.003	0.003	0.002	0.002	0.002
	50%	-0.25	0.003	0.003	0.002	0.002	0.002
		-0.50	0.003	0.003	0.002	0.002	0.002
		-1.00	0.003	0.003	0.003	0.003	0.003

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

RMSE for Classification Consistency - 3,000 Examinees

Method	Drifted Items	Drift Magnitude	Ability Distribution				
			N(0, 1)	N(0.5, 1)	N(1, 1)	S(0.5, 1)	S(1, 1)
SL	None	None	0.003	0.003	0.003	0.003	0.002
	25%	-0.25	0.003	0.006	0.006	0.005	0.005
		-0.50	0.003	0.007	0.009	0.007	0.008
		-1.00	0.004	0.009	0.013	0.009	0.012
	50%	-0.25	0.003	0.008	0.009	0.007	0.008
		-0.50	0.003	0.011	0.015	0.011	0.013
		-1.00	0.007	0.014	0.023	0.015	0.023
HB	None	None	0.003	0.004	0.003	0.003	0.002
	25%	-0.25	0.003	0.005	0.006	0.004	0.004
		-0.50	0.003	0.005	0.007	0.005	0.006
		-1.00	0.009	0.003	0.008	0.003	0.007
	50%	-0.25	0.003	0.007	0.009	0.006	0.007
		-0.50	0.005	0.008	0.012	0.008	0.011
		-1.00	0.017	0.004	0.014	0.005	0.014
LAV	None	None	0.003	0.004	0.004	0.003	0.002
	25%	-0.25	0.004	0.005	0.005	0.003	0.003
		-0.50	0.004	0.004	0.004	0.003	0.003
		-1.00	0.004	0.003	0.003	0.003	0.002
	50%	-0.25	0.004	0.006	0.009	0.005	0.007
		-0.50	0.008	0.004	0.009	0.004	0.006
		-1.00	0.025	0.007	0.006	0.008	0.004
CC	None	None	0.006	0.005	0.003	0.003	0.002
	25%	-0.25	0.006	0.007	0.006	0.005	0.005
		-0.50	0.005	0.010	0.010	0.008	0.008
		-1.00	0.003	0.014	0.017	0.012	0.015
	50%	-0.25	0.006	0.010	0.010	0.008	0.008
		-0.50	0.005	0.015	0.017	0.013	0.015
		-1.00	0.003	0.023	0.032	0.021	0.030
FPC	None	None	0.004	0.003	0.003	0.002	0.002
	25%	-0.25	0.004	0.006	0.005	0.003	0.003
		-0.50	0.003	0.008	0.008	0.005	0.006
		-1.00	0.003	0.010	0.013	0.008	0.010
	50%	-0.25	0.004	0.008	0.008	0.006	0.006
		-0.50	0.003	0.012	0.014	0.009	0.012
		-1.00	0.005	0.017	0.024	0.014	0.022

SL = Stocking Lord; HB = Haebara; LAV = Least Absolute Values; CC = Concurrent Calibration; FPC = Fixed Parameter Calibration

APPENDIX G

EMPIRICAL ITEM ESTIMATES

Base Form				New Form			
Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
1	0.999	-0.396	0.281	1	1.195	-0.386	0.247
2	1.310	-0.173	0.339	2	1.078	-0.300	0.289
3	0.916	-2.104	0.340	3	1.056	-1.982	0.306
4	0.596	-1.201	0.311	4	0.963	-0.615	0.315
5	1.210	-3.068	0.329	5	0.813	-4.258	0.318
6	0.922	-2.173	0.301	6	1.258	-1.491	0.346
7	1.183	-0.129	0.316	7	1.357	0.248	0.407
8	0.736	-2.906	0.288	8	1.016	-2.335	0.286
9	0.889	0.482	0.425	9	0.548	-0.157	0.222
10	0.626	-1.719	0.300	10	0.778	-1.391	0.279
11	0.928	-3.161	0.312	11	1.041	-3.038	0.307
12	1.035	-2.144	0.312	12	0.945	-2.293	0.303
13	0.760	-3.294	0.312	13	0.827	-2.727	0.331
14	0.868	-1.363	0.246	14	0.784	-1.280	0.266
15	0.834	-4.430	0.330	15	1.211	-3.134	0.330
16	0.860	-3.580	0.318	16	0.814	-3.455	0.328
17	0.573	-3.172	0.299	17	0.595	-3.366	0.311
18	0.744	-3.210	0.312	18	0.607	-4.221	0.313
19	0.356	-4.254	0.306	19	0.254	-5.207	0.318
20	0.789	0.309	0.204	20	0.587	0.445	0.189
21	1.179	-1.360	0.329	21	1.062	-1.374	0.408
22	1.538	-2.002	0.375	22	1.082	-2.865	0.311
23	0.555	1.400	0.270	23	0.570	0.934	0.307
24	0.868	-4.450	0.336	24	0.902	-4.473	0.329
25	0.308	-3.528	0.317	25	0.192	-6.485	0.319
26	0.463	-5.906	0.318	26	0.573	-5.075	0.316
27	1.256	-3.526	0.330	27	0.998	-4.042	0.328
28	0.835	-3.055	0.297	28	0.674	-3.323	0.313
29	0.309	3.399	0.298	29	0.295	4.135	0.318
30	0.625	-0.137	0.263	30	0.680	-0.717	0.313
31	0.285	-2.855	0.328	31	0.263	-2.183	0.319
32	0.486	-2.795	0.325	32	0.503	-2.777	0.331
33	1.232	-2.995	0.349	33	1.332	-2.849	0.284
34	0.943	-1.401	0.312	34	0.889	-1.426	0.292
35	0.722	-1.388	0.251	35	0.659	-1.271	0.269
36	0.497	-3.427	0.294	36	0.619	-2.745	0.307
37	0.971	-0.156	0.360	37	0.934	-0.061	0.435
38	0.999	-0.591	0.253	38	0.926	-0.971	0.280
39	0.817	-1.827	0.278	39	0.848	-1.602	0.300
40	0.957	-1.910	0.280	40	1.370	-1.670	0.299

Base Form				New Form			
Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
41	0.732	-1.669	0.319	41	0.767	-1.851	0.291
42	0.412	-3.281	0.305	42	0.542	-2.785	0.302
43	0.753	-1.044	0.322	43	0.856	-1.323	0.303
44	0.384	-2.124	0.312	44	0.590	-1.283	0.334
45	0.802	-0.859	0.355	45	0.613	-1.635	0.298
46	0.384	-1.746	0.311	46	0.414	-1.463	0.272
47	0.928	-0.126	0.228	47	1.065	0.395	0.327
48	0.960	-2.644	0.264	48	0.709	-3.436	0.305
49	0.669	-0.742	0.231	49	0.704	-0.822	0.240
50	0.582	-0.267	0.276	50	0.679	0.234	0.259
51	1.438	-2.216	0.330	51	1.126	-2.805	0.285
52	0.883	-1.726	0.321	52	0.866	-1.542	0.258
53	0.704	-2.399	0.325	53	0.726	-2.489	0.309
54	0.560	-2.867	0.291	54	0.776	-2.018	0.292
55	0.869	-1.027	0.334	55	0.980	-1.579	0.337
56	1.188	0.318	0.198	56	1.221	0.949	0.206
57	0.299	-6.732	0.315	57	0.363	-5.506	0.315
58	0.847	-1.574	0.402	58	0.691	-2.146	0.343
59	0.668	0.813	0.364	59	0.626	0.503	0.290
60	0.506	-1.905	0.351	60	0.451	-2.418	0.313
61	0.685	-2.075	0.296	61	0.768	-1.898	0.269
62	0.480	-2.391	0.338	62	0.553	-2.254	0.330
63	1.007	-2.701	0.308	63	1.288	-2.171	0.348
64	0.577	-0.748	0.286	64	0.457	-0.906	0.314
65	0.899	-2.517	0.334	65	0.866	-2.443	0.366
66	1.396	-1.504	0.386	66	1.343	-1.818	0.240
67	0.365	7.693	0.275	67	1.091	-0.220	0.359
68	0.469	-3.274	0.295	68	0.691	-2.680	0.316
69	0.607	-3.117	0.300	69	0.608	0.659	0.295
70	0.474	-2.058	0.304	70	0.611	-6.416	0.333
71	1.067	-0.557	0.252	71	0.557	-0.608	0.298
72	0.713	-0.102	0.236	72	0.775	-1.391	0.255
73	0.613	-3.491	0.321	73	0.613	-1.111	0.273
74	1.125	-2.811	0.317	74	0.699	-1.462	0.334
75	0.668	-0.410	0.282	75	0.962	-0.462	0.325
76	0.470	-3.902	0.303	76	1.110	-2.384	0.328
77	0.856	-1.259	0.318	77	0.867	0.750	0.286
78	0.577	-1.376	0.307	78	0.647	-1.726	0.318
79	0.568	-2.243	0.312	79	0.743	-6.774	0.353
80	0.499	-3.424	0.301	80	0.856	-4.220	0.330

Base Form				New Form			
Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
81	0.759	0.152	0.194	81	0.593	-4.787	0.317
82	1.180	-2.052	0.319	82	0.388	-2.886	0.291
83	0.408	-2.745	0.288	83	0.673	-3.950	0.308
84	0.997	-2.764	0.307	84	0.990	-2.079	0.295
85	1.018	-0.092	0.517	85	0.635	-3.327	0.317
86	0.796	-2.016	0.308	86	0.926	-1.627	0.267
87	0.365	-11.136	0.337	87	0.605	-0.346	0.270
88	0.897	1.109	0.197	88	0.790	-2.542	0.314
89	0.485	0.778	0.333	89	0.499	-2.745	0.310
90	0.213	-2.157	0.328	90	0.582	-0.767	0.256
91	0.391	-1.427	0.293	91	0.534	2.458	0.303
92	0.607	0.785	0.310	92	0.637	0.225	0.327
93	0.282	-0.455	0.302	93	0.939	-1.520	0.276
94	0.775	-0.331	0.237	94	0.697	0.884	0.226
95	0.552	1.059	0.236	95	1.042	-1.479	0.252
96	0.239	0.339	0.377	96	0.561	-1.915	0.278
97	0.893	-3.299	0.299	97	0.725	-4.211	0.312
98	0.619	-2.540	0.278	98	0.684	-3.574	0.317
99	0.484	-1.399	0.281	99	0.636	1.208	0.258
100	0.837	-0.618	0.292	100	1.303	-1.967	0.306
101	0.317	-4.077	0.313	101	0.453	-3.149	0.314
102	0.368	-2.789	0.301	102	0.186	-9.179	0.316
103	0.996	0.277	0.212	103	1.007	-2.269	0.261
104	0.231	-3.774	0.330	104	0.536	-3.126	0.310
105	1.158	-0.717	0.364	105	0.713	-1.652	0.329
106	0.752	-2.786	0.300	106	1.002	-2.603	0.327
107	0.867	1.069	0.176	107	0.681	-0.132	0.277
108	0.399	-4.599	0.310	108	0.815	-3.435	0.299
109	0.529	-3.598	0.314	109	0.905	-0.424	0.216
110	1.065	0.542	0.394	110	0.701	-1.633	0.294

Note: The first 66 highlighted rows are common items. Linked item estimates are from the SL method.