9-2013

Application of Item Response Theory Models to the Algorithmic Detection of Shift Errors on Paper and Pencil Tests

Robert Joseph Cook
University of Massachusetts Amherst

APPLICATION OF ITEM RESPONSE THEORY MODELS TO THE ALGORITHMIC DETECTION OF SHIFT ERRORS ON PAPER AND PENCIL TESTS

A Dissertation Presented

by

ROBERT JOSEPH COOK

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
Of the requirements for the degree of

DOCTOR OF EDUCATION

September 2013

Education

APPLICATION OF ITEM RESPONSE THEORY MODELS TO THE ALGORITHMIC DETECTION OF SHIFT ERRORS ON PAPER AND PENCIL TESTS

A Dissertation Presented

By

ROBERT JOSEPH COOK

Approved as to style and content by:

_____

Lisa A. Keller, Chairperson

_____

Craig S. Wells, Member

_____

Richard S. Ellis, Member

_____

Christine B. McCormick, Dean

School of Education

DEDICATION

For Karen, who makes everything possible.

ACKNOWLEDGMENTS

Getting a doctorate after stubbornly refusing to get any degree at all for the first thirty-five years of my life does not happen without some blame going around. There are so many people who share some responsibility for my getting to this point.

I would first like to thank the members of my dissertation committee, Lisa Keller, Craig Wells, and Richard Ellis, for their guidance and support as I have pursued this research agenda. Lisa has been my advisor for the last four years and my committee chair for the last two. She deserves a lot of credit for providing direction through my studies and research, patience while I found a dissertation topic I could sink my teeth into, and wisdom in allowing me the space I need to thrive. Or maybe she was just busy all the time. Either way, thanks. Much of the research I have done at UMass has been under Craig's mentorship and he has been instrumental in my learning to develop and structure research plans. Additionally, he has been like a co-advisor to me in my time here, helping my navigate through some of the challenges that are undoubtedly a part of anyone's graduate school experience. Richard was kind enough to join my committee after my taking a one-semester probability course with him but my gratitude toward him goes beyond that. His enthusiasm for the subject gave me some much needed energy at a point when I was tired of being in the classroom and his teaching gave me the confidence to pursue a dissertation topic that centered on probabilistic ideas.

One of the strengths of the measurement program at UMass is that it is like having five mentors. The three who did not serve on my committee, Ron Hambleton, Jennifer Randall, and Steve Sireci, all deserve my heartfelt thanks for sharing their knowledge and wisdom inside and outside the classroom. Outside UMass, I was fortunate enough to be able to work with Howard Wainer as a summer intern, though his mentorship has continued in the two years since. The wealth of knowledge, advice, and opportunity he has made available to me go beyond reasonable expectation and he has my gratitude.

My cohort has been like a family to me. Chris Foster, despite constantly distracting me and drinking all of my water, gets a giant thanks for handing me my dissertation topic. Chris and I have been bouncing ideas back and forth since we arrived at UMass, but when he threw the idea of detecting shift errors out, I couldn't stop thinking about it and he was kind enough to let me run with it and make it my own. Jerome Clauser, Katrina Crotts, and Amanda Soto get my thanks for being willing to play as hard as we worked. Not only do we have a wall of trivia trophies to show for it, but it turned my time at UMass into one that I will truly treasure.

I also need to acknowledge the people who got me to graduate school. Frank Padellaro has been telling me the things I don't necessarily want to hear for almost twenty years now, been unwavering in his friendship, and found the Petersham Curling Club for me, a source of great community and respite from the sometimes insanity of graduate school life. When I decided to start behaving like an adult, two people were instrumental in helping me find and stay on that course. Julia Halevy, who may think of her role in my life as a brief and difficult boss, did much to help me better understand myself and give me self-confidence at a point when that was what I needed most. At that time, I also found myself sharing a neighborhood with my cousin, Stephen Lapointe, and his penchants for kindness and honesty helped me to make much better decisions than I might have otherwise. I have a great family top to bottom, but my parents and brothers specifically have put up with a lot of stubbornness and unconventional decision-making over the years and have never stopped being patient and supportive.

My wife, Karen, uprooted her life in Texas, moved half of a country away from her family and friends, allowed us to cut our income nearly in half, dealt with my incredibly busy grad school schedule and the toll it often took on my moods for the last four years, and acted like I have somehow been doing her a big favor the whole time. I don't know how to adequately thank her for everything she does beyond dedicating this work to her and making this far too brief acknowledgment of everything she does.

Somehow, three of my dearest friends, Rob Keller, Lisa Keller, and Peter Baldwin, all found their way into field of psychometrics. I'm not sure if it was legitimate interest in - or merely a desire to understand - what they were talking about that had me follow them into this field, but I owe them my thanks for pulling me in this direction. Fraser Stowe gets an apology instead of thanks for my abandoning him as the last bastion of sanity amidst conversations of IRT, CTT, SEM, and other TLA's from which we both used to need escape.

Lastly, I would like to thank the Town of Greenfield for putting whatever they did in the water and for being a great place to live.

ABSTRACT

APPLICATION OF ITEM RESPONSE THEORY MODELS TO THE ALGORITHMIC DETECTION OF SHIFT ERRORS ON PAPER AND PENCIL TESTS

SEPTEMBER 2013

ROBERT JOSEPH COOK, B.S., UNIVERSITY OF MASSACHUSETTS LOWELL

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Lisa A. Keller

On paper-and-pencil multiple-choice tests, the potential for examinees to mark their answers in incorrect locations presents a serious threat to the validity of test score interpretations. When an examinee skips one or more items (i.e., answers out of sequence) but fails to accurately reflect the size of that skip on their answer sheet, that can trigger a string of misaligned responses called shift errors. Shift errors can result in correct answers being marked as incorrect, leading to possible underestimation of an examinee's true ability. Despite movement toward computerized testing in recent years, paper-and-pencil multiple-choice tests are still pervasive in many high-stakes assessment settings, including K-12 testing (e.g., MCAS) and college entrance exams (e.g., SAT), leaving a continuing need to address issues that arise within this format.

Techniques for detecting aberrant response patterns are well-established but do little to recognize reasons for the aberrance, limiting options for addressing the misfitting patterns. While some work has been done to detect and address specific forms of aberrant response behavior, little has been done in the area of shift error detection, leaving great room for improvement in addressing this source of aberrance. The opportunity to accurately detect construct-irrelevant errors and either adjust scores to more accurately reflect examinee ability or flag examinees with inaccurate scores for removal from the dataset and retesting would improve the validity of important decisions based on test scores, and could positively impact model fit by allowing for more accurate item parameter and ability estimation.

The purpose of this study is to investigate new algorithms for shift error detection that employ IRT models for probabilistic determination as to whether misfitting patterns are likely to be shift errors. The study examines a matrix of detection algorithms, probabilistic models, and person parameter methods, testing combinations of these factors for their selectivity (i.e., true positives vs. false positives), sensitivity (i.e., true shift errors detected vs. undetected), and robustness to parameter bias, all under a carefully manipulated, multifaceted simulation environment. This investigation attempts to provide answers to the following questions, applicable across detection methods, bias reduction procedures, shift conditions, and ability levels, but stated generally as: 1) How sensitively and selectively can an IRT-based probabilistic model detect shift error across the full range of probabilities under specific conditions?, 2) How robust is each detection method to the parameter bias introduced by shift error?, 3) How well does the detection method detect shift errors compared to other, more general, indices of person-fit?, 4) What is the impact on bias of making proposed corrections to detected shift errors?, and 4) To what extent does shift error, as detected by the method, occur within an empirical data set?

Results show that the proposed methods can indeed detect shift errors at reasonably high detection rates with only a minimal number of false positives, that detection improves when detecting longer shift errors, and that examinee ability is a huge determinant factor in the effectiveness of the shift error detection techniques. Though some detection ability is lost to person parameter bias, when detecting all but the shortest shift errors, this loss is minimal. Application to empirical data also proved effective, though some discrepancies in projected total counts suggest that refinements in the technique are required. Use of a person fit statistic to detect examinees with shift errors was shown to be completely ineffective, underscoring the value of shift-error-specific detection methods.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS v

LIST OF TABLES xiii

LIST OF FIGURES xvii

CHAPTER Page

1. INTRODUCTION 1

1.1 Background 1
1.1.1 A Brief Explanation of Shift Error Detection Methods 3
1.1.2 Person Fit Methods vs. Shift Error Detection 3
1.1.3 How Undetected Shift Errors Threaten Validity 4
1.2 Statement of Problem 5
1.3 Purpose of Study 7

2. REVIEW OF LITERATURE 9

2.1 Overview of Literature Review 9
2.2 Aberrant Test Response Behavior and Measures of Person Fit 9
2.2.1 Aberrant Response Behavior and its Impact on Validity 9
2.2.2 Overview of Person-Fit Indices 11
2.2.3 Specific Forms of Aberrant Response Behavior and Methods for their Detection 14
2.2.3.1 Detection of Cheating Behavior 21
2.2.3.2 Detection of Shift Errors. 23
2.2.4 Addressing Aberrant Test Response Behavior 288
2.3 Issues of Validity and Differential Performance Dependent on Test Format 29
2.4 Probability and Measurement Models 34
2.5 Conclusions Based on a Review of Literature 36

3. METHODOLOGY 388

3.1 Overview 388
3.2 Shift Detection Algorithms 38
3.2.1 Misaligned Response Detection 39
3.2.2 Most Probable Correction Detection 40
3.3 Probability Models 41
3.3.1 3PL Model 41
3.3.2 Nominal Response Model 42
3.4 Person Parameter Estimation Techniques 43

3.5 Decision Criteria 44
3.5.1 Probability Threshold 45
3.5.2 Person-Fit Indices 45
3.5.3 Change in Bias 46
3.6 Simulation Study Based on Empirical Data 46
3.7 Empirical Application of Simulation Study Results 50
3.8 Simulation Study Based on Stratified Ability Levels 51
3.9 Comparison of Shift Error Detection Methods to the H^T Person-Fit Statistic 53
3.10 Calibration 51

4. RESULTS 55

4.1 Overview 55
4.2 Simulation Study Based on Empirical Data 55
4.2.1 Short Shifts 55
4.2.2 Medium Shifts 58
4.2.3 Long Shifts 60
4.2.4 Mixed Length Shifts 62
4.3 Empirical Application of Simulation Study Results 64
4.3.1 Short Shifts 64
4.3.2 Medium Shifts 64
4.3.3 Long Shifts 65
4.3.4 Mixed Length Shifts 65
4.4 Simulation Study Based on Stratified Ability Levels 666
4.4.1 Short Shifts 666
4.4.2 Medium Shifts 68
4.4.3 Long Shifts 72
4.4.4 Mixed Length Shifts 74
4.4.5 Other Results 77
4.5 Comparison of Shift Error Detection Methods to the H^T Person-Fit Statistic 77

5. DISCUSSION 78

5.1 Overview 78
5.2 Simulation Study Based on Empirical Data 78
5.2.1 Shift Error Length 79
5.2.2 Shift Detection Algorithms 79
5.2.3 Probability Models 81
5.2.4 Person Parameter Estimation Methods 82
5.3 Empirical Application of Simulation Study Results 85
5.3.1 The Progressive Approach 86
5.3.2 The Mixed-Length Approach 87
5.3.3 Lack of Simulation Applicability 887
5.4 Simulation Study Based on Stratified Ability Levels 88
5.5 Comparison of Shift Error Detection Methods to the H^T Person-Fit Statistic 88
5.6 Summary of Findings 89
5.7 Implications 90
5.7.1 Empirical Application of of Shift Error Detection Methods 90
5.7.2 Simulation Studies of Shift Error Detection Methods 91
5.7.3 Person Fit Research 92
5.7.4 Fairness of Shift Error Detection and Correction 93

5.7.5 Other Applications 94
5.8 Limitations 95
5.8.1 Empirical Data 95
5.8.2 Shift Error Lengths 96
5.8.3 Bias Correction 96
5.8.4 Person Fit Comparison 97
5.8.5 Unfairness 97
5.8.6 Simulation Methods 98
5.9 Future Directions 98
5.9.1 The Multiple Choice Model 98
5.9.2 Shift Error Lengths 99
5.9.3 Shift Error Saturation Analysis 99
5.9.4 Shift Quantities and Distances 99
5.9.5 Bias Control Measures 100
5.9.6 Other Person Fit Measures 100
5.9.7 Fairness Analysis 101
5.9.8 Application to Other Empirical Data 101
5.10 Conclusion 102

APPENDICES

I: TABLES 103

II: FIGURES 123

REFERENCES 181

LIST OF TABLES

Table Page

1. Quotes on Invalidity of Aberrant Response Patterns (Petridou & Williams, 2010) 103
2. Person Fit Statistics (Meijer & Sijtsma, 2001) 104
3. Results for Single Scan Detection of Shift Errors (Skiena & Sumazin, 2000a) 104
4. Thresholds for 100% shift classification accuracy (Cook & Foster, 2012) 105
5. Shift error detection rates and thresholds, false discovery rate = .00, shift length 3 105
6. Shift error detection rates and thresholds, false discovery rate = .05, shift length 3 105
7. Mean Change in Absolute Bias, false discovery rate = .00, shift length 3 105
8. Mean Change in Signed Error, false discovery rate = .00, shift length 3 106
9. Mean Change in Absolute Bias, false discovery rate = .05, shift length 3 106
10. Mean Change in Signed Error, false discovery rate = .05, shift length 3 106
11. Shift error detection rates and thresholds, false discovery rate = .00, shift length 7 106
12. Shift error detection rates and thresholds, false discovery rate = .05, shift length 7 106
13. Mean Change in Absolute Bias, false discovery rate = .00, shift length 7 107
14. Mean Change in Signed Error, false discovery rate = .00, shift length 7 107
15. Mean Change in Absolute Bias, false discovery rate = .05, shift length 7 107
16. Mean Change in Signed Error, false discovery rate = .05, shift length 7 107
17. Shift error detection rates and thresholds, false discovery rate = .00, shift length 10 107
18. Shift detection rates and thresholds, false discovery rate = .05, shift length 10 108
19. Mean Change in Absolute Bias, false discovery rate = .00, shift length 10 108
20. Mean Change in Signed Error, false discovery rate = .00, shift length 10 108
21. Mean Change in Absolute Bias, false discovery rate = .05, shift length 10 108
22. Mean Change in Signed Error, false discovery rate = .05, shift length 10 108
23. Shift detection rates and thresholds, false discovery rate = .00, mixed length shifts 109
24. Shift detection rates and thresholds, false discovery rate = .05, mixed length shifts 109

Page

25. Mean Change in Absolute Bias, false discovery rate = .00, mixed length shifts 109
26. Mean Change in Signed Error, false discovery rate = .00, mixed length shifts 109
27. Mean Change in Absolute Bias, false discovery rate = .05, mixed length shifts 109
28. Mean Change in Signed Error, false discovery rate = .05, mixed length shifts 110
29. Empirical true positives at simulation thresholds, shift lengths of 3 or less..... 110
30. Empirical true positives at simulation thresholds, shift lengths of 7 or less..... 110
31. Empirical true positives at simulation thresholds, shift lengths of 10 or less..... 110
32. Empirical true positives at simulation thresholds, mixed shift lengths 110
33. Agreement rates between methods, false discovery rate = .05, mixed length shifts 111
34. Mean Absolute Difference, empirical data, mixed length shifts 111
35. Mean Signed Difference, empirical data, mixed length shifts 111
36. Shift error detection rates with true person parameters, FDR = .00, shift length 3..... 111
37. Shift error detection rates with true person parameters, FDR = .05, shift length 3..... 111
38. Mean Change in Absolute Bias, false discovery rate = .00, shift length 3..... 112
39. Mean Change in Signed Error, false discovery rate = .00, shift length 3..... 112
40. Mean Change in Absolute Bias, false discovery rate = .05, shift length 3..... 112
41. Mean Change in Signed Error, false discovery rate = .05, shift length 3..... 112
42. Shift error detection rates with true person parameters, FDR = .00, shift length 7..... 112
43. Shift error detection rates with true person parameters, FDR = .05, shift length 7..... 113
44. Mean Change in Absolute Bias, false discovery rate = .00, shift length 7..... 113
45. Mean Change in Signed Error, false discovery rate = .00, shift length 7..... 113
46. Mean Change in Absolute Bias, false discovery rate = .05, shift length 7..... 113
47. Mean Change in Signed Error, false discovery rate = .05, shift length 7..... 113
48. Shift detection rates with true person parameters, FDR = .00, shift length 10..... 114
49. Shift detection rates with true person parameters, FDR = .05, shift length 10..... 114

Page

50. Mean Change in Absolute Bias, false discovery rate = .00, shift length 10 114
51. Mean Change in Signed Error, false discovery rate = .00, shift length 10 114
52. Mean Change in Absolute Bias, false discovery rate = .05, shift length 10 114
53. Mean Change in Signed Error, false discovery rate = .05, shift length 10 115
54. Shift detection rates with true person parameters, FDR = .00, mixed-length shifts 115
55. Shift detection rates with true person parameters, FDR = .05, mixed-length shifts 115
56. Mean Change in Absolute Bias, false discovery rate = .00, mixed-length shifts 115
57. Mean Change in Signed Error, false discovery rate = .00, mixed-length shifts 115
58. Mean Change in Absolute Bias, false discovery rate = .05, mixed-length shifts 116
59. Mean Change in Signed Error, false discovery rate = .05, mixed-length shifts 116
60. Shift detection rates with estimated person parameters, FDR = .00, shift length 3 116
61. Shift detection rates with estimated person parameters, FDR = .05, shift length 3 116
62. Shift detection rates with estimated person parameters, FDR = .00, shift length 7 116
63. Shift detection rates with estimated person parameters, FDR = .05, shift length 7 117
64. Shift detection rates, estimated person parameters, FDR = .00, shift length 10 117
65. Shift detection rates, estimated person parameters, FDR = .05, shift length 10 117
66. Shift detection rates with estimated person parameters, FDR = .00, mixed-lengths 117
67. Shift detection rates with estimated person parameters, FDR = .05, mixed-lengths 117
68. Shift detection rates with bias-corrected parameters, FDR = .00, shift length 3 118
69. Shift detection rates with bias-corrected parameters, FDR = .05, shift length 3 118
70. Shift detection rates with bias-corrected parameters, FDR = .00, shift length 7 118
71. Shift detection rates with bias-corrected parameters, FDR = .05, shift length 7 118
72. Shift detection rates with bias-corrected parameters, FDR = .00, shift length 10 118
73. Shift detection rates with bias-corrected parameters, FDR = .05, shift length 10 119
74. Shift detection rates with bias-corrected parameters, FDR = .00, mixed-lengths 119

Page

75. Shift detection rates with bias-corrected parameters, FDR = .05, mixed-lengths 119

76. Detection rate differences between algorithms, estimated parameters, FDR = .00 119

77. Detection rate differences between algorithms, estimated parameters, FDR = .05 119

78. Detection rate differences between IRT models, estimated parameters, FDR = .00 120

79. Detection rate differences between IRT models, estimated parameters, FDR = .05 120

80. Differences between parameter estimation methods, CMP/3PL, FDR = .00 120

81. Differences between parameter estimation methods, CMP/NRM, FDR = .00 120

82. Differences between parameter estimation methods, SCIP/3PL, FDR = .00 120

83. Differences between parameter estimation methods, SCIP/NRM, FDR = .00 121

84. Differences between parameter estimation methods, CMP/3PL, FDR = .05 121

85. Differences between parameter estimation methods, CMP/NRM, FDR = .05 121

86. Differences between parameter estimation methods, SCIP/3PL, FDR = .05 121

87. Differences between parameter estimation methods, SCIP/NRM, FDR = .05 121

88. Counts and projected total shift errors in empirical data, mixed length shifts 122

LIST OF FIGURES

Figure Page

1. ROC curve using false discovery rate 123

2. Misaligned response string, misaligned 1 forward starting at item 7 123

3. Misaligned response string, misaligned 1 backward starting at item 8 123

4. ROC Curves, CMP/3PL, all person parameter methods, shift length 3 124

5. ROC Curves, CMP/NRM, all person parameter methods, shift length 3 124

6. ROC Curves, SCIP/3PL, all person parameter methods, shift length 3 125

7. ROC Curves, SCIP/NRM, all person parameter methods, shift length 3 125

8. ROC Curves, all methods, true person parameters, shift length 3 126

9. ROC Curves, all methods, estimated person parameters, shift length 3 126

10. ROC Curves, all methods , bias-corrected person parameters, shift length 3 127

11. ROC Curves, CMP/3PL, all person parameter methods, shift length 7 127

12. ROC Curves, CMP/NRM, all person parameter methods, shift length 7 128

13. ROC Curves, SCIP/3PL, all person parameter methods, shift length 7 128

14. ROC Curves, SCIP/NRM, all person parameter methods, shift length 7 129

15. ROC Curves, all methods, true person parameters, shift length 7 129

16. ROC Curves, all methods, estimated person parameters, shift length 7 130

17. ROC Curves, all methods , bias-corrected person parameters, shift length 7 130

18. ROC Curves, CMP/3PL, all person parameter methods, shift length 10 131

19. ROC Curves, CMP/NRM, all person parameter methods, shift length 10 131

20. ROC Curves, SCIP/3PL, all person parameter methods, shift length 10 132

21. ROC Curves, SCIP/NRM, all person parameter methods, shift length 10 132

22. ROC Curves, all methods, true person parameters, shift length 10 133

23. ROC Curves, all methods, estimated person parameters, shift length 10 133

24. ROC Curves, all methods , bias-corrected person parameters, shift length 10 134

Page

25. ROC Curves, CMP/3PL, all person parameter methods, mixed-length shifts 134
26. ROC Curves, CMP/NRM, all person parameter methods, mixed-length shifts 135
27. ROC Curves, SCIP/3PL, all person parameter methods, mixed-length shifts 135
28. ROC Curves, SCIP/NRM, all person parameter methods, mixed-length shifts 136
29. ROC Curves, all methods, true person parameters, mixed-length shifts 136
30. ROC Curves, all methods, estimated person parameters, mixed-length shifts 137
31. ROC Curves, all methods , bias-corrected person parameters, mixed shifts 137
32. ROC Curves, CMP/3PL, true person parameters, shift error length 3 138
33. ROC Curves, CMP/NRM, true person parameters, shift error length 3 138
34. ROC Curves, SCIP/3PL, true person parameters, shift error length 3 139
35. ROC Curves, SCIP/NRM, true person parameters, shift error length 3 139
36. ROC Curves, all methods, true person parameters = -1, shift error length 3 140
37. ROC Curves, all methods, true person parameters = 0, shift error length 3 140
38. ROC Curves, all methods , true person parameters = 1, shift error length 3 141
39. ROC Curves, CMP/3PL, true person parameters, shift error length 7 141
40. ROC Curves, CMP/NRM, true person parameters, shift error length 7 142
41. ROC Curves, SCIP/3PL, true person parameters, shift error length 7 142
42. ROC Curves, SCIP/NRM, true person parameters, shift error length 7 143
43. ROC Curves, all methods, true person parameters = -1, shift error length 7 143
44. ROC Curves, all methods, true person parameters = 0, shift error length 7 144
45. ROC Curves, all methods , true person parameters = 1, shift error length 7 144
46. ROC Curves, CMP/3PL, true person parameters, shift error length 10 145
47. ROC Curves, CMP/NRM, true person parameters, shift error length 10 145
48. ROC Curves, SCIP/3PL, true person parameters, shift error length 10 146
49. ROC Curves, SCIP/NRM, true person parameters, shift error length 10 146

Page

50. ROC Curves, all methods, true person parameters = -1, shift error length 10 147

51. ROC Curves, all methods, true person parameters = 0, shift error length 10 147

52. ROC Curves, all methods , true person parameters = 1, shift error length 10 148

53. ROC Curves, CMP/3PL, true person parameters, mixed-length shifts 148

54. ROC Curves, CMP/NRM, true person parameters, mixed-length shifts 149

55. ROC Curves, SCIP/3PL, true person parameters, mixed-length shifts 149

56. ROC Curves, SCIP/NRM, true person parameters, mixed-length shifts 150

57. ROC Curves, all methods, true person parameters = -1, mixed-length shifts 150

58. ROC Curves, all methods, true person parameters = 0, mixed-length shifts 151

59. ROC Curves, all methods , true person parameters = 1, mixed-length shifts 151

60. ROC Curves, CMP/3PL, estimated person parameters, shift length 3 152

61. ROC Curves, CMP/NRM, estimated person parameters, shift length 3 152

62. ROC Curves, SCIP/3PL, estimated person parameters, shift length 3 153

63. ROC Curves, SCIP/NRM, estimated person parameters, shift length 3 153

64. ROC Curves, all methods, estimated person parameters = -1, shift length 3 154

65. ROC Curves, all methods, estimated person parameters = 0, shift length 3 154

66. ROC Curves, all methods , estimated person parameters = 1, shift length 3 155

67. ROC Curves, CMP/3PL, estimated person parameters, shift length 7 155

68. ROC Curves, CMP/NRM, estimated person parameters, shift length 7 156

69. ROC Curves, SCIP/3PL, estimated person parameters, shift length 7 156

70. ROC Curves, SCIP/NRM, estimated person parameters, shift length 7 157

71. ROC Curves, all methods, estimated person parameters = -1, shift length 7 157

72. ROC Curves, all methods, estimated person parameters = 0, shift length 7 158

73. ROC Curves, all methods , estimated person parameters = 1, shift length 7 158

74. ROC Curves, CMP/3PL, estimated person parameters, shift length 10 159

Page

75. ROC Curves, CMP/NRM, estimated person parameters, shift length 10 159

76. ROC Curves, SCIP/3PL, estimated person parameters, shift length 10 160

77. ROC Curves, SCIP/NRM, estimated person parameter levels, shift length 10 160

78. ROC Curves, all methods, estimated person parameters = -1, shift length 10..... 161

79. ROC Curves, all methods, estimated person parameters = 0, shift length 10 161

80. ROC Curves, all methods , estimated person parameters = 1, shift length 10 162

81. ROC Curves, CMP/3PL, estimated person parameters, mixed shifts 162

82. ROC Curves, CMP/NRM, estimated person parameters, mixed shifts 163

83. ROC Curves, SCIP/3PL, estimated person parameter levels, mixed shifts 163

84. ROC Curves, SCIP/NRM, estimated person parameters, mixed shifts 164

85. ROC Curves, all methods, estimated person parameters = -1, mixed shifts 164

86. ROC Curves for all methods for estimated person parameters = 0, mixed shifts 165

87. ROC Curves for all methods for estimated person parameters = 1, mixed shifts 165

88. ROC Curves, CMP/3PL, bias-controlled person parameters, shift length 3..... 166

89. ROC Curves, CMP/NRM, bias-controlled person parameters, shift length 3 166

90. ROC Curves, SCIP/3PL, bias-controlled person parameters, shift length 3..... 167

91. ROC Curves, SCIP/NRM, bias-controlled person parameters, shift length 3 167

92. ROC Curves, all methods, bias-controlled parameters = -1, shift length 3 168

93. ROC Curves, all methods, bias-controlled person parameters = 0, shift length 3 168

94. ROC Curves, all methods , bias-controlled person parameters = 1, shift length 3 169

95. ROC Curves, CMP/3PL, bias-controlled person parameters, shift length 7..... 169

96. ROC Curves, CMP/NRM, bias-controlled person parameters, shift length 7 170

97. ROC Curves, SCIP/3PL, bias-controlled person parameters, shift length 7..... 170

98. ROC Curves, SCIP/NRM, bias-controlled person parameters, shift length 7 171

99. ROC Curves, all methods, bias-controlled parameters = -1, shift length 7 171

Page

100. ROC Curves, all methods, bias-controlled parameters = 0, shift length 7 172

101. ROC Curves, all methods, bias-controlled parameters = 1, shift length 7 172

102. ROC Curves, CMP/3PL, bias-controlled parameters, shift length 10 173

103. ROC Curves, CMP/NRM ,bias-controlled parameters, shift length 10 173

104. ROC Curves, SCIP/3PL, bias-controlled parameters, shift length 10 174

105. ROC Curves, SCIP/NRM, bias-controlled parameters, shift length 10 174

106. ROC Curves, all methods, bias-controlled parameters = -1, shift length 10..... 175

107. ROC Curves, all methods, bias-controlled parameters = 0, shift length 10..... 175

108. ROC Curves, all methods, bias-controlled parameters = 1, shift length 10..... 176

109. ROC Curves, CMP/3PL, bias-controlled person parameters, mixed shifts 176

110. ROC Curves, CMP/NRM, bias-controlled person parameters, mixed shifts..... 177

111. ROC Curves, SCIP/3PL, bias-controlled person parameters, mixed shifts 177

112. ROC Curves, SCIP/NRM, bias-controlled person parameters, mixed shifts..... 178

113. ROC Curves, all methods, bias-controlled parameters = -1, mixed shifts..... 178

114. ROC Curves, all methods, bias-controlled parameters = 0, mixed shifts 179

115. ROC Curves, all methods, bias-controlled parameters = 1, mixed shifts 179

116. ROC Curves using H^T for all shift error length scenarios..... 180

CHAPTER 1

INTRODUCTION

1.1 Background

On paper-and-pencil multiple-choice tests, the potential for examinees to mark their answers in incorrect locations presents a serious threat to the validity of test score interpretations. When an examinee skips one or more items (i.e., answers out of sequence) but fails to accurately reflect the size of that skip on their answer sheet, that can trigger a string of misaligned responses (Skiena & Sumazin, 2000a, 2000b, 2004). This phenomenon, referred to as a shift error, can result in correct answers being marked as incorrect and the examinee's score underestimating his or her true ability. While erasure analysis and answer-changing behavior studies (e.g., Matter, 1985, McMorris & Weideman 1986; Shatz & Best, 1987; van der Linden & Jeon, 2012) show that many shift errors are detected and corrected by the examinees, Skiena and Sumazin estimated that approximately 2% of paper-based tests have undetected shift errors. Despite movement toward computerized testing in recent years, paper and pencil multiple-choice tests are still pervasive in many high-stakes assessment settings, including K-12 testing (e.g., Massachusetts Comprehensive Assessment System; MCAS) and college entrance exams (e.g., SAT). Honest mismarking is not the only way that shift errors may occur, however. Another possible cause relates to cheating behavior. One examinee may look to another examinee's sheet for answers and, in the process, copy the answers to the wrong position on his or her own form, misaligning the copied responses by a position or two or perhaps even an entire column. In such cases, the threat to validity is not presented by the shift error itself, the stolen responses not being reflective of examinee ability and thus the more critical validity threat, but detection of such errors could provide a method for detection for this pattern of cheating behavior as well as a means for removing these invalid responses, thereby improving item parameter estimates. Methods for detecting copying behavior, with their reliance on pattern matching (Holland, 1996; Frary,

Tideman & Watts, 1977, Wollack, 1997), may also help inform the development of good methods for shift error detection.

Shift errors, regardless of cause, represent a form of aberrant item score pattern (Meijer & Sijtsma, 1995). Much research has been done investigating aberrant item score patterns, with methods for their detection and indices of appropriateness and person fit long- and well-established (Levine & Rubin, 1979; Drasgow & Guertler, 1987, Meijer & Sijtsma, 2001; Karabatsos, 2003). While these indices may be effective and reliable in detecting aberrant response patterns, "finding an aberrant pattern does not provide the explanation for this aberrance. The application of person-fit analysis techniques may easily lead to the detection of aberrant patterns, whereas the reasons for this aberrance is poorly understood." (Meijer & Sijtsma, 1995). Shift error analysis, instead of or in addition to person-fit analysis, provides a means for detection and understanding of this particular form of aberrant test behavior. Few methods for such analysis are offered in extant literature, though a series of studies by Skiena and Sumazin (2000a, 2000b, 2004) offer three such methods, two of which they found capable of detecting shifts with adequate selectivity but different degrees of sensitivity depending on the exam characteristics and lengths of the shifts present in the response vectors. While robust enough for careful use on real test data, their methods leave room for improvement. Based on the probabilities of aberrant patterns within full response sets, the methods either ignore item characteristics in calculating these probabilities or only factor in item difficulty as understood within a classical test theory framework, suggesting at least one avenue for possible improvement: application of item response theory models for calculating response probabilities, thereby incorporating item characteristics and examinee ability into the calculations.

The opportunity to accurately detect construct-irrelevant errors and either adjust scores to more accurately reflect examinee ability or flag examinees with inaccurate scores for removal from the dataset and retesting would improve the validity of important decisions based on test

scores, and could positively impact model fit by allowing for more accurate item parameter and ability estimation.

1.1.1 A Brief Explanation of Shift Error Detection Methods

Probabilistic shift error detection methods must consist of two general steps: 1) calculation of response pattern probabilities for detection of improbable response vectors, and 2) evaluation of a proposed alternative response vectors. Skiena and Sumazin (2000a, 2000b, 2004), employed these two steps, finding substrings with improbable patterns (i.e., having a disproportionate number of incorrect answers given the total number correct) then looking at improvements in fit based on shifting those substrings. In a preliminary study trying out the 3PL (Cook & Foster, 2012), the steps were reversed, response patterns corresponding to but misaligned with the answer key were first detected and the realigned pattern was proposed as the alternative vector, then probabilities were calculated that these patterns were not misaligned. With Skiena and Sumazin's methods, the probabilities of both the misalignment and the corrected substring play factors in identification of shift errors, whereas Cook and Foster's is based solely on the probability that the response string could have occurred in its place absent a shift error. Regardless of the order the steps, the result of the process is a list of shift errors that may be evaluated against previously determined acceptable thresholds of error.

1.1.2 Person Fit Methods vs. Shift Error Detection

Numerous methods are available for evaluating the fit of a person's performance on a test to the measurement model being used to score that test. These methods, referred to historically as appropriateness measurement (Levine & Drasgow, 1983), but more currently as person-fit methods (Meijer & Sijtsma, 1995), provide indices based on how well individuals' response patterns fit with expected patterns based on the given test model (Meijer & Sijtsma, 2001). Indices have been developed to fit CTT models (e.g., personal point-biserial and biserial, Donlon & Fischer, 1968), the Rasch model (e.g., M, Molenaar & Hoijtink, 1990), 2PL and 3PL models (e.g., l_z, Drasgow, Levine & Williams, 1985), and to CAT models (e.g., T statistics, van

Krimpen-Stoop & Meijer, 2000). Meijer & Sijtsma's 2001 meta-analysis compared 24 different person-fit statistics, each applicable to one or more measurement models. Karabatsos (2003) tested 36 person-fit statistics and found one (H^T, Sijtsma & Meijer, 1992) to greatly outperform the others. Drasgow, Levine, and Zickar (1996) proposed statistically optimal methods for detection of person-misfit in which probabilities of misfit are dependent on the specific type of misfit being looked for. Trabin and Weiss (1983) looked at person response curves, comparing expected to observed in order to detect certain types of misfit depending on specific differences between the two. None of these were determined to be effective in specifically identifying shift error though Drasgow, Levine, and Zickar's statistically optimal methodology offered a framework given an appropriate shift-error misfit model.

The shift error detection methods of Skiena and Sumazin (2000a, 2000b, 2004) and those proposed herein represent efforts to pinpoint this one specific type of person misfit, developing optimal indices that can be compared against error thresholds that produce acceptable levels of accuracy while maximizing the detection rate. Unlike a more general person-fit index, which can flag misfitting persons for exclusion from test analysis, detection of shift errors has the potential to provide more alternatives for dealing with the resultant misfit, including, given adequate confidence in the results, correction of the response string.

1.1.3 How Undetected Shift Errors Threaten Validity

According to the Standards for Educational and Psychological Testing, "Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests," going on to say that "… validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use" (AERA, APA, & NCME, 1999, p. 9). Kane (2006) suggests that this definition is reflective of the principles inherent in the construct-validity model. Unless the validity argument suggests that attention to detail or something like it provides important evidence in support of proposed test uses, answers arrived at correctly but entered incorrectly

introduce variance that is irrelevant to the construct and that undermines said validity argument. In the case of the would-be cheater, assuming him or her to be of low ability, the misalignment of copied answers may better reflect the examinee's ability than were they in the correct location, but that is not to say that the responses are an accurate reflection of that examinee's ability or that one could make accurate inferences based on the examinee's score. Further, if detection of the shift error leads to accurate detection of cheating behavior, removal of the examinee prior to final parameter estimation, scaling and setting of cut scores enhances the validity of all of those steps.

1.2 Statement of Problem

Methods for the general detection of aberrant response patterns are well-established to the point that the most recent studies on the topic are either meta-analyses of previously developed methods (Miejer & Sijtsma, 2001; Karabatsos, 2003) or have focused only on application to new testing formats, such as CAT (van Krimpen-Stoop & Meijer, 2000). While numerous person-fit methods are demonstrated to effectively identify aberrant test behavior, they offer little to nothing in pinpointing its nature. Sources of both spuriously high and spuriously low scores are discussed throughout the literature (e.g., Levine & Rubin, 1979, Meijer, 1996) but little research (Trabin & Weiss, 1983; Drasgow, Levine & Zickar, 1996) has been done on identification of specific types of aberrance. On achievement and aptitude tests, examinees' scores can be spuriously high by copying neighbors' answers or obtaining correct answers prior to the test somehow. Assuming these to be the cause ignores the possibility of the aberrance being due simply to a series of lucky guesses. On attitude scales, higher scores can be achieved simply by faking good. Spuriously low scores can have several root sources, including poor alignment of curriculum to test content, low motivation, unusual interpretation of items, and shift errors. The resultant misclassifications can, in the case of spuriously high scores, lead to unqualified candidates being undeservingly selected into jobs or academic programs, while spuriously low scores could result in qualified people being denied said opportunities.

Absent causal explanation, options for action after detection of an aberrant response pattern are limited. In the case of spuriously high scores, the choice between accepting the score, retesting, or outright penalization/disqualification of the examinee would depend on proof that the score was based on some form of cheating behavior. To that end, much research is currently being done on cheating detection methods and strategies. One such avenue, an exception to the dearth of type-specific person-fit research, is the development of indices for detection of copying on tests (e.g., K-Index, Holland, 1996; ω, Wollack, 1997). These look for agreement beyond chance between a suspected source and examinees who may have had the opportunity to copy from said source.

No one is accusing examinees with spuriously low scores of cheating, but appropriate action in the face of such scores remains complicated when the cause of the aberrance remains unknown. If the cause is due to inferior test-taking strategies, a failure to understand instructions, a different source of construct-irrelevant variance (e.g., language affecting scores on math word problems), or low motivation, the underlying problem may not go away simply by retesting (Drasgow & Guertner, 1987). In the case of shift error, however, retesting (with appropriate cautions) is likely to avoid a repeat of spuriously low performance. What's more, detection of shift errors with a high enough level of confidence could result in saving the time and expense of retesting altogether, since shift error detection has the capacity to determine not just the presence of such an error, but of the exact location and length of said error. In instances where shift errors are a byproduct of cheating behavior, shift error detection may provide another means toward exposing cheaters and taking appropriate action.

It might be tempting to minimize the present impact of undetected shift errors due to current and future inroads toward computer based testing, but that would be wrong. Firstly, educational testing and credentialing exams are taking on more importance than ever with educational reform, and much of this will continue to be done with paper and pencil, and mostly on the bubble sheets that can result in the shift errors this research attempts to address. Validity of

results is critical, and so every effort should be made to reduce the systematic threat to validity introduced with shift errors. Secondly, as some of the current educational testing is being done with younger children who are much more likely to mismark answer sheets, every effort must be made to reduce this threat to validity in the test scores of younger children. Thirdly, new statistical indices will be of interest to many testing agencies that may not currently be identifying these shift error problems. Finally, shift errors do not only affect those who make them; the impact extends to the identification of pass rates, and the estimation of item statistics.

1.3 Purpose of Study

Studies into answer-changing behavior on tests (e.g., Holland, 1996; Frary, Tideman & Watts, 1977; Wollack, 1997) attribute as much as 16% of answer-changing behavior to clerical errors detected and corrected mid-test by the examinee. Skiena and Sumazin (2000a, 2000b, 2004) claim that about two percent of paper and pencil tests contain undetected shift errors. Absent optimal methods for detecting shift errors, this remains a guess. Even if this is a severe overestimate, some simple math applied to a large-scale high-stakes testing program like the Massachusetts Comprehensive Assessment System (MCAS) can demonstrate the potential impact of shift errors. In 2012, 552,549 Massachusetts public school students were tested in 3 subjects each (Massachusetts Department of Education, 2012), meaning approximately 1.5 million tests were administered that year. If the two percent figure is accurate, that represents 30,000 examinees with spuriously low scores due to an undetected shift error. Cut the rate to one percent, that's still 15,000 examinees in one state in one year with test scores underrepresenting their ability due to shift errors. Whether or not that remains a gross overestimate, we can't know without optimal detection methods. Other paper-and-pencil tests with large stakes attached include the SAT, with 3 million examinees yearly, and the ACT, with 1.6 million examinees yearly, both of which are used by universities in making admissions decisions. As to the impact on the individual examinees, parameter estimates, cut scores and pass rates, that also cannot be determined unless the shift errors can be accurately detected. In short, part of the importance of

the problem is knowing exactly how important a problem it is. Making this determination has the beneficial side effect of simultaneously providing a solution.

The purpose of this study is to investigate new methods for shift error detection that employ IRT models, examining the methods for their selectivity (i.e., true positives vs. false positives), sensitivity (i.e., true shift errors detected vs. undetected), and robustness to parameter bias, all under a carefully manipulated, multifaceted simulation environment. This investigation should provide answers to the following questions, applicable across detection methods, bias reduction procedures, shift conditions, and ability levels, but stated generally as: 1) How sensitively and selectively can an IRT-based probabilistic model detect shift error across the full range of probabilities under specific conditions?, 2) How robust is each detection method to the parameter bias introduced by shift error?, 3) How well does the detection method detect shift errors compared to other, more general, indices of person-fit?, 4) What is the impact on bias of making proposed corrections to detected shift errors?, and 5) To what extent does shift error, as detected by the method, occur within an empirical data set?

What follows is a literature review of issues and sources of aberrant test response behavior and methods for its detection, previous research on shift error detection and other specific forms of aberrant test response behavior, issues related to the paper-and-pencil test format, and a breakdown of the measurement models underlying the probability calculations used for detecting shift errors within this study. Following that are an outline of the methods for a series of studies with a breakdown of all study conditions, including descriptions of probability models, detection algorithms, and employed person-fit statistics, a report of the results from performing these studies, and a discussion of the meaning of these findings and their impact on future measurement practice and research.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Overview of Literature Review

This chapter is a review of the literature pertinent to undetected shift error and methods for its detection. The chapter will be broken down by topic into the following sections:

2.2 Aberrant Test Response Behavior and Measures of Person Fit.

2.3 Issues of Validity and Differential Performance Dependent on Test Format.

2.4 Probability and Measurement Models.

2.2 Aberrant Test Response Behavior and Measures of Person Fit

2.2.1 Aberrant Response Behavior and its Impact on Validity

On a multiple-choice test, an examinee whose response pattern differs greatly from other examinees taking the same test could render that person's test score an inappropriate measure of ability. "Even with the best tests and testing procedures, at least a few anomalies are likely to occur in any very large test administration and … an effort should be made to identify the resulting defective test scores" (Levine & Rubin, 1979). To that end, a multitude of appropriateness measures, called person-fit statistics in modern parlance, have been developed. Using the term "appropriateness index", Levine and Rubin (1979) define a person-fit statistic as:

> … a measure of goodness of fit of a very general psychometric model to the individual examinee's item-by-item pattern of responses. An appropriateness index is expected to be high if the examinee's answer sheet is like that of similarly able examinees, and expected to be low if unlike similarly able examinees. Like the test score, the examinee's appropriateness score is solely a function of the examinee's item responses. Appropriateness indices thus summarize the internal evidence in the examinee's answer sheet indicating whether he or she approaches the test as do other examinees with the same ability (p. 271).

Paraphrasing Smith (1986), Petridou and Williams (2010) describe how person misfit threatens validity:

> when examinees take a test their responses are expected to conform to some standard of reasonableness: they are expected to generally answer the easier items correctly and answer the more difficult ones incorrectly. Patterns that

> contradict this expectation violate this standard of reasonableness and may be regarded as misfitting. Such response patterns, if they are unusual enough, signal what we call aberrant performances, which may indicate aberrant examinees. We suspect that such aberrance may imply the examinees are mismeasured, thus their score may be invalid" (p. 43).

Petridou and Williams (2010) describe at length how person misfit is assumed to threaten validity, including a table summarizing quotes about the relationship between person-fit and test score validity, reprinted here in Table 1. In particular, according to Wolfe and Smith's (2007) taxonomy of sources of evidence to support validity arguments, person misfit would provide evidence for substantive validity, defined as the extent to which theory explains differences in item response. Petridou and Williams (2010) investigated whether an aberrant response pattern was inherently indicative of mismeasurement and, thus, enough to render a test score invalid for its purpose. In a study of 674 individuals taking a 45-item mathematics test, two external sources of validity – interviews of 31 examinees and teacher assessments of approximately 400 examinees – were applied to determine whether unexpected item responses accurately reflected examinee knowledge. When either the interview or the teacher assessment agreed with the misfit statistic, they concluded this was evidence of mismeasurement and responses were changed. They concluded that aberrance as indicated by a fit statistic is often evidence of mismeasurement but that it does not automatically imply mismeasurement. As Wright's (1995) quote suggests, aberrance, whether truly indicative of mismeasurement or not, casts doubt on the estimate of a person's ability, and as such, person misfit should be seen as a threat to validity.

Hendrawan, Glas, and Meijer (2005) examined the effect of person-fit on a specific validity issue: classification decisions. With simulated samples of 400 and 1000 taking tests of 30 and 60 items under various conditions in which ten percent of the sample were either guessing on some easy items or gained access to the answer on some difficult items, the effect of these misfitting examines on theta estimates using MLE and EAP and resulting classification decisions were investigated. They found that the effect of the misfitting examinees on the classification of

the normal examinees was minimal, though precision in classifying aberrant examinees approached zero as the aberrance became more extreme.

2.2.2 Overview of Person-Fit Indices

The quantity of available person-fit statistics is large, but they are nicely summarized and compared in Meijer & Sijtsma (2001) and Karabatsos (2003). For most common measurement models, multiple person-fit statistics are available. Meijer and Sijtsma (2001) provide a methodological review looking at person-fit statistics based on Classical Test Theory, the Rasch Model, and the 2PL and 3PL models, as well as person-fit statistics designed specifically to address aberrant behavior in computer adaptive testing. Table 2 is a reprinting of their compiled list of person-fit statistics, demonstrating the abundance of available indices and breaking them down by underlying model.

Most group-based person-fit statistics are based on the Guttman (1944, 1950) model, in which it is assumed that a correct answer on a given item occurs at the exclusion of incorrect answers on any easier items. If items are arranged in order of difficulty, permitted patterns under the Guttman model are (1, 0), (0, 0), and (1, 1) where 1 is a correct response and 0 is an incorrect response. These are known as Guttman patterns. A response pattern of (0, 1) violates the Guttman model and is labeled an "error" or "inversion". Group-based statistics are all based on weighted counts of these inversion patterns and differ only in the weighting scheme applied. Criteria for selection of a group-based person-fit statistic are 1) low correlation with the number correct score and 2) detection rate (Harnisch & Linn, 1981, Rudner, 1983). An issue with group-based statistics is that the distribution of values for most of them and, in turn, the probability of classifying a score pattern as misfitting are dependent on the test score distribution (Meijer & Sijtsma, 2001).

For IRT models, be they one of the dichotomous or polytomous models, person-fit statistics take the form of residual-based statistics or likelihood-based statistics. Residual-based IRT person-fit statistics can be expressed generally as the sum of weighted differences between expected and actual outcomes on the items on a test. This form sets the expectation of a person-fit

statistic at 0. Squaring residuals before summation allows them to accumulate rather than risk them canceling each other out. Indices available based on mean-squared residuals include *U* and *W* (Wright & Masters, 1982), and *UB* and *UW* (Smith, 1985).

Likelihood-based person fit statistics are all based on the log-likelihood function

$$l_0 = \sum_{g=1}^{k} \{X_g \ln P_g(\theta) + (1 - X_g) \ln[1 - P_g(\theta)]\}$$

applied by Levine & Rubin for assessing person fit, in which k is the number of items, X is the response for item g (1 being a correct response, 0 being an incorrect one), and $P_g(\theta)$ is the probability of a correct response on item g for someone of ability level θ. Because l_0 is not standardized and has an unknown null distribution, several refinements were to follow, most notably l_z (Drasgow, Levine & Williams, 1985), which standardized l_0 by simply subtracting its expected value and dividing the result by the standard deviation, and for the Rasch model, *M* (Molenaar & Hoijtink, 1990), a component of l_0 both easy to approximate and known to follow the same order as l_0 given an item-score pattern $\boldsymbol{X}$. Group-based (i.e., CTT-based) fit statistics can also be used to detect misfitting patterns under an IRT model but may yield different results than those based on IRT parameters (Meijer & Sijtsma, 2001).

One type of person fit index that has both CTT- and IRT-based incarnations is the caution index. The caution index (Harnisch & Linn, 1981) is expressed as:

$$C_i = 1 - \frac{Cov(\mathbf{X}_i, \mathbf{n})}{Cov(\mathbf{X}_i^*, \mathbf{n})}$$

where $\mathbf{X}_i$ is the item score vector for examinee *i*, $\mathbf{X}_i^*$ is the expected Guttman vector given the number correct for examinee *i*, and **n** is a vector of correct responses per item across all examinees. All caution index-based fit indices weigh item score vectors in terms of their agreement with a Guttman pattern in this way. For IRT-based caution indices, the expected Guttman vector is replaced by the vector $\mathbf{P}(\theta)$ for a given ability level θ and IRT model, providing a more nuanced but preserved version of the Guttman pattern (Meijer & Sijtsma, 2001).

Karabatsos (2003), in a comparison of 36 person-fit indices across conditions including 1) test length: short (17 item), medium (33 item), and long (65 item), 2) proportion of respondents that are aberrant: .05 to .5, and 3) respondent types: cheaters, creative responders, lucky guessers, careless responders and random responders, found that one fit statistic performed better than all of the others: H^T (Sijtsma & Meijer, 1992). H^T works similarly to a caution index in that it looks at covariance of an examinee response in relation to other examinee responses, but instead of comparing it to a vector of number correct, it sums the covariances of an individual respondent and each other respondent individually, then does the same for the maximum possible covariance in each case and takes the ratio:

$$H_i^T = \frac{\sum_{j \neq i} \sigma_{ij}}{\sum_{j \neq i} \sigma_{ij}^{\max}}$$

H^T has a couple theoretical disadvantages that could lead one to choose a different measure of person fit. Firstly, it is not standardized to the Guttman pattern such that perfect fit would be given by $H_i^T = 1$ and perfect misfit is given by $H_i^T = 0$. As such, even when an item-score pattern fits perfectly to the Guttman pattern, H_i^T may still be less than 1 (Sijtsma & Meijer, 2001). Additionally, though we are working with IRT models, H^T is a CTT-based statistic. It may be theoretically desirable to choose a statistic consistent with the measurement model actually being used. The advantages may outweigh the disadvantages, the primary advantage being that H^T simply works the best. The effectiveness of all IRT-based person fit statistics are, to some extent, undermined by their dependence on $\hat{\theta}$, the person parameter estimate, rather than θ, the true person score. Using estimates rather than true scores can shrink the variance of IRT-based person fit statistics such that the distribution is no longer standard normal, with an empirical Type I error rate smaller than the nominal Type I error rate. Under these circumstances, it becomes more difficult to differentiate between true and false positives

(Molenaar & Hoijtink, 1990; Meijer & Sijtsma, 2001). In terms of relationship to the Guttman pattern, since the effectiveness of the statistic is based on finding a maximum ratio between true and false positives and not on this interesting theoretical feature of some person fit statistics, it is empirically irrelevant to their purpose.

2.2.3 Specific Forms of Aberrant Response Behavior and Methods for their Detection

Several forms of aberrant response behavior are defined in the literature. Levine and Rubin (1979) note that aberrant response behavior takes the general form of a disproportionate number of easy items answered incorrectly or difficult items answered correctly. They cite four specific forms of aberrant behavior: 1) improperly obtaining items, 2) shift errors, 3) creative responding, and 4) suboptimal test-taking strategies. Meijer (1996) expands upon, refines, and defines these forms into six categories of aberrant behavior. It should be noted that these proposed forms of aberrance have not necessarily been detected in empirical scenarios:

1) Sleeping behavior: The examinee is slow to adapt to the task of taking the exam and, after adjusting, does not check answers to some of the easier items. This results in proportion correct on easy items being smaller than expected given proportion correct on the items of medium to high difficulty.

2) Guessing behavior: The examinee guesses blindly at items that are beyond that individual's ability. Proportion correct for items of a difficulty at or below that of the examinee's ability is high while proportion correct for more difficult items is approximately equal to one over the number of response options.

3) Cheating behavior: A person of low ability is likely to do well on the easy items but may, after struggling on some more difficult items, copy answers for the most difficult items. As such, cheating behavior may appear as a high proportion of easy and difficult items answered correctly while items of medium difficulty are answered incorrectly.

4) Plodding behavior: Based on the assumption that response behavior is probabilistic, a response vector that perfectly follows a Guttman pattern may be considered "too good to be true."

This is a pattern that may be generated by a methodical worker taking the time to answer each item correctly before moving onto the next.

5) Alignment error: When an examinee reaches an item of high difficulty, the examinee may move on to the next item, intending to go back to the difficult item only after completing the easier items. If this person fails to skip a position on the answer sheet, this can result in an alignment or shift error of several items as the examinee fills in a string of responses intended to be in a given position but filled in in an incorrect location. Meijer (1996) conceptualizes this as several more difficult answers being scored incorrectly, though this seems a simplification of the nature of a shift error. While a shift error may be more likely to begin on a more difficult item, the rest of the misaligned string could take place on items of any difficulty.

6) Extremely creative behavior: A high ability examinee may look at some easy items as too simple and apply a creative reinterpretation to those but not the better fitting medium and difficult items. As a result, proportion correct on the easier items will be lower than expected compared to the medium and high difficulty items.

7) Subability deficiency: On a test made up of more than one subdomain, if an otherwise high ability examinee is weak in one particular subdomain, the examinee will perform disproportionately poorly on all items within that subdomain regardless of their difficulty level.

While different forms of aberrance have been defined, identification of specific forms of aberrance remains difficult. Some behaviors, while very different, produce similar patterns of aberrance. Some aberrant patterns defy easy classification. Other aberrant behaviors produce response vectors that are closely aligned with the Guttman pattern and thus would not be flagged by a person-fit index (Meijer, 1996). Meijer and Sijtsma (2001) further note that while person-fit indices provide indications that item-score patterns are misfitting, most “do not allow the recovery of the mechanism that created the deviant item-score patterns” (p. 108). Karabatsos (2003) found that some forms of aberrance are easier to detect than others depending on the person fit index that has been selected but, whereas a particular form of aberrance may be more

likely to be detected, the indices offer no indication as to what form was detected beyond its being spuriously high or spuriously low. Meijer and Sijtsma (1995) note:

> "Finding an aberrant pattern does not provide the explanation for this aberrance. The application of person-fit analysis techniques may easily lead to the detection of aberrant patterns, whereas the reasons for this aberrance are poorly understood. Therefore, a full person-fit analysis requires additional research into the motives, the strategies, and the background of those examinees that deviate from the statistical norm set by the model or the group." (p. 270)

Hulin, Drasgow, and Parsons (1983) further note that when underlying causes of aberrance are unknown, little meaning can be given to the aberrant individual's test score. Levine and Rubin (1979) acknowledged this stating that person-fit indices "will be most useful as broad, low-power screening devices for identifying a proportion of examinees requiring specific more powerful procedures" (p. 271).

A couple of methods, however, do allow for identification of some forms of aberrance. Trabin and Weiss (1983) developed a method for identifying aberrant behavior by comparing expected and observed person response functions. A person response function is similar to a test response function except, rather than presenting probable test performance across the person parameter scale, it presents (expected or observed) performance at different difficulty levels for an individual of a given ability. Employing an IRT model, by stratifying items on an exam according to location for a given ability level, an expected proportion correct can be calculated for each of those strata, the result being a curve that represents expected performance along the range of difficulties for that ability level. When an individual's observed performance curve is compared to this expected curve, when the relationship between the two takes certain forms, a certain type of aberrance may be inferred. If, for instance, a person's observed response function is significantly lower than expected on the easier items, that may represent a form of careless behavior: sleeping or creative responding. A person whose curve has a very steep slope at their ability level is showing alignment to the Guttman pattern beyond what might be expected given a test's probabilistic nature, indicating plodding behavior. A pattern that levels off at a certain

difficulty level such that it is approximately one over the number of response options is exhibiting guessing behavior. Sijtsma (1998; Meijer & Sijtsma, 2001) developed this into a person-fit index, $D(\theta)$ by adding the average differences at each strata together, a method which Karabatsos (2003) found to be the most effective of the IRT- based approaches, though by itself it would potentially hide the form of aberrance as one can envision a scenario in which a person who is perfectly misfitting could have those average differences add up to zero, hiding that misfit (Armstrong & Shi, 2009). An approach in which differences between observed and expected are squared (Nering & Meijer, 1998) may serve as a better index. This chi-square statistic for $D(\theta)$ may be calculated as follows:

$$\chi^2 = \sum_{g=1}^{G} \frac{\left\{ \left(\sum_{k \in A_g} X_k \right) - \left(\sum_{k \in A_g} P_k(\hat{\theta}) \right) \right\}^2}{\left(\sum_{k \in A_g} P_k(\hat{\theta}) \right)}$$

where $\mathbf{A}_g$ is the vector of stratified subsets, G being the number of subsets. $\mathbf{X}_k$ is the response vector while $P_k(\hat{\theta})$ is the probability of getting a correct response on item *k*.

Drasgow and Levine (1986; Drasgow, Levine, & Zickar, 1996) proposed a statistically optimal method for detecting aberrant response vectors. These are methods for detecting specific forms of aberrance because the statistical optimization requires a model specific to the form of aberrance being detected. Within this method, given statistically optimal models for both normal and aberrant responses, the likelihood of each response pattern under those models is calculated and the likelihood ratio of each pattern is calculated. Once this is done, the minimum likelihood ratio at which the acceptable Type I error rate is achieved provides the statistically optimal point at which to base your decisions to accept or reject the null hypothesis that a given response vector is not aberrant. The key to the method, then, is selecting appropriate models for calculating likelihoods that the vector is fitting and that it is misfitting. For calculating the likelihood that it is

fitting, within an IRT framework, this is a matter of applying the appropriate IRT model given the data and what number of parameters are appropriate to the item characteristics. For calculating the likelihood of an aberrant pattern, the specific form of hypothesized aberrance will indicate the appropriate calculation. They model several forms of aberrant behavior: cheating on both known and unknown item sets, dissimulation ("faking good") on personality scales and biographical inventories, format unfamiliarity (contextualized as unfamiliarity with computers, but their description seems as though it would hold for any unfamiliar context), and the Levine-Rubin Spuriously Low Model (1979).

The Levine-Rubin model fits for any of several contexts that result in spuriously low performance on some portion of a response vector, including shift error (Drasgow, Levine, & Zickar, 1996). According to this model, in a test of *n* items, *m* are answered with one of *J* possible response options selected at random. The remaining *n minus m* items have responses based on the normal (i.e., non-aberrant) model. The probability for response vector **u** under the normal model is given by the formula:

$$\mathrm{P}(\mathbf{u}) = \prod_{i \in \mathbf{u}} P_i^{\mathbf{u}_i} (1 - P_i)^{1-\mathbf{u}_i}$$

in which P_i is the probability of a correct response on item *i* and $\mathbf{u}_i$ is the response to item *i* in response vector **u (**1 for a correct response, 0 for an incorrect response), whereas for the response vector **u*** under the aberrant condition given a subset S_k subject to the aberrant condition and the remaining items subject to the normal condition, the probability is given by the formula:

$$\mathrm{P}(\mathbf{u}^* \mid S_k) = \prod_{i \in S_k} \frac{1}{J}^{\mathbf{u}_i^*} \left(\frac{J-1}{J} \right)^{1-\mathbf{u}_i^*} * \prod_{i \notin S_k} P_i^{\mathbf{u}_i} (1 - P_i)^{1-\mathbf{u}_i}$$

in which all symbols are as for the previous formula and *J*, as noted above, is the number of possible response options.

A pair of studies (Drasgow & Levine, 1986; Drasgow, Levine & McLaughlin, 1987) found misfit detection rates under this model given an 85 item test with 9 misfitting items to be

.08, .13, and .17 at alphas of .01, .03, and .05 while increasing the number of misfitting items to 13 and looking only at the top 8% of examinees (under the premise that high-ability examinees are most severely affected by spuriously low responding) obtained detection rates of .81, .86, and .88. Increasing number of misfitting items to 26 while continuing to look only at the top 8% improved detection rates to 0.97, 0.98, and 0.99. Drasgow et al. (1996) are careful to note that optimal indices are only truly statistically optimal if the underlying aberrance model is correctly modeling the underlying aberrance, a condition easy to achieve under simulation but impossible to know when using empirical data.

Armstrong and Shi (2009) propose a method that, while not designed specifically to allow identification of the causes of aberrance, builds the person-fit statistic in a cumulative fashion such that it can identify the location of the aberrance within an item vector. While Miejer (1996) defines different types of misfit in terms of how such misfit might be reflected in spurious highness or lowness, Armstrong and Shi believe confining aberrance to consecutive items is a more reasonable assumption when looking at causes such as fatigue, distraction, cheating, or special knowledge. Shift error, though not specifically mentioned by Armstrong and Shi, is a form of misfit that will occur over consecutive items and may be due to fatigue or distraction. The statistic they propose, a parametric cumulative sum statistic (CUSUM), sums differences between expected and observed responses one item at a time so as to detect runs of either positive or negative deviations that could balance each other out and remain undetected if one calculation is made on the entire response vector. The authors note that this cumulative procedure can be applied with any person-fit statistic, including Drasgow and Levine's (1986) optimal detection method. Thus, it is possible that both location and type may be ascertainable through a hybridization of the CUSUM method and a type-specific model for misfit detection. In a simulation study comparing detection performance of CUSUM to several person-fit statistics under conditions with 100 and 300 aberrant examinees out of 10,000, aberrances of 0, 8, 10, and

12 introduced, and alpha levels of .01 and .05 examined, found the CUSUM to detect the aberrant examinees at rates 3 to 4 times that of the non-cumulative person fit statistics.

In evaluating person-fit statistics, while a couple of the studies (e.g., Armstrong & Chi, 2009; Hendrawan, Glas & Meijer, 2005) examine the effectiveness only at specific alpha levels, best practice, as defined by Levine & Rubin (1979) and used to great effect by Karabatsos (2003), is to employ receiver operating characteristic (ROC) curves (Green & Swets, 1966). Rather than presenting the detection rate at only a couple of alpha levels, the ROC curve shows the tradeoff between sensitivity and selectivity across the full type I error spectrum from 0 to 1. An example is shown in Figure 1. As the threshold of probability at which the detection method represents an identified error is made more permissive, sensitivity increases and can be seen by the curve increasing in the vertical dimension. This increased sensitivity can come at a cost in terms of selectivity, a tradeoff represented by increase along the horizontal dimension. A perfectly selective method with no type I error would hug the y-axis perfectly whereas a method incapable of detecting anything but false positives would hug the x-axis perfectly. By employing an ROC curve, one can determine the best type I error rate for any detection method in terms of what is an acceptable compromise between sensitivity to true positives and lack of specificity in eliminating false positives. While a typical ROC curve uses the characteristics of sensitivity (i.e., true positive rate) and 1 – selectivity (i.e., false positive rate). The example uses true positive rate but shows a different characteristic for presenting false positives: the false discovery rate (Benjamini & Hochberg, 1995). Whereas a false positive rate gives the ratio of negatives falsely identified as positive to total negatives, false discovery rate gives the number of negatives falsely identified as positives to total identified positives. Whereas a nominally low false positive rate of .01 or .05 can still bury what seems like a good true positive rate under a mountain of false positives, a false discovery rate of the same guarantees that true positives outnumber false positives at exactly the rate specified, a more meaningful and useful finding when negatives greatly outnumber positives in the data.

2.2.3.1 Detection of Cheating Behavior

Of all specific sources of person misfit, perhaps the most attention has been paid to cheating behavior. Whereas general aberrance detection generally features methods highlighting patterns of unusual disagreement with the Guttman scale, cheating detection historically has focused on unusual agreement between two examinees, particularly in their wrong answers. Whereas agreement in correct answers may simply indicate that both examinees are of similar ability, when incorrect answers agree to an unusual extent, this may indicate copying behavior (Angoff, 1974; Holland, 1996; Belov, 2011). One may consider shift error in an analogous way: as response vectors in unusual agreement with a misaligned portion of the answer key. As such, the indices developed to detect unusual agreement as indicators of cheating behavior may inform development of indices to detect unusual agreement as indicators of shift error.

Angoff (1974) conducted the earliest research into this idea, comparing examinees on the SAT, comparing examinees known to be in different geographical locations (making copying impossible) to create a norm group and then in the same geographical locations to see if there was any behavior that differed significantly from these norms. Comparisons were made on 12 variables, which were then used to produce eight indices "of copying" through bivariate comparisons. By controlling for independent variables, such as number answered incorrectly by the compared respondents, value on a dependent variable, such as number answered incorrectly in the same way by both respondents, could be assessed in terms of deviations from the mean in the normed group and, if significantly different from that mean, could be flagged as exhibiting copying behavior. When calculated on a sample of 50 examinee pairs known to have exhibited copying behavior, calculated t-values were at least 3.0 for at least one of the indices. An index measuring consecutive shared omits and incorrect responses while controlling for total omits and incorrect responses identified 41 of 50 known copying cases at that threshold. Two of the eight indices, including this last one, were put into operational use by Educational Testing Service.

Holland (1996) employed the K-Index to assess the degree of unusual agreement between incorrect multiple-choice answers in two examinees. Holland derives the probability that two examinees will have some number of matching incorrect responses and determines that the distribution of a binomial variable serves as a reasonable approximation of that derived probability. Since unusual agreement is defined by number of incorrect responses being at or above a certain level, the formula appropriate for stating this is:

$$P(B \geq m) = \sum_{a=m}^{n} \binom{n}{a} p^{a} (1-p)^{n-a}$$

where B is a binomial random variable, m is the count of items in agreement, n is the number of incorrect items that the source examinee got incorrect, and p is the probability that an incorrect responses in the comparison group that is in agreement with the incorrect response of the source. The K-index is $P(B \geq m)$ where p is defined by making comparison groups of examinees with the same number of items wrong and calculating the proportion whose wrong answers agree with the source's. Note that this method's rooting in proportions makes it analogous to CTT.

Frary, Tideman & Watts (1977) has a CTT-based method considered to have a clear theoretical and statistical advantage over K-index and other CTT methods (Wollack, 1997). Their method compares the number of identically answered items for a pair of examinees to the expected number of identically answered items. The g_2 index simply takes the difference between these two values and divides it by the standard deviation of the difference, giving it an approximately standard normal distribution. By treating one of the examinees in the pair as a source and the other as a copier, one can consider the source's answers to be fixed and need only come up with reasonable probability estimates that the copier would select S's answers under non-copying conditions. When correct responses agree, this would simply be the CTT p-value for each item:

$p_c = e_c/e_t$

where e_c is the number of examinees getting the item correct and e_t is the total number of examinees. For incorrect responses, the probability would be:

$$p_w = (1-p_c)/(n-1)$$

in which p_w is the probability of an incorrect response, p_c is the probability of a correct response, as given above, and n is the number of response options. Summing these probabilities for all items would give you an expected agreement rate. Summing $p_c(1-p_c)$ for all items would give the standard deviation. Unlike the K-index, this method is not dependent only on incorrect answers, instead including correct and incorrect answers in its model.

Wollack (1997) expands upon g_2 by incorporating the nominal response model (NRM; Bock, 1972) rather than relying on proportion correct to generate response probabilities. NRM is discussed in section 2.6. As with g_2, Wollack's ω will be approximately standard normal, so long as the distribution of observed agreements between sources and copiers is also normal, a condition that will be met according to the central limit theorem so long as the number of items is sufficiently large.

Several other indices (e.g., Belov, 2011; van der Linden & Sotaridona, 2004; Sotaridona & Meijer, 2002) are available for detection of copying behavior, all based on a comparison of observed and expected levels of agreement, either of all responses or correct responses. Those reported herein are the ones most influential in how shift error detection methods were developed in this paper.

2.2.3.2 Detection of Shift Errors.

Though alluded to as a source of person misfit (Levine & Rubin, 1979; Meijer & Sijtsma, 1995; Meijer, 1996; Dodeen & Darabi, 2009) and while some of the methods may be suitable for their detection, shift error has received little direct attention in the psychometric literature in terms of specific methods for its detection. Optimal statistical detection (Drasgow & Levine, 1986) as a concept is applicable given an appropriate misfit model. CUSUM's focus on sequential items (Armstrong & Chi, 2009) is in concert with the sequential nature of shift errors. Beyond

acknowledging, defining, and perhaps classifying shift error as a subset of a more general category of person misfit, the psychometric literature offers little in terms of directly addressing shift error detection. Even if a method is especially adept at detecting shift errors, it may not be capable of classifying them as such so that appropriate action may be taken.

Skiena and Sumazin (2000a, 2000b, 2004) lay out the challenges in detecting shift errors, noting that it is inadequate to simply identify answer sequences whose scores increase when shifted, as it would unfairly reward random guessers and other poor performers who arrived at such a sequence purely by chance. Even hunting for large blocks of incorrect answers that become correct when shifted is not entirely appropriate since shifted blocks could still have correct answers by chance and long strings of entirely correct answers may be as unlikely as long strings of incorrect ones. Factors they note as essential for consideration in detecting shift errors are 1) examinee performance, 2) exam difficulty, 3) change in number of correct/incorrect answers, 4) extent to which scoring method encourages or discourages random guessing, and 5) answer key pattern. They approach the issue of shift errors head-on as a computer science problem Skiena and Sumazin (2000a, 2000b, 2004). With a goal of discriminating between exams that contain shifted responses and those that do not, they approach the problem algorithmically, developing three approaches for shift detection. Each scans individual exams for patches representing potential shifts, differing in what information should be considered in scoring the exams. Their three models, in order of increasing complexity, are: 1) the dynamic programming model, 2) the single scan model, and 3) the double scan model. A breakdown of each follows.

The dynamic programming model performs a string alignment between the answer key and the exam, calculating edit distance of optimal alignment of the two strings. Edit distance is a count of the number of operations required to align two strings (Gusfield, 1997). In the context of correcting for shift errors, when few edits result in better alignment of a response string to an answer key string, this is evidence that a shift error may have occurred. Assuming at most one shift error per response set, the dynamic programming model looks for the misalignment with the

greatest improvement at the lowest cost, improvement determined by number of items contained in the shift, cost determined by how many places the substring has been displaced and thus must be moved in order to achieve optimal alignment. A threshold is set by weighting the benefit vs. cost in a way that keeps false detections to an acceptable minimum while allowing as many true positives to be found as possible.

The single scan model analyzes each substring within an exam probabilistically in order to determine its likelihood given the examinee's overall score. Given a particular raw score, the method calculates a probability of a substring of size n having a given number of wrong answers. Patches of likelihood below predetermined probability thresholds based on the raw score are flagged as suspicious so that they may be tested for improvement (in likelihood) after a corrective shift.

The double scan model is an extension of the single scan model that calculates the substring probabilities via a more complex algorithm, one in which probability distributions are determined for all items based on all examinees. Such a method is more capable of identifying easier patches of items and thus can treat such areas with more suspicion than when it assumes all items to be of the same difficulty, as with the single scan model.

Whereas the dynamic programming model need only find the shift with the greatest benefit to cost ratio and determine if it meets the predefined threshold that minimizes false positives, the single and double scan models involve two steps, one in which suspicious patches are identified as per the model descriptions above and one in which those patches are scored in order to be classified as shift errors or not. Suspicious substring identification involves a series of probability calculations of the form $P(N,n,k,m)$ in which *N* is the total number of items in the exam, *n* is a substring length, *k* is the number of wrong answers in substring *n*, and *m* is the examinee's raw score on the exam. The probability being calculated is the probability of getting *k* wrong answers in a substring of length *n* given a raw score of *m* on an exam of length *N*. When

interested in substrings for many different *n*'s, this involves a large number of operations that could prove prohibitive. Probability of a correct answer P_c = number correct *m* over total number of items *N*, the probability of getting *k* incorrect in a substring of length is expressed by the formula:

$$n = P_c^{n-k} (1-P_c)^k.$$

The double scan model employs the same method except it adds weights for item difficulty at the group level in its calculation for P_c. Because this method yielded no practical benefit within their study, specifics of the weighting method are not further discussed here.

Because detection alone is not enough to discriminate between true and false shifts, Skiena and Sumazin further analyzed the suspicious substrings by determining significance of the score increase caused by correcting the shift. Models for calculating the significance are based on the null hypothesis that the suspicious substring is not a shift error and, thusly, the result of shifting it should relate to the answer key in a random fashion. This being the case, the probability of the shifted substring occurring by chance can be calculated. If that probability is below the specified alpha level, the null hypothesis may be rejected and the substring may be considered a shift error. Skiena and Sumazin test this hypothesis via two models, one independent of answer key and one dependent on it. The independent model calculates the probability that a block of length *N* with α response options will yield *B* correct answers by the following:

$$P(N,B,\alpha) = \sum_{k=B}^{N} \binom{N}{k} \left(\frac{1}{\alpha}\right)^B \left(\frac{\alpha-1}{\alpha}\right)^{N-B}$$

The dependent model considers the structure of the answer key and number of response options to factor in increases and decreases in probability due to repetition in the key.

In testing their methods, Skiena and Sumazin (2000a, 2000b, 2004) name but do not describe three detection levels: Permissive, Proper, and Restrictive. Results in Skiena and Sumazin (2000a) are reported based on the permissive detection level. In their studies, they found that the permissive level excessively rewarded poor performers and might be adequate for their

university exams but not the SAT. They reported few false detections at the proper detection level while identifying 902 examinees with shift errors on SAT forms. Using a restrictive method reduced this finding to 159 with nearly no false detections but an inadequately low detection probability. Exact false detection rates are not reported.

In Skiena and Sumazin (2000a), they looked at five exams from university courses, each with different difficulty levels (p=.62 to .81), sample sizes (N=66 to 204), numbers of response options (4 or 5), and number of items (30, 33, or 50). Into the response sets for these tests, they simulated shift errors of lengths 3 to 10 then ran their three detection methods on those sets plus the unshifted response sets. Under permissive conditions, they found that the dynamic programming method performed significantly worse than the single and double scan methods, while the double scan method did not perform sufficiently better than the single scan method to warrant its extra algorithmic complexity and processing time. Results for their single scan method are reprinted in Table 3. When no shifts are introduced, shifts are still detected within the data (representing either false positives or pre-existing real shift errors) at a rate between .000 and .019. With shifts of length 3 introduced, they are detected at rates between .127. As shifts of increasing lengths are introduced, detection rates steadily increase for all exams such that shifts of length 10 are detected at rates between .750 and .943.

Skiena and Sumazin (2000b) replicates the previous study but uses Scholastic Amplitude [sic] Test (SAT) data. In this study, they conducted experiments to answer two questions: 1) how well do their methods detect shifted exams and 2) how often do actual exams contain uncorrected shift errors. Within this study, using the same methods as for the previous, they report results at all three detection levels. The simulation study demonstrates the same trends, with detection rates increasing as length increases. In the empirical data, containing approximately 1830 shift errors, at the restrictive level 159 shift errors were detected, 902 were found at the proper level, and 3611 were found at the permissive level, demonstrating clearly that the permissive level is overly permissive while the proper and restrictive methods are sacrificing sensitivity in favor of

selectivity. A further breakdown of shift errors by race and income showed no significant differences based on either of those criteria.

Cook and Foster (2012) tested two simple algorithms for shift error detection, one employing the 3PL model and one under which all responses are assumed equally probable. Both followed the same procedure, only differing in method of calculating probability. Each begins by detecting sequences of responses that coincide exactly with the answer key except for a misalignment either one position forward or one position backward, then calculate the probability that the response sequence is in the correct location given the probabilistic model. Probabilities are compared against a threshold set to minimize false positives to an acceptable Type I error rate. Methods were tested in simulation under nine shift error conditions. The first eight simulated shifts of one length, between three and ten, introduced to 100% of examinees while the ninth introduced shifts of all lengths from three to ten within the same group, each length introduced to 1% of examinees. For each scenario, 100 replications of 2000 examinees were simulated based on a normal ability distribution and using 45 items, parameters taken from an actual state K-12 assessment and responses simulated using the nominal response model (Bock, 1972). Application of the detection algorithms yielded 100% selectivity rates with varying degrees of sensitivity, dependent on shift condition for all conditions under the 3PL while not reaching 100% selectivity and always performing inferiorly under the equal probability model. Table 4, reproduced from Cook and Foster (2012), shows those results. Results were encouraging given that the method did not incorporate incorrect answers into the algorithm, potentially disguising shift errors that had one or more wrong answers contained within the shift.

2.2.4 Addressing Aberrant Test Response Behavior

While few studies have been conducted investigating the usefulness of applying person-fit statistics to empirical data (Meijer & Sijtsma, 2001), Smith (1985) recommends four actions that may be taken when a response vector is determined to be misfitting: 1) report multiple ability estimates for an examinee rather than just one, 2) correct the item-score pattern so

that it better fits the model (i.e., assume the misfit represents a correctable error), 3) discard the test results and retest the examinee, or 4) conclude that the error resulting from the misfit is small enough that the ability estimate is accurate enough without correction. Which action is most appropriate is highly dependent on context.

2.3 Issues of Validity and Differential Performance Dependent on Test Format

Shift error is a problem most specific to paper and pencil tests in which answers are given on a separate sheet rather than within the question booklet. As such, understanding the extent to which test format can influence performance helps to contextualize the problem of shift errors. A great deal of research has been conducted regarding the influence of test format on validity and differential performance. As computer-based testing has increased in prevalence, research on the topic has focused on determining whether paper-and-pencil tests and computer-based formats are equivalent, with a distinct emphasis on whether conversion to computer-based testing will be so different as to call into question whether the tests are measuring the same things (e.g., Mead and Drasgow, 1993; Pomplun & Custer, 2005; Kingston, 2009). The need for measurement equivalence is raised in the *Guidelines for Computer-Based Tests and Interpretations* (APA, 1986), the *Standards for Educational and Psychological Testing* (APA, AERA, NCME, 1999) and in guidelines set forth by the International Testing Commission (2005), all promoting the position that "delivery mode should not affect examinee performance on any measure, and evidence of measurement equivalence or non-equivalence should be reported" (Rowan, 2010, p. 3). Kingston (2009) highlights the importance of this point, stating:

> "Often making a traditional test available on computer is seen as a first step on a path that will lead to future larger improvements. However, because of technology access issues (for example, not all schools have sufficient computers or the necessary Internet bandwidth), concern over equity, or general political issues, many testing programs find it necessary to offer their constituencies (districts, school, or individuals) choice. Thus it becomes imperative to demonstrate the comparability of scores from computer and paper administrations." (p. 23)

Russell, Goldberg, and O'Connor (2003) found that computer familiarity plays a large role in differential performance on tests offered in both paper-and-pencil and computer-based formats, with those familiar with computers scoring higher on the computer-based version and those unfamiliar with computers performing better on the paper version. If format proficiency is not part of the construct of interest but is affecting test performance, this is a threat to validity that can render comparison of scores between modes of administration impossible (Rowan, 2010). Huff and Sireci (2001) note that introducing tutorials has helped to combat construct irrelevant variance caused by lack of computer familiarity but it is unclear from the literature what is being done to reduce error introduced exclusively by the paper format. In a 1993 meta-analysis, Mead and Drasgow found that the largest differences between these two modes of administration were for speeded tests, a phenomenon they attributed in part to difference in response format, particularly pressing a key vs. marking a bubble sheet. This study solely analyzed research on adult populations, though a study using cognitive test scores for students in grades 4 to 12 (Ito & Sykes, 2004) reached a similar conclusion. Pomplun, Frey, and Becker (2002) found, in a test of score equivalence between paper and computer versions of a speeded reading test, found higher scores on the computer version as well as more students completing the computer version. A format that is more likely to result in speededness and incompleteness logically allows for less time to check answers and correct errors.

Format differences, however, predate computer-based testing. With the advent of machine scoring of tests, concerns arose as to how separating the answer sheet would affect reliability and validity of test results (Dunlap, 1940). Dunlap cites one source of error that could arise from separation of the answer sheet: failure to record the answer in the intended place, stating, "it often happens that he omits an item and, as a result, vertically displaces all remaining answers" (p. 5). In a series of experiments in which fourth graders were asked to complete tests using answer sheets with different numbering formats and to underline answers within the booklets as well as fill out the answer sheets, Dunlap found that approximately two-thirds of the

exams had discrepancies between booklet underlining and answer sheet marking with between 2.5% and 16.0% (depending on conditions) of the discrepancies being of a type that would be consistent with a shift error.

Several studies have investigated potential impact of using optically scored answer sheets rather than having answers embedded within a test booklet. Gaffney and Maguire (1971) tested 840 students between grades two to nine and looked at differences between those answering within a booklet and those answering on a separate answer sheet in number correct of easy items. Given seven items for which class means were 6.6 when answering within a booklet, second-graders had mean scores of 3.4, 5.0, and 5.1 when answering on a separate sheet under conditions of minimal instruction and no practice, maximum instruction and no practice, and maximum instruction plus practice. Third-graders also performed significantly below within-booklet class means under all conditions whereas fourth- and fifth-graders only performed below the within-booklet means when not allowed to practice. Above fifth grade, students performed as well on the separate answer sheets as those answering within the booklets.

These results concur with two studies by Cashen and Ramseyer (1969; Ramseyer & Cashen, 1971), in which they found significant differences based on grade and level of instruction. In the first, approximately 120 first to third graders were given the 1963 California Test of Mental Maturity in both booklet and answer sheet formats and were found to have decreasing performance differences dependent on format as grade increased. First graders had raw scores 23.67 points higher on the booklet format than answer sheet, a difference that decreased to 10.79 points in second graders and a non-significant 3.32 point difference in third graders. In the 1971 follow-up, first and second graders were first given instructions and found score differences decreased to less than 10 points in both groups. In 1985, Ramseyer and Cashen performed yet another study of score differences dependent on test-format, this time investigating interaction of this effect with eye-hand coordination levels. In first graders, they found a mean difference of 12 points between booklet and answer sheet formats for those with low or middle

eye-hand coordination but a difference of only 4.5 for those in the high eye-hand coordination group. No second-graders were in the low eye-hand coordination group and those in both the middle and high coordination groups had differences in scores of about 5 points in favor of within-booklet answering.

Muller, Orling, and Calhoun (1972) investigated differences in score reliability and variability dependent on whether answers were within-booklet or on a separate sheet. Examining groups of students at the third-, fourth-, and sixth- grade levels. At all three levels, mean number of errors tripled when responses moved from within booklet to a separate answer sheet while variability doubled in third graders and tripled for fourth and sixth graders, suggesting that moving to a separate answer sheet weakens test reliability.

Wise, Duncan, and Plake (1985) examined differences in test scores for 53 third graders divided into low, medium, and high abilities taking the Iowa Test of Basic Skills under three testing conditions: 1) answering within booklet, 2) answering on a separate sheet with no prior practice, and 3) answering on a separate sheet after several practice sessions. They found significant differences only in the low ability group and only between the separate sheet, no practice and the other two conditions. No significant difference was found between in-booklet answering and separate answer sheet with prior practice, suggesting that low-ability third graders can overcome format effects with training whereas they are not significant at other ability levels.

In their investigation into shift errors at Stony Brook University, Skiena and Sumazin (2000a) found that between 1-2% of paper-based tests contained shift errors, causing a loss of 10% of the student's grade, on average. An extension of the study to the SAT found 1.8% of 101,265 tests contained unrecognized shift errors, costing examinees up to 210 points, a penalty that could be devastating to one's chance of admission. According to the College Board website, over 3 million students sit for the SAT each year. The ACT, also used as a criterion for college admission and only administered in paper-and-pencil format, saw over 1.6 million examinees in 2012. Other high-stakes tests available, at least in part, in a paper-and-pencil format include the

Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL), Law School Admissions Test (LSAT), the Multistate Bar Examination (MBE), and the majority of statewide academic assessments (e.g., Massachusetts Comprehensive Assessment System (MCAS); Proficiency Assessment for Wyoming Students (PAWS)), including those in which a passing score is required for graduation. While studies related to answer sheet format seemed to show no significant effect after fifth grade, this specific effect was detected in students of high-school and college age. It is logical to conclude that, if young students are more generally prone to error based on test format, they will be more prone to this specific form of error as well.

Matter (1985), in an investigation of answer-changing behavior, observed 633 runs of 3 or more consecutive answer changes in a dataset expected to produce no more than 20 under an assumption of independent errors. Skiena and Sumazin (2000a) extrapolated from this that a substantial portion of all tests contain corrected shift errors and that, even if 90% of all shift errors are corrected, a significant problem remains concerning undetected shift errors.

Several articles including student surveys on answer-changing behavior provide further evidence of shift error as a source (McMorris & Weideman 1986; McMorris, DeMers, & Schwarz, 1987; Shatz & Best, 1987; Schwarz, McMorris & DeMers, 1991; van der Linden & Jeon, 2012). Shatz and Best (1987) interviewed 65 students upon their completion of a 62 item test and found that 20 changed answers after putting them initially in the wrong space, resulting in 70% changes from wrong to right. McMorris and Weideman (1986) found that 8% of all answer changes were due to mismarking of answer sheets. McMorris et al. (1987) found willingness to change answers based on clerical error was high (1.7 on a 1 to 2 scale) and that clerical error accounted for 18% of all changes made within their study. Schwarz et al (1991) found results concurrent with McMorris and Weideman, that, based on a series of personal interviews, 8% of changes made were due to clerical errors. van der Linden & Jeon (2012) discussed the rationale behind erasures and answer changing behavior and found that when changes are made to correct clerical error, the majority are from right to wrong.

2.4 Probability and Measurement Models

The shift detection methods proposed in this paper will principally employ item response theory (IRT) models, probabilistic models in which response probabilities are calculated based on a combination of the latent trait being measured in an examinee and characteristics of an item (e.g., item difficulty) (Hambleton, Swaminathan & Rogers, 1991). Methods employed will feature the dichotomous 3-parameter logistic (3PL; Birnbaum, 1968) model and the polytomous nominal response model (NRM; Bock, 1972). This section discusses each of these models as well as concerns that may need to be addressed in using them within the shift detection methods.

The 3PL model is a dichotomous model for determining the probability of a specific result given two possible outcomes (e.g., correct given possibilities of correct/incorrect, true given true/false options) on an exam item. The formula for its calculation is given by:

$$P(\theta) = c + (1-c)\frac{e^{1.7a(\theta-b)}}{1+e^{1.7a(\theta-b)}}$$

where $P(\theta)$ is the probability that an examinee with ability (θ) answers the item correctly, *a* is the item discrimination parameter, *b* is the item difficulty parameter, and *c* is the pseudo-chance parameter, which incorporates the probability of a low-ability examinee guessing the correct answer (Hambleton, Swaminathan & Rogers, 1991). This parameter provides a lower asymptote for probability of arriving at a correct answer, making the 3PL especially appropriate for use in multiple-choice exams in which even an examinee of no ability has a better than zero chance of guessing the correct response (de Ayala, 2009).

The NRM is a polytomous model that provides probabilities of all available responses for a given item, rather than a simple right/wrong. Consider a multiple-choice question with four response options. Though only one of those options may be considered correct, there is a probability associated with each of the available responses that depends on examinee's ability in the latent trait as well as the item characteristics. Whereas the 3PL requires only one calculation

per item, the NRM requires a calculation for each response option, the formula for which is given by:

$$P(u = k \mid \theta) = \frac{e^{a_k\theta + c_k}}{\sum_{i=1}^{n} e^{a_i\theta + c_i}}$$

in which $P(u = k \mid \theta)$ is the probability that response *u* is in category *k* given examinee θ and category slope and intercepts are given by a_k and c_k, respectively, for the specific responses and given gain for all responses from $i = 1$ to n by a_i and c_i, respectively.

The potential advantage of the NRM over the 3PL is that it provides probabilities for the incorrect responses as well as the correct ones. When employing the 3PL, the probability that the response is incorrect is simply 1-*P(θ)*, but this doesn't tell the probability of the specific incorrect response among the options. Instead one must divide the overall probability of an incorrect response by the number of incorrect response options:

$$P_w(\theta) = \frac{1 - P(\theta)}{(n-1)}$$

This makes an assumption that all incorrect responses are equally likely, an assumption that is unlikely to hold but perhaps adequate for accurate shift error detection.

Another possible concern employing the 3PL for shift error detection is that shift error detection methods must be tested and calibrated in simulation that depends on the NRM for data generation. As such, it becomes important that the 3PL is capable of adequately accurate parameter recovery of data generated using the NRM. "When data are generated with a particular model, scores based on that model will tend to be most accurate" (DeMars, 2008, p. 9). Because of differential scoring of incorrect responses, the NRM will tend to provide more information than the 3PL in the ability range in which an item is difficult for a given examinee (DeMars, 2008), but Thissen (1976) found that this does not result in an increase in score reliability because this offers no additional information gain for middle- and high-ability examinees. In DeMars'

(2008) study, results showed that overall, the NRM is more reliable than the 3PL in recovering parameters from data generated via the NRM, though differences in reliability were .04 across several conditions. For ability levels below zero, the NRM took advantage of extra information and showed a reliability difference of .1. For ability levels above zero, 3PL outperformed the NRM with reliability difference of .02 with large sample size (n=2500) and .06 for small sample size (n=250). Lower scores were, overall, more reliable than higher scores, perhaps making the increased reliability of the 3PL in the higher range of more value despite the 3PL's overall inferiority. While concluding that use of polytomous models may be large enough to be meaningful in certain contexts, DeMars does not conclude that use of the 3PL is inappropriate for parameter recovery of data simulated under the NRM.

2.5 Conclusions Based on a Review of Literature

When answers on paper and pencil tests are recorded on separate answer sheets rather than within test booklets, this can lead to inferior examinee performance, especially but not exclusively among children of early primary school age. Clerical errors under such test conditions are shown to be a problem for examinees of all ages, and one of the more common and most severe forms of clerical error that an examinee can make is to commit a shift error, a mistake in which an item is skipped and a series of responses misaligned to their intended items is produced. The longer the shift error, the more severely this can bias examinee ability estimates, threating the validity of any inferences one might wish to make based on these estimates. Shift errors are identified as a source of aberrance in the person-fit literature, but while person-fit statistics, even statistically optimal detection methods, are capable of detecting test pattern aberrance caused by shift errors, none offers a mechanism for identifying the cause of the misfit as a shift error. Absent the ability to identify specific causes of person misfit, the best course of action for remedying its presence remains unclear. Algorithms developed for the detection of shift errors have shown some capability for accurately detecting shift errors and differentiating them from false positives but leave room for improvement. IRT measurement models offer a clear avenue

for such improvement, as they are capable of providing accurate probabilities that item responses are being given in the correct location or shifted by a specific distance.

A preliminary study applying the 3PL model to strings of misaligned correct responses demonstrated promise for identifying shift errors, but realizing the full potential of IRT models in shift error detection requires employing algorithms that incorporate incorrect answers as well as correct ones. Further, the 3PL does not differentiate between incorrect responses, assigning equal probabilities to all of them within a given item. Employing the NRM, which provides probabilities for all response options on an item, could prove even more accurate in its detection of shift errors. Determining how sensitively and selectively these IRT-based models can detect shift errors, how robust they are to parameter bias, and how they compared to more general indices of person-fit offers the potential to identify and remedy this specific and potentially severe form of person misfit. Further applying the methods to empirical data will provide answers about just how prevalent this form of misfit is within actual tests.

This dissertation attempts to answer several specific research questions, all of which can be expressed more generally as variations on the following research questions:

1) Which combination of algorithm and model provides the highest shift error detection rates?
2) How robust is each method/model combination to the parameter bias introduced by the presence of shift errors?
3) Does the given algorithm/model combination outperform an appropriate person-fit statistic in identifying a candidate as misfitting the data?
4) How prevalent is shift error, as detected by the method/model combination, within an empirical data set?

CHAPTER 3

METHODOLOGY

3.1 Overview

With a goal of providing practical knowledge to improve upon the validity of test score interpretations through the detection and proper classification of one form of aberrant response behavior, namely undetected shift error on paper-and-pencil tests, this dissertation examined a small three-dimensional matrix of detection algorithms, probabilistic models, and person parameter estimation techniques, testing permutations of these factors for their selectivity and sensitivity in detecting shift errors under different simulated and empirical conditions. Additionally, these methods were compared to a more traditional person-fit statistic for their relative ability to identify, if not classify, this type of misfit.

The dissertation was broken down into four studies: 1) a simulation study based on empirical data, 2) an application of the results of the simulation study to the empirical data on which it was based, 3) a simulation study designed to determine if shift error detection methods perform differentially based on person parameter levels, and 4) a comparison of shift error detection methods and the H^T person-fit statistic for detecting shift errors.

What follows are the specifics of the detection algorithms, probabilistic models, and person parameter estimation techniques that make up the shift error detection methods followed by a detailed breakdown of the data and methods used in the four studies.

3.2 Shift Detection Algorithms

Two algorithms were employed for the error detection step within this study: (1) misaligned response detection, and (2) most probable correction detection. Both algorithms involve iteration through examinee response strings to detect anomalous patterns that may indicate that a shift error has been committed. The specifics of each method and the indices based on them are described next.

3.2.1 Misaligned Response Detection

The misaligned response detection algorithm involves comparison of the test answer key to a given examinee response string to find strings that correspond with but are misaligned to the answer key. For instance, looking for shifts forward in direction with a distance of 1, the algorithm will compare answer key item 1 to examinee response 2. If there is no match, the algorithm simply moves forward to compare answer key item 2 to examinee response 3. If they match, the item is flagged as the beginning of a misaligned response string. Answer key item 2 is then compared to examinee response 3 for a potential match. If there is a match, the next misaligned pair is compared, a process which continues until a mismatch occurs, at which point the examinee number, starting item, shift length, distance, and direction are recorded. The algorithm moves through the entire examinee response string recording all such misaligned response strings. Figure 2 shows an example of a misaligned response string forward in direction with a distance of 1, starting at item 7 and having a length of seven. The algorithm can do this for all examinees, shift directions and distances. Figure 3, for example, shows a response string with a misalignment that is backward with a distance of 1, starting at item 8 with a length of 6. Once the algorithm has processed all examinees, directions and distances, the resultant list can be evaluated against the probabilistic models to determine the likelihood of each having occurred in the correct position. The resultant probability is called the coincident misalignment probability (CMP) because it represents the probability that the misaligned agreement between substrings is due to coincidence rather than due to a shift error. Selectivity and sensitivity of CMP at different thresholds will be determined by classifying candidate substrings with CMP below the threshold as shift errors and candidates with CMP above the threshold as merely coincident, lower probabilities of coincidence being indicative of higher probabilities of a shift error. Under simulation conditions, shifts will be introduced that are a distance of only one forward or backward, and thus detection within the simulation using this algorithm will be limited to a

distance of one as well. This reduces the processing workload without undermining the investigation into the effectiveness of the shift error methods.

3.2.2 Most Probable Correction Detection

The algorithm for most probable correction detection requires no comparisons between an examinee's response string and the answer key, instead making the determination as to whether a substring is a shift error solely on changes in response probabilities when items are shifted. It does so by calculating three probability vectors for each examinee response string: 1) in-place probability: the probabilities associated with the responses in place as they occur on the answer sheet, 2) forward-shifted probability: the probabilities associated with the responses if each is shifted forward one position, and 3) backward-shifted probability: the probabilities associated with the responses if each is shifted backward one position. In practice, these shifts could be expanded to include shifts of larger distances, but were limited within this study to minimize the processing workload. In-place probability is then subtracted from forward-shifted probability for each item within each examinee's response string to obtain a vector of change in probability when responses are shifted forward and then from backward-shifted probability to obtain a vector of change in probability when responses are shifted backward. Subvectors of all lengths have their changes in probability summed within both the forward and backward change-in-probability vectors to find the subvectors with the largest total. The resultant index is called summed change in probability (SCIP). Because SCIP is a measure of improvement in probability when a shift error candidate is corrected, candidates are classified as true shift errors when SCIP is higher than a given threshold but classified as non-shift errors when SCIP is below the threshold. For the purposes of this study, because no more than one shift error was introduced in simulation within a given examinee response vector, flagging was, in turn, limited to only one shift error candidate: the subvector with the largest SCIP for that examinee.

3.3 Probability Models

For determining probabilities of suspicious substrings being in the correct location, two IRT models were applied: the 3PL model and the NRM. One problem with applying IRT models when there is potential misfit is that the misfit gets built into the parameters themselves, biasing the parameter estimates and, therefore, any probabilities calculated based on those estimates. This suggests two applications of each of these models: one in which item and person parameters are all treated as known (using the known parameters employed in the simulation) and one in which they are estimated after simulation. While known parameters are not available in empirical situations, by evaluating methods using both known and estimated ability parameters allows us to not only see how much bias is introduced to the ability estimates by the shift errors, but to also see how that bias impacts the shift error detection. Additionally, when methods are taken to control that bias, comparison to the baseline based on the known parameters becomes essential.

3.3.1 3PL Model

The 3PL Model for calculating probabilities of response strings incorporates both item and person parameters into the calculation. Within the 3PL, described in more detail within the literature review, the probability of a correct response is given by:

$$P(\theta) = c + (1-c)\frac{e^{1.7a(\theta-b)}}{1+e^{1.7a(\theta-b)}}$$

where $P(\theta)$ is the probability that an examinee with ability (θ) answers the item correctly, *a* is the item discrimination parameter, *b* is the item difficulty parameter, and *c* is the pseudo-chance parameter, which incorporates the probability of a low-ability examinee guessing the correct answer. Probability of an incorrect response can be expressed as:

$$P_w(\theta) = \frac{1-P(\theta)}{(n-1)}$$

where $P_w(\theta)$ is the probability of an incorrect response and *n* is the number of response options. To calculate the probability of a particular response string of length *l*, one need multiply all of the calculated probabilities associated with that string:

$$P_{substring}(\theta) = \prod_{i=1}^{l} P_i(\theta)$$

3.3.2 Nominal Response Model

The Nominal Response Model (NRM; Bock, 1972), like the 3PL, incorporates person and item parameters into calculation of response probabilities. Unlike the 3PL, which only calculates the probability of an examinee choosing the correct response, the NRM calculates the probabilities of the examinee choosing each response option. Within the NRM, described in more detail within the literature review, the probability of each response for an item is given by:

$$P(u = k \mid \theta) = \frac{e^{a_k\theta + c_k}}{\sum_{i=1}^{n} e^{a_i\theta + c_i}}$$

in which $P(u = k \mid \theta)$ is the probability that response *u* is in category *k* given examinee θ and category slope and intercepts are given by a_k and c_k, respectively, for the specific responses and given gain for all responses from $i = 1$ to n by a_i and c_i, respectively. Unlike the 3PL, which produces only the probability of an incorrect response and, in order to calculate individual incorrect response probabilities, requires the unsubstantiated assumption that all distractors are equally attractive regardless of examinee ability, application of the NRM to multiple choice items results in calculation of unique probabilities for both the correct and incorrect responses that consider the measurement properties of each response option.

As with the 3PL model, to calculate the probability of a particular response string of length *l* using the NRM, one need multiply all of the calculated probabilities associated with that string, and may use the same formula.

3.4 Person Parameter Estimation Techniques

Person parameter estimation was performed in three ways within this study: 1) using the true parameters, obtained through calibration of the empirical data and treated as truth during response simulation, 2) using the parameter estimates obtained during calibration of the simulated response sets, and 3) correcting those estimates for bias introduced by the presence of shift error candidates. Person parameters were obtained for all technieques using expected a posteriori (EAP) estimation. This was determined to be the best compromise between accuracy relative to maximum likelihood estimation (MLE), given that large amounts of data were treated as missing for bias control, and efficiency relative to maximum a posteriori (MAP) estimation, whose iterative procedure proved too time-intensive given the large data set and great number of conditions and replications within this series of studies.

Simulation of the data set used in this study included the use of known person parameters estimated from an empirical data set then treated as known. While these would remain unknown in empirical studies, we can use this knowledge to evaluate our data for shift errors under an assumption that we are able to remove all bias before estimation. This may be an unrealistic scenario and is not technically estimation, but provides a sense of the best case scenario and allows us to examine the influence of person parameter bias due to shift errors on the process of shift error detection.

Under empirical conditions, item and person parameters are unknown; they are estimated through the calibration process. Though person parameters were known under the simulation conditions, no such luxury would exist when applying shift error detection methods to empirical data. As such, estimates obtained based on the simulated data better represent how these methods would necessarily be applied under empirical conditions. By comparing the effectiveness of the shift error detection methods using both known and estimated parameters, we can get a sense of how well these methods work under realistic conditions as well as how much detection ability is lost to the bias created by the shift errors themselves. When person parameter estimates are biased

based on systematic error, such as undetected shift error, use of the parameters in detecting said shift error will be less than optimal. Underestimates of ability will, in addition to presenting threats to validity, result in less powerful probability calculations, making detection of shift errors more difficult.

Under the assumption that the bias introduced through estimation when shift errors are present would noticeably impact the estimation and impair the shift error detection, a purification method for reducing that bias was employed under the hope that this would improve shift error detection accuracy. The purification method for reducing the bias involved running each shift error detection twice, treating shift error candidates identified during the first run based on the estimated person parameters as missing, re-estimating item and person parameters, and basing a second shift error detection on these bias-corrected parameters for comparison to detection using the uncorrected estimated person parameters. The threshold for making the determination of which candidates to omit was the threshold at which all true positives were included along with any false positives that fell on the correct side of the threshold, while candidates falling on the wrong side of the threshold were not treated as missing.

3.5 Decision Criteria

Once substrings are flagged as potential shift errors, they must be compared against some criterion or criteria in order to classify them as either shift errors or coincident misaligned patterns. Three criteria are suggested for possible classification of shift errors: 1) probability threshold, 2) person-fit indices, and 3) change in bias of ability estimates. It may be that a simple comparison of substring probabilities to a probability threshold may prove adequate in providing accurate and inclusive shift error detection, but application of these other criteria could improve accuracy and inclusivity or provide additional support for conclusions drawn based on probability.

3.5.1 Probability Threshold

The foundation of the proposed detection methods is, in all cases, a probability calculation, whether it is a single calculation of probability that a suspicious response substring would occur in its place, as per the misaligned response detection method, or it takes two calculated probabilities, that of the substring occurring in place to identify it as suspicious and a second to evaluate the probability of a proposed corrective shift. Classifying a given substring as either a shift error or merely a coincident misalignment depends on the calculated probability or probabilities falling on the correct side of a thresholds set to ensure an acceptable level of accuracy. In the case of the single calculation method, probabilities would need to fall at or below the threshold. In the case in which two calculations are performed, optimal thresholds for both flagging substrings as suspicious and for classification as shifts must be determined. Because optimal probabilities are dependent on acceptable ratios of true to false positives, ROC curves presenting the trade-off between true positive and false discovery rates were used to make this determination. Additionally, because some procedures involve making evaluations at specific thresholds, these were set based on arbitrarily acceptable false discovery rates of .00, allowing no false positives, and .05, allowing 5% of all classified shift errors to be falsely identified negatives.

3.5.2 Person-Fit Indices

In addition to meeting the probability criterion, it may be that a person fit statistic or a change in person fit may provide a more sensitive and/or selective indicator of a shift error than a probability threshold. Rather than setting probability thresholds at levels that optimize selectivity and sensitivity, it could be that change in person fit serves as a better index. Rather than varying probability thresholds, thresholds of person fit or change in person fit can be assessed in the same way, via ROC curves, to determine the tradeoff between Type I and Type II errors. To this end, an appropriate person-fit index needed to be selected. Because it was found to be superior under all conditions (Karabatsos, 2003), H^T (Sijtsma & Miejer, 1992) described in section 2.2.2 above, was selected for this purpose. Because H^T involves calculation of covariances between all

examinees, a task too large given the scope of the first simulation substudy, comparisons between the proposed shift error detection methods and person-fit were conducted in a separate substudy using a small subsample of examinees from that first substudy.

3.5.3 Change in Bias

When person parameters are known or estimated without shift errors, they may be compared against the ability estimates obtained including the response substring identified as a possible shift error to see how much bias that potential shift error has produced. Under simulation conditions, pre-shift correction and post-shift correction theta estimates may be obtained and compared against the known thetas to determine which response pattern biases the estimate less. Under empirical conditions (or in an attempt to emulate empirical conditions), ability estimates may be obtained with identified candidate substrings shifted, unshifted and removed from estimation altogether. Provided that bias-correction methods are effective, comparison of pre-shift and post-shift estimates to the estimates with substrings removed would be safe under the assumption that those represent the best available estimates for comparison. As will be seen in the first study, attempting to correct bias by removing potential shift errors from examinee response vectors provided no predictable pattern of improvement in ability estimation or shift error detection, so the assumption that bias-corrected estimates were the best available was unsafe and this was discarded from the subsequent studies.

3.6 Study 1: Simulation Study Based on Empirical Data

The first study was designed to evaluate the accuracy and effectiveness of the proposed shift error detection methods as they might be applied under empirical conditions. To that end, item and person parameter estimates were estimated from an empirical data set using the NRM and treated as true parameters for the sake of simulation. Data was obtained from the administration of a paper-and-pencil K-12 proficiency exam with 45 items given to a set of approximately 40,000 examinees. In order to build the response sets for each scenario, responses were first simulated without shift errors using the NRM. Simulation was performed by calculating

probabilities of all response options for each examinee taking each item using the empirical item and person parameters under the NRM then comparing a randomly generated number to those response probabilities to determine the response for each examinee/item combination. The NRM was selected because of its ability to probabilistically differentiate between the different response options.

Four scenarios were simulated, consisting of short, medium, long, and mixed-length shift errors. For the short, medium, and long shift error scenarios, shift lengths were set to 3, 7, and 10 respectively. For the mixed-length scenario, shifts were evenly split between every length from 3 to 10. These lengths were selected for their comparability to the Skiena and Sumazin (2000a, 2000b, 2004) series of studies and because they place reasonable bounds on lengths below which shift error detection is unlikely to behave reliably and above which shift error detection trends are likely to be predictable. For the fixed-length scenarios, rather than look at all lengths between 3 and 10, as Skeina and Sumazin did, choosing the shortest and longest and a midpoint for examination served to inform as to general patterns in regard to the research questions without overburdening the studies with time-intensive replications producing nearly similar results. Where inconsistencies surfaced could inform future research on other lengths. For all scenarios, five percent of examinees received a shift error, shift errors were limited to one per examinee, and shifts could occur anywhere within an examinee's response sequence so long as the starting point occurred early enough to leave room for the entire shift before the end of the sequence. While five percent is larger than was projected by previous research, it provided a larger sample of shift errors within the simulation, allowing larger numbers of shift errors to be examined in fewer replications than if Skiena and Sumazin's two percent was used. For each scenario, 100 replications of the 45 item by 40,000 examinee matrix were produced and shift errors of the appropriate lengths were then introduced. Once data sets were generated, the matrix of IRT probability models, person parameter estimation method, and shift detection algorithm were applied to obtain lists of candidate substrings and their associated shift detection indices.

To classify candidate substrings as shift errors, indices must be compared to thresholds that result in acceptable type I and type II error rates. Because of the disproportionate ratio (19 to 1) of examinees without shift errors to those with them within this simulation, using false positive rates, which compare the number of false positives to the maximum number of false positives (i.e., actual negatives), could provide misleading results. If, for instance, a false positive rate of .05 were deemed acceptable, within this study approximately 2000 false positives would be deemed acceptable. If, at the same time, 1000 true positives were found out of 2000 total positives, one could conclude that there is a 50% hit rate at an alpha of .05. Meanwhile, for any given positive, there is only a one in three chance that it is a true positive. For this reason, this and the following studies instead used false discovery rates (Benjamini & Hochberg, 1995) in contextualizing type I errors. False discovery rate is the ratio of false positives to total positives. In the given example, 2000 of 3000 positives are false, a false discovery rate of .67. This provides a much more meaningful indicator of the effectiveness of the shift error detection than the false positive rate of .05. Additionally, stating that a false discovery rate of .05 is acceptable would be more clearly meaningful, indicating that 1 in 20 positives being false is an acceptable rate for type I errors.

Because acceptable error rates are dependent on application, curves were constructed to show the tradeoff between true positive rates and false discovery rates over the entire threshold range. These curves are similar to receiver operating characteristic (ROC) curves (Green & Swets, 1966), which plot false positive rates on the *x* axis and true positive rates on the *y* axis, the difference being that false positive rates are replaced by false discovery rates on the *x* axis. Additionally, thresholds with false discovery rates of .00 and .05 were established to determine what percent of shift errors were found both without error and at an arbitrary ratio of true to false positives. These arbitrary false discovery rates were essential to applying the detection methods back to the empirical data in the second study.

In addition to determining true positive rates at the thresholds with false discovery rates of .00 and .05, calculation of the effect of correcting detected shift errors on person parameter bias was performed. Bias is simply the difference between the examinee's true ability and the examinee's estimated ability:

$$\text{Bias}(\hat{\theta}) = \hat{\theta} - \theta$$

For change in bias for a set of examinees, mean change in absolute bias (MCAB) and mean change in signed error (MCSE)were calculated. MCAB is given by the following formula:

$$MCAB = \frac{\sum_{i=1}^{n} \left| \text{Bias}(\hat{\theta})_{i\,post} \right| - \sum_{i=1}^{n} \left| \text{Bias}(\hat{\theta})_{i\,pre} \right|}{n}$$

where n is the number of examinees, $\text{Bias}(\hat{\theta})_{i\,pre}$ is the value of $\text{Bias}(\hat{\theta})$ for examinee i based on the ability estimates made prior to shift error detection and $\text{Bias}(\hat{\theta})_{i\,post}$ is the value of $\text{Bias}(\hat{\theta})$ for examinee i after re-estimating thetas after shift error detection and correction has been performed. MCAB is a reasonable calculation of change in bias in this study because, whereas we expect the bias introduced by shift errors to be negative, our goal in correcting shift errors is reducing the overall magnitude of the bias to the ability estimates. Strictly measuring change is not as meaningful since large changes may not be indicative of improvement if they merely swing underestimates to equally large overestimates, for instance. Since MCAB is calculated for each of the 100 replications, the average MCAB over all replications is obtained by summing them and dividing by 100 and one figure for MCAB is reported for a given scenario.

MCSE is given by the following formula:

$$MCSE = \frac{\sum_{i=1}^{n} \text{Bias}(\hat{\theta})_{i\,post} - \text{Bias}(\hat{\theta})_{i\,pre}}{n}$$

where individual terms are the same as for MCAB. MCSE gives an indication as to the direction in which bias moves. In conjunction with MCAB, this helps to understand the effect of shift error

correction. One would expect correction of shift errors to increase scores (resulting in positive MCSE) while reducing the magnitude of bias (resulting in negative MCAB). Since MCSE is calculated for each of the 100 replications, the average MCSE over all replications is obtained by summing them and dividing by 100 and one figure for MCSE is reported for a given scenario.

3.7 Study 2: Empirical Application of Simulation Study Results

The second study was an application of the simulation results to the empirical data from which the item and person parameters were derived. Should any of the shift error methods prove useful operationally, these are logical first and second steps for conducting such an application: performing a study based on the test you wish to investigate for shift errors, introducing shift errors in simulation to determine thresholds for their detection at reasonable type I and type II error rates, then applying the results back to the empirical data to find real shift errors within that data.

Within this study, a two-dimensional matrix of detection algorithms and probability models was employed. Person parameter estimation method was discarded as a factor since true person parameters remain unknown and bias control methods proved unsatisfactory in study one, leaving uncorrected person parameter estimates as the only working measure of ability. The four algorithm/model combinations were applied using these uncorrected person parameters to calculate the shift detection indices, which were then compared to the thresholds from the fixed and mixed-length shift error scenarios of the simulation study at false discovery rates of zero and .05 in order to obtain true and false positive counts within the empirical data. Additionally, based on the true positive rates found in the simulation and the empirical true positive counts, projections of total positive counts were calculated at those thresholds and, for the mixed-length scenarios, across the false discovery range in order to determine the consistency of projections across the full range.

As in the simulation study, it is interesting to know the benefit of correcting the shift errors in terms of change in bias for those whose shift errors were corrected. Unfortunately, bias

is incalculable because, whereas in simulation calculations are straightforward because examinee's true ability (θ) is known, with empirical data, examinee's true ability is unknown. Because simulation conditions are modeled after the empirical data, MCAB obtained through simulation can be used as an indicator of expected benefit when applied empirically, but to get a sense of the actual impact of shift errors on ability estimates in empirical data, a change in ability estimate is a sensible alternative. For change in ability, Mean Absolute Difference (MAD) and Mean Signed Difference (MSD) statistics were calculated. MAD is given by the formula:

$$MAD(\hat{\theta}) = \frac{\sum_{i=1}^{n} \left| \hat{\theta}_{ipost} - \hat{\theta}_{ipre} \right|}{n}$$

where n is the number of examinees, $\hat{\theta}_{ipre}$ is the ability estimate for examinee shift error detection and $\hat{\theta}_{ipost}$ is the ability estimate for examinee *i* re-estimated after shift error detection and correction has been performed. Note that whereas MCAB represents an improvement, MAD represents change. Given that the majority of changes made to response strings were presumed to be on true shift errors and only a small percentage were adjustments to false positives, it was assumed that this change was an improvement, but this remains unknown.

MSD is given by the formula:

$$MSD(\hat{\theta}) = \frac{\sum_{i=1}^{n} \hat{\theta}_{ipost} - \hat{\theta}_{ipre}}{n}$$

where notation is the same as for MAD. Whereas MAD gives a measurement of the magnitude of movement in person parameter estimates, MSD gives a measurement of the direction of that movement.

3.8 Study 3: Simulation Study Based on Stratified Ability Levels

Shift error detection is likely to behave differentially dependent on examinee ability. Because shifted correct responses are more likely to appear as incorrect responses and lower

examinees are more likely to answer items incorrectly, the higher probabilities associated with incorrect responses in low ability examinees should also result in higher probabilities associated with misaligned response strings. In the case of the misaligned substring detection method, which looks only for coincident correct responses, stopping at the first disagreement between misaligned responses, an incorrect response within a true shift error will shorten the detected misalignment, further inflating the probability of their occurrence and reducing the likelihood that they will be properly identified as shift errors. As such, it becomes interesting to see how the methods work at different ability levels. This study will look at the methods at different examinee ability levels.

In order to determine whether the proposed shift error methods function differentially dependent on examinee ability, the third study replicates the procedure from study one using smaller samples and fixed, stratified person parameters. Three samples of 2000 examinees were simulated, the first having person parameter fixed at negative one, the second having a person parameter fixed at zero, and the last having a person parameter fixed at one. Item parameters were the 45 items taken from the same paper-and-pencil K-12 proficiency exam and treated as true parameters for the purpose of simulating responses. In order to build the response sets for each scenario, responses were first simulated without shift errors using the NRM.

Four shift error length scenarios were simulated, consisting of short, medium, long, and mixed-length shift errors. For the short, medium, and long shift error scenarios, shift lengths were set to 3, 7, and 10 respectively. For the mixed-length scenario, shifts were evenly split between every length from 3 to 10. For all scenarios, five percent of examinees received a shift error, shift errors were limited to one per examinee, and shifts could occur anywhere within an examinee's response sequence so long as the starting point occurred early enough to leave room for the entire shift before the end of the sequence. For each person-parameter/shift-length combination, 100 replications of the 45 item by 2000 examinee matrix were produced and shift errors of the appropriate lengths were then introduced. Once data sets were generated, the matrix of IRT

probability models, person parameter estimation method, and shift detection algorithm were applied to obtain lists of candidate substrings and their associated shift detection indices.

Results were reported as for study one, using false discovery rates rather than false positive rates, building modified ROC curves, and reporting the true positive rates and thresholds with zero false positives and at the false discovery rate of .05. Change in mean absolute bias was also calculated based on shift errors detected at those false discovery rates.

3.9 Comparison of Shift Error Detection Methods to the H^T Person-Fit Statistic

The proposed shift detection indices, SCIP and CMP, are both person-fit statistics directed toward one specific form of person-misfit: shift errors. While more general person-fit statistics may lack the capability to identify the specific location of misfit within an examinee's response string or the specific nature of the misfit, it is possible that it may do a better job of identifying the misfitting individuals who have committed a shift error. This study compares SCIP and CMP to H^T, the person-fit statistic found in Karabatsos' (2003) analysis to perform superiorly to all other person-fit statistics across test lengths, sample sizes, misfit types, and misfit saturations. Additionally, because the shift-error detection methods stand to improve person fit, this study will investigate the change in the H^T statistic based on correcting the detected shift errors at certain thresholds.

In order to evaluate the relative performances of SCIP, CMP, and H^T, this study uses the simulated responses from the first study and calculates the H^T person-fit index in order to determine its misfit detection rate for comparison to the shift error detection rates of the shift error detection methods. This was done for each of the four shift error length scenarios: short, medium, long, and mixed-length shift errors. Results were reported as for study one, using false discovery rates rather than false positive rates, building modified ROC curves for all three indices in all scenarios, and reporting the true positive rates and thresholds with zero false positives and at the false discovery rate of .05. Additionally, corrections to the response strings were made based on shift errors detected at those false discovery rates and H^T was recalculated in order to

determine what effect, if any, correcting shift errors has on person fit. Calculation of changes in person-fit allow for evaluation of the benefit realized by applying a given shift error detection method. Overall change in person-fit can be expressed by the formula:

$$\sum_{i=1}^{N}(H_{i\ pre}^{T} - H_{i\ post}^{T})$$

where N is the number of examinees, $H_{i\ pre}^{T}$ is the value of H^T for examinee *i* prior to correction based on the shift error detection result and $H_{i\ post}^{T}$ is the value of H^T for examinee *i* after correction of a detected shift error, if any. Overall change in person fit was calculated for all examinees in each scenario as well as only for those who had shift errors.

3.10 Calibration

Calibration in all studies was performed using the FlexMIRT software package (Cai, L., 2012) using the 3PL and NRM as appropriate to each scenario's selected probability model. Where the 3PL model was used, *c*-parameters were held fixed at .25, a reasonable estimate of each item's guessability that stabilized the calibration process, a necessity given the study size and the lack of stability of the *c*-parameters that was found when they were allowed to be estimated freely. In study three, which detected shift errors at fixed person parameter levels, all item parameters were fixed for both the 3PL and NRM in estimating the person-parameters after responses were simulated.

CHAPTER 4

RESULTS

4.1 Overview

The dissertation was broken down into four studies: 1) a simulation study based on empirical data, 2) an application of the results of the simulation study to the empirical data on which it was based, 3) a simulation study designed to determine if shift error detection methods perform differentially based on person parameter levels, and 4) a comparison of shift error detection methods and the H^T person-fit statistic for detecting shift errors. Results for each of the four studies will be reported separately. For all studies, index thresholds for given scenarios and false discovery rates are set based on the mean index value across the replications for the given scenario. Devations in false positive counts at this mean index were small enough at all false discovery rates as to be considered insignificantly different from the expected counts at those rates.

4.2 Study 1: Simulation Study Based on Empirical Data

The first study was designed to evaluate the accuracy and effectiveness of the proposed shift error detection methods as they might be applied under empirical conditions. Results of the study are broken down into sections by shift length, and within those sections, results from the probability model, parameter estimation method, and shift detection algorithm matrix are reported.

4.2.1 Short Shifts

Short shift errors were of length 3. Figures 4 to 7 show ROC curves separated by scoring algorithm and probability model while Figures 8 to 10 shows all algorithm/model combinations using the true, estimated, and bias-corrected person parameters. A comparison of the ROC curves shows that for these short shifts, CMP is more sensitive to true positives than SCIP at the lowest false discovery rates but at false discovery rates of .04 and higher, SCIP surpassed CMP using the NRM, and at the false discovery rates of .07 and higher, SCIP surpassed CMP using the 3PL. For

both SCIP and CMP, the NRM was consistently more sensitive to true positives than the 3PL. Loss in sensitivity due to estimation bias is greater using CMP at these short lengths, though it is also consistent along most of the ROC curves for 3PL misaligned detection. For SCIP, there is little loss of sensitivity due to bias and at some points the estimated person parameters even outperform the true person parameters. Controlling for bias hindered the detection methods at this short length for all algorithm/model combinations except for CMP using the NRM, which saw it recover most of the loss due to estimation bias. For the SCIP index, bias control rendered shift errors undetectable or nearly so below false discovery rates of .05.

Table 5 shows detection rates of all method combinations for the short shifts at the false discovery rate of .00 and Table 6 shows detection rates with a false discovery rate of .05. These confirm what the ROC curves tell us: When allowing no false positives, CMP outperformed SCIP, with the NRM detecting 10.0% of the true shift errors using true person parameters, 2.8% using uncorrected estimates, and 4.6% using bias-corrected estimates, while the 3PL detected shift errors at rates of 7.8%, 5.0% and 4.4%, respectively. For both IRT models, SCIP detected approximately 2% of true shift errors using both true and estimated thetas and detected nearly zero shift errors when employing the bias-control method. At the false discovery rate of .05, SCIP outperformed CMP and NRM continues to outperform 3PL. SCIP using NRM detected 23.2% of shift errors of length 3 using true person parameters and barely dropped off when using estimated person parameters, detecting 22.4% of shift errors, though attempting to correct for bias yielded only minimally better results than when attempting to do when no false positives were allowed. Also of interest at the .05 false discovery rate is that estimated person parameters had higher shift error detection rate than true person parameters for SCIP using the 3PL , with a true positive rate of 18.1% compared to 13.5%.

Tables 7 and 8 give the MCAB and MCSE for shifts of length 3 when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of .00. Applying the 3PL model gave an MCAB for all

examinees of slightly more than 0.020 across all models and person parameters while MCAB for shifted examinees showed a reduction in absolute bias, with values between -0.020 and -0.041 using CMP and between -0.009 and -0.017 using SCIP. MCSE is 0.000 across models and person parameters when all examinees are considered while those with shift errors had an MCSE between 0.041 to 0.083. Applying the NRM produced superior results using the CMP algorithm, reducing absolute bias in all examinees by 0.001 to 0.003 and improving on reduction in absolute bias using true and bias-corrected parameters while losing some reductive power using estimated person parameters. SCIP resulted in either no change in magnitude bias or a decrease of 0.001 for all examinees and showed slight bias reduction using true and estimated person parameters but attempts at bias correction were ineffective. MSE under all algorithms and person parameters was 0.000 for all examinees, showing no directional tendency to at least three decimal places. For the shifted group, attempts to correct the shift showed a tendency to moved signed errors toward the positive, indicating that shift error correction was tending to raise scores.

Tables 9 and 10 give the MCAB and MCSE for shifts of length 3 when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of .05. Results were nearly identical to when the false discovery rate is zero with a slight reduction in magnitudes across the board, making the positive MCAB slightly less damaging under the 3PL at this threshold and the negative MCAB slightly more favorable under the NRM. MCSE also behaved similarly to the lower false discovery rate, showing no tendency either way when applied to all examinees but a positive tendency when applied to the shifted examinees.

The result of introducing shift errors of length 3 into the data was an MCAB of 0.007 compared to the unshifted data. While attempts to correct shift errors using the 3PL at the false discovery rates of .00 and .05 increased the MCAB, applying the NRM cut the amount of bias introduced by the shift errors by between 14% and 57%.

4.2.2 Medium Shifts

Medium shift errors were of length 7. Figures 11 to 14 show ROC curves separated by scoring algorithm and probability model while Figures 15 to 17 show all algorithm/model combinations for true, estimated, and bias-controlled person parameters, respectively. A comparison of the ROC curves shows that, for medium shifts, behavior is much more predictable across detection algorithms, probability models, and person parameter estimation methods than it was for the short shifts. Along the false discovery rate range, SCIP proves more sensitive to shift errors than CMP and the NRM is more sensitive than the 3PL. Bias caused by using estimated rather than true person parameters consistently reduces selectivity, though SCIP using NRM proves most robust to this bias. The bias control effort proves minimally effective, except for CMP using NRM, which showed considerable improvement under bias-controlled conditions at very low false discovery rates and, less usefully, bias control was very effective for the SCIP algorithm using the 3PL model at false discovery rates greater than .20. Across the board, true positive rates were higher for medium length shifts than for short shifts at comparable false discovery rates.

Table 11 shows detection rates of all scenarios for the medium shifts without false positives and Table 12 shows detection rates with a false discovery rate of .05. Again, these support what the ROC curves tell us: that SCIP outperformed CMP and NRM outperformed 3PL. When no false positives were allowed, SCIP using NRM detected shifts of length 7 at rates of 49.2%, 44.6% and 37.9% using true, estimated, and bias-corrected person parameters respectively. SCIP using 3PL yielded true positive rates of 43.2%, 40.0%, and 33.4%, CMP using NRM had rates of 39.0%, 28.6%, and 33.1% and CMP using 3PL had rates of 34.2%, 26.7%, and 25.0%. At a false discovery rate of .05, SCIP using NRM continued to outperform the other algorithm/model combinations for all ability estimation methods, with rates of 70.0% using true person parameters, 66.7% estimated, and 65.8% bias-corrected. SCIP using 3PL had true positive

rates of 64.7%, 59.7%, and 60.5%, CMP using NRM had rates of 54.4%, 45.8%, and 49.3% and CMP using 3PL had rates of 48.5%, 42.3%, and 38.2%.

Tables 13 and 14 give the MCAB and MCSE for shifts of length 7 when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of zero. Applying the 3PL model resulted in an MCAB for all examinees of around 0.010 across all models and person parameters while MCAB was negative for the shifted examinees. Results were slightly better using SCIP compared to CMP. MCSE was 0.000 across models and person parameters when all examinees were considered while those with shift errors had MCSE that was systematically positive with values between 0.262 and 0.378. Applying the NRM reduced the MCAB across algorithms and person parameters. When the CMP algorithm was employed, MCAB was - 0.010 to -0.013 while SCIP results in an MCAB between -0.012 and -0.015. Shifted examinees saw larger reductions that were similar but slightly better than those using the 3PL. MCSE under all algorithms and person parameters was 0.000 under the NRM, showing no directional tendency, while MCSE for the shifted examinees was positive, showing the tendency to raise scores when correcting shift errors.

Tables 15 and 16 give the MCAB and MCSE for shifts of length 7 when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of .05. The more permissive threshold shows all of the same trends between algorithms and models as when zero false positives are allowed but was uniformly better at reducing the magnitude of bias while raising scores for the shifted examinees.

The result of introducing shift errors of length 7 into the data was an MCAB of 0.018. While attempts to correct shift errors using the 3PL at the false discovery rates of .00 and .05 increased the absolute bias, applying the NRM cut the amount of bias introduced by the shifts by between 61% and 94%.

4.2.3 Long Shifts

Long shift errors were of length 10. Figures 18 to 21 show ROC curves separated by scoring algorithm and probability model while Figures 22 to 24 show all algorithm/model combinations using true, estimated, and bias-corrected person parameters, respectively. A comparison of the ROC curves shows that, for long shifts, as for medium, behavior was much more predictable across detection, person parameter estimation methods, and probability models algorithms than for the short shifts. Along the full selectivity range, SCIP proved more sensitive to shift errors than CMP and the NRM was more sensitive than the 3PL. Sensitivity for detecting long shifts continued the trend of improvement across probability models, detection algorithms, and person parameter estimation methods as length increases. Bias caused by using estimated rather than true person parameters consistently reduces selectivity, though once again, SCIP using the NRM proved most robust to this bias. At this shift error length, bias correction yielded consistently better results for the SCIP algorithm, was minimally effective at small false discovery rates for the CMP algorithm using the NRM, and performed worse than the estimated person parameters for the CMP using the 3PL model.

Table 17 shows true positive rates of all scenarios for the long shifts without false positives and Table 18 shows true positive rates with a false discovery rate of .05. Once again, these support what the ROC curves tell us and they present a very similar message to what they did for the medium length shifts in terms of how methods performed relative to one another while continuing the trend of improved sensitivity as shift errors get longer. When no false positives were allowed, SCIP using the NRM detected shifts of length 10 at true positive rates of 65.7%, 58.2% and 58.3% using true, estimated, and bias-corrected person parameters, respectively. SCIP using the 3PL yielded rates of 60.2%, 51.6% and 53.6%, CMP using the NRM had rates of 49.3%, 36.5%, and 38.9%, and CMP using the 3PL had rates of 42.9%, 31.2%, and 30.6%. At the false discovery rate of .05, SCIP/NRM continued to outperform the other algorithm/model combinations for all ability estimation methods, with sensitivity rates of 81.3% using true person

parameters, 77.8% estimated, and 79.1% bias-corrected. SCIP/3PL was almost as good, with rates of 77.6%, 70.0%, and 75.5% respectively. CMP sensitivity was considerably worse for both of the probability models, with true positive rates of 65.8%/55.3%/57.0% for the NRM and 59.9%/48.8%/45.5% for the 3PL model.

Tables 19 and 20 give the MCAB and MCSE for shifts of length 10 when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of zero. Applying the 3PL model and the CMP algorithm resulted in an MCAB for all examinees of 0.008 when using estimated or bias-corrected person parameters and 0.005 with true person parameters. For shifted examinees, reductions in bias were substantial, with MCAB ranging from -0.236 to -0.300. The 3PL with the SCIP algorithm showed even more favorable results, increasing bias in all examinees with MCAB between 0.000 and 0.030 while decreasing bias in shifted examinees with MCAB between -0.337 and -0.392. Applying the NRM showed further improvements on MCAB, even showing small decreases when applied to all examinees. When the CMP algorithm is employed, MCAB shows a reduction in absolute bias, with values between -0.015 and -0.018 while SCIP results in further decreases, with MCAB between -0.021 and -0.023 for all examinees. For the shifted examinees only, CMP had MCAB between -0.256 and -0.313 while SCIP decreased absolute bias by even more, having an MCAB between -0.376 and -0.410. MCSE under all conditions was 0.000 for the all examinees group but continued the trend of moving toward the positive when correcting examinees with shift errors.

Tables 21 and 22 give the MCAB and MCSE for shifts of length 10 when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of .05. As for shifts of length 7, improvement over the zero false positive rate is uniform across conditions while trends between those conditions are consistent with those at the more restrictive rate.

The result of introducing shift errors of length 10 into the data was an increase in MCAB of 0.025. While attempts to correct shift errors using the 3PL at the false discovery rates of .00 and .05 increased or only minimally reduced the magnitude of bias, applying the NRM cut the amount of shift-induced bias by between 86% and 98%.

4.2.4 Mixed Length Shifts

Mixed length shifts included an equal number of shifts of every length from 3 to 10. Figures 25 to 28 show ROC curves separated by scoring algorithm and probability model while Figures 29 to 31 show all algorithm/model combinations using the true, estimated, and bias-corrected person parameters, respectively. A comparison of the ROC curves show a continuation of the trends established in the fixed-length shift scenarios and are especially reminiscent of the medium-length shift error ROC's. SCIP continued to prove more sensitive to shift errors than CMP and the NRM was again more sensitive than the 3PL model. Bias caused by using estimated rather than true person parameters consistently reduced sensitivity, though once again, SCIP using NRM proved most robust to this bias. The bias control performed well for CMP using the NRM at all false discovery rates and for SCIP using 3PL at false discovery rates over .3.

Table 23 shows detection rates of all scenarios for the mixed length shifts without false positives and Table 24 shows detection rates with a false discovery rate of .05. Detection rates were similar but slightly smaller than for the medium shifts. When no false positives were allowed, SCIP with the NRM detected shifts at rates of 39.9%, 35.7% and 30.0% using true, estimated, and bias-corrected person parameters, respectively. SCIP using the 3PL model yielded rates of 35.2%, 32.3% and 26.7%, CMP with NRM had rates of 33.6%, 22.3%, and 27.1%, and CMP with the 3PL had rates of 29.1%, 21.4% and 21.1%. At the false discovery rate of .05, SCIP/NRM continued to outperform the other algorithm/model combinations for all ability estimation methods, with sensitivity rates of 60.7% using true person parameters, 57.4% estimated, and 55.1% bias-corrected. SCIP/3PL was nearly as good, with rates of 56.0%, 51.5%,

and 50.6% respectively. CMP did not perform as well, with rates of 47.9%, 39.2% and 41.2% using the NRM, and rates of 41.1%, 35.0%, and 32.1% using the 3PL.

Tables 25 and 26 give the MCAB and MCSE for shifts of mixed length when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of .00. Applying the 3PL model and the CMP algorithm increased the mean magnitude of bias for all examinees with MCAB between 0.011 and 0.014 while reducing absolute bias, MCAB being between -0.135 and -0.178 for the shifted examinees. SCIP improved slightly on those results, MCAB being between 0.010 and 0.013 for all examinees group and between -0.156 and -0.204 for the shifted examinees group. Applying the NRM decreased the magnitudes of bias across the board. When the CMP algorithm was employed, MCAB was between -0.008 and -0.011 for all examinees while SCIP resulted in an MCAB between -0.010 and -0.012. Shifted examinees improved slightly on MCAB values when the NRM was used instead of the 3PL except when CMP was applied to estimated person parameters. MCSE continued to demonstrate the trend of not tending toward the positive or negative when applied to all examinees, but showing a positive tendency in shifted examinees.

Tables 27 and 28 give the MCAB and MCSE for shifts of mixed length when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of .05. As with the medium and long shift scenarios, improvement over the zero false discovery rate is uniform across conditions while trends between those conditions are consistent with those at the more restrictive rate.

The result of introducing shift errors of mixed length into the data was an increase in mean absolute bias of 0.017. While attempts to correct shift errors using the 3PL at the false discovery rates of .00 and .05 increased the magnitude of bias, applying the NRM cut the amount of shift error-induced bias by between 47% and 76%.

4.3 Study 2: Empirical Application of Simulation Study Results

The second study was an application of the simulation results to the empirical data from which the item and person parameters were derived. Results of the study are broken down into sections by shift length, and within those sections, results from the algorithm/model matrix are reported.

4.3.1 Short Shifts

Short shifts were of length 3. Table 29 gives the counts and projected counts of shifts of length 3 or less that are detectable at the thresholds associated with each model, algorithm, and false discovery rate obtained in simulation using estimated thetas. When no false discoveries were allowed, only the CMP index using the 3PL detected any shift errors, finding 3, which projected to 60 total positives. When the false discovery rate was raised to .05, all methods detected at least some shift errors, with CMP using 3PL detecting 17, projecting to 145 total positives, CMP using NRM detecting only 2, projecting to 20 total, SCIP using 3PL detecting 11, projecting to 58 total, and SCIP using NRM detecting 7, projecting to 34.

4.3.2 Medium Shifts

Medium shifts were of length 7. Table 30 gives the counts and projected counts of shifts of length 7 or less, which was inclusive of shift errors detected when looking only for short length shifts, that were detectable at the thresholds associated with each model, algorithm, and false discovery rate obtained in simulation using estimated thetas. When no false discoveries were allowed, the CMP index using the 3PL continued to detect at a higher rate than the other methods, finding 6 shift errors, which would project to 22 total shift errors while the others detect 1 or 2 shift errors each, projecting to between 2 and 7 total shift errors. When the false discovery rate is raised to .05, again the CMP index using the 3PL model detected more shift errors at a nominally lower detection rate, leading to more extreme projections, detecting 118 shift errors which projects to 265 total. CMP using NRM detected 70, projecting to 174 total, SCIP using 3PL

detected 81, projecting to 129 total, and SCIP using NRM detected 86, which projects to 122 total.

4.3.3 Long Shifts

Long shifts were of length 10. Table 31 gives the counts and projected counts of shifts of length 10 or less detectable at the thresholds associated with each model, algorithm, and false discovery rate obtained in simulation using estimated thetas. Again, this is inclusive of the shifts of shorter length counted using those thresholds. When no false discoveries are allowed, all model/algorithm combinations yielded identical counts but because they should be capturing a higher percentage of true positives, this suggests lower projected numbers of shift errors, though only minimally lower. When the false discovery rate is raised to .05, the count obtained by CMP using 3PL is more in line with the other model/algorithm combinations, detecting 115 shift errors as compared to 101 using CMP with the NRM, 106 using SCIP with the 3PL and 108 using SCIP with the NRM, projecting to 174, 149, and 136 total positives, respectively.

4.3.4 Mixed Length Shifts

Table 32 gives the counts and projected counts of shift errors of all lengths that are detectable at the thresholds associated with each model, algorithm, and false discovery rate obtained in simulation using estimated thetas from the mixed-length scenario. When no false discoveries are allowed, results and projections are minimally different from the medium and long shift error lengths. When the false discovery rate is raised to .05, counts for all models are fairly similar for all method/model combinations, with 86 shift errors detected by CMP with the 3PL, 63 by CMP with the NRM, 79 by SCIP with the 3PL, and 88 by SCIP with the NRM. Based on true positive rates, three of the four combinations yield very similar projected counts. CMP using the 3PL continues to produce larger projections than the rest, with 233 shift errors projected while CMP using the NRM projects 153 shift errors and both models under the SCIP index project 146.

Table 33 shows agreement rates between detections under the mixed-length thresholds and shows that agreement between SCIP methods was high, having 77.2% of detected shift errors in common. CMP methods agreed on only 34.9% of shift errors. All four indices agreed on only 7.9% of the identified shift errors, largely due to differences between CMP and SCIP, which had very little crossover.

Using the thresholds for no false positives and for the false discovery rate of .05 from the mixed-length simulation, candidates classified as shift errors were corrected and person parameters were re-estimated. Mean absolute difference (MAD) and mean signed difference (MSD) were calculated for all conditions. Tables 34 and 35 show the results using the thresholds based on person parameters estimated in simulation. Results for all 3PL conditions are identical, with MAD of 0.112 and MSD of 0.000 while all NRM conditions are nearly identical with MAD of 0.000 or 0.001 and MSD of 0.000.

4.4 Study 3: Simulation Study Based on Stratified Ability Levels

In order to determine whether the proposed shift error methods function differentially dependent on examinee ability, the third study replicated the procedure from study one using smaller samples and stratified person parameters fixed at negative one, zero, and one, each then subjected to the same procedure as in the first study. Results are reported here by shift error length, showing how the different algorithm/model combinations perform for examinees at different person parameter levels. Although the study performed shift error analyses based on true, estimated, and bias-corrected person parameters, trends between ability levels were consistent independent of parameter estimation method, so the results reported herein will include only the true person parameters.

4.4.1 Short Shifts

Short shift errors were of length 3. Figures 32 to 35 show ROC curves at the three person parameter levels for each of the four scoring algorithm/probability model calculations while Figures 36 to 38 show how the algorithm/model combinations differ at each of the three levels. A

comparison of the ROC curves shows that for these short shift errors, ability level was a huge determinant factor in the effectiveness of all four of the algorithm/model combinations. When the person parameter was negative one, shift errors were undetectable via any of the combinations until false discovery rates were above .3, failed to detect even 10% of true positives until at or close to a false discovery rate of 0.9 and reached maximum detection rates of near 30% for CMP and 60% for SCIP. When the person parameter was one, on the other hand, all four algorithm/model combinations were capable of detecting some shift errors without false positives and peaked between 85% and 98%. SCIP using the 3PL model had the lowest true positive rate when false discovery rate was zero and person parameters were one. When person parameters were zero, detection rates were little better at low false discovery rates than when person parameters were negative one, NRM using CMP being the best at around 10%, but for the most part they split the difference between the higher and lower ability groups fairly evenly along the whole false discovery range.

Table 36 shows detection rates of all method combinations for the short shifts at the false discovery rate of .00 and Table 37 shows detection rates with a false discovery rate of .05. These confirm what the ROC curves tell us: when allowing no false positives, shift errors are nearly undetectable for people with low person parameters, ranging from 0.1% using CMP with the NRM up to 0.5% with SCIP using the 3PL model. For middle and high person parameters, the trend reverses, with 12.8% and 45.9% of shift errors detected at those levels using the CMP/NRM combination, 5.7% and 36.2% detected for SCIP using NRM, 6.4% and 33.9% using CMP with the 3PL, and 3.6% and 7.6% using SCIP with the 3PL. When the false discovery rate was increased to .05, person parameters of zero saw no gains in detection rate. Person parameter zero increased slightly, with a detection rate of 13.8% for CMP using NRM, 10.0% for SCIP using NRM, 7.8% for CMP using the 3PL, and 5.8% for SCIP using the 3PL. Person parameter one shift error detection rates increased by at least 10% for all models, with CMP using NRM detecting 55.7% of shift errors of length 3, SCIP using NRM detecting 52.6%, CMP using 3PL

detecting 45.2%, and SCIP using 3PL, which only detected 7.6% at false positive rate zero detecting 45.0% at false discovery rate of .05.

Tables 38 and 39 give the MCAB and MCSE for shifts of length 3 using estimated person parameters when shift errors are corrected for all model/algorithm combinations for each of the three person parameter groups at false discovery rates of .00. For examinees with a person parameter of negative one, using the 3PL resulted in increased magnitudes of bias for the all examinees group and little difference between algorithms, having MCAB of 0.059 for the all-examinees group and near 0.050 for the shifted-only examinees group. MCSE also trended toward the positive, suggesting that the increased magnitudes of bias tended toward overestimation. Using the NRM had such low detection rates that it produced few corrections and no meaningful changes in bias.

Examinees with a person parameter of zero showed decreases in the overall magnitude of bias using the 3PL with minimal differences between CMP and SCIP. For all examinees, MCAB was - 0.001 while for the shifted examinees, it was - 0.030. MCSE was 0.061 for the all examinees group and approximately 0.130 for the shifted examinees, showing that changes in bias tended to the positive direction. Using the NRM, attempts to correct these short shift errors increased MCAB, though only slightly. The all-examinees group had an MCAB of 0.001 while shifted examinees had an MCAB of 0.015 when NRM was used with CMP and 0.009 when used with SCIP. MCSE using NRM was 0.002 for all examinees and 0.031 to 0.035 for the shifted group.

Examinees with a person parameter of one showed decreased magnitude of bias (i.e., negative MCAB) across all models and algorithms but showed the greatest improvements when using the 3PL. The all-examinees group had an MCAB around -0.020 using the 3PL but less than -0.010 using the NRM. The shifted examinees saw a reduction in absolute bias of -0.137 using CMP and -0.152 using SCIP while using the 3PL. Using the NRM, CMP reduced absolute bias by -0.144, the only case where NRM performed superiorly to the 3PL, and by -0.125 using the

SCIP algorithm. MCSE tended toward the negative for the all-examinees group but the positive for the shifted examinees when using the 3PL but was positive for both groups when using the NRM.

Tables 40 and 41 give the MCAB and MCSE for shifts of length 3 when shift errors are corrected for all model/algorithm combinations using true, estimated, and bias-corrected person parameters at a false discovery rate of .05. For examinees with a person parameter of -1, results were identical to those for the false discovery rate of .00. For examinees with person parameters of zero and one, differences were very small and tended toward larger negative MCAB values, smaller negative tendencies in MCSE for the all-examinees group, and larger positive tendencies in MCSE for the shifted examinees.

4.4.2 Medium Shifts

Medium shift errors were of length 7. Figures 39 to 42 show ROC curves at the three person parameter levels for each of the four scoring-algorithm/probability-model calculations while Figures 43 to 45 show how the algorithm/model combinations differ at each of the three levels. A comparison of the ROC curves shows that, as with the short shift errors, ability level was a large determinant factor in the effectiveness of all four of the algorithm/model combinations for shifts errors of length 7. When the person parameter was negative one, true positive rates were between 0% and 10% at low false discovery rates and peaked near 60% for the CMP algorithm and 80% for the SCIP algorithm for both probability models, though only achieving those higher detection rates once false discovery rates were over .90. For the group with a person parameter of one, only CMP using the 3PL model ever dipped below a true positive rate of 90% and then only at false discovery rates at or very close to zero. For the SCIP algorithm, using NRM or 3PL, true positive rates approached but not quite reached 100% even at low false discovery rates. When person parameters were zero, detection rates were closer to the high-ability group than the low- ability group across the false discovery range for all algorithms and

probability models, nearly reaching the same rate as for the high-ability group at higher false discovery rates using SCIP.

Table 42 shows detection rates of all method combinations for the medium shifts at the false discovery rate of .00 and Table 43 shows detection rates with a false discovery rate of .05. Once again, these are confirmatory of what the ROC curves tell us: when allowing no false positives, shift errors for people with low person parameters range from 2.3% to 8.5%. For person parameter of zero, the lowest true positive rate was 47.3% for CMP using the 3PL model and the highest was 73.3% for SCIP using the NRM. At person parameter of one, the low was 86.7% and the high was 96.1%. At the false discovery rate of .05, person parameters of zero saw no gains in detection rate using CMP and only minimal gains using SCIP, from 5.0% to 5.1% using the 3PL model and from 8.5% to 10.3% with the NRM. Person parameter zero saw the largest increases using SCIP, with detection rates of 80.1% using the 3PL and 86.2% using NRM. CMP detection rates also increased at theta of zero, with detection rates of 56.1% and 64.0% for the 3PL and NRM, respectively. Person parameter one shift error detection rates, already high when no false detections were allowed, increased under all model/algorithm combinations, the lowest being 93.4% for CMP using 3PL and the highest being 98.8% for SCIP using the NRM. Regardless of person parameter level an false discovery rate, all detection rates favored SCIP over CMP and NRM over 3PL.

Tables 44 and 45 give the MCAB and MCSE for shifts of length 7 using estimated person parameters when shift errors are corrected for all model/algorithm combinations for each of the three person parameter groups at false discovery rates of .00. For examinees with a person parameter of -1, using the 3PL resulted in positive MCAB with little difference between SCIP and CMP, both algorithms having MCAB of approximately 0.060 for the all-examinees group and approximately 0.070 for the shifted-only examinees group. MCSE using the 3PL showed a trend for bias changing toward the positive for the all-examinees group with a change in MCSE of 0.032 while trending negatively for the shifted examinees with rates of -0.014 using CMP

and -0.010 using SCIP. Using the NRM on these low-ability examinees again produced low detection rates and, consequently had small impact on bias, though it did inflate it rather than reduce it. NRM using CMP showed no change in absolute bias for the all-examinees group while using SCIP had an MCAB of only 0.001. For the shifted examinees, NRM with CMP and SCIP had MCAB of 0.004 and 0.011, respectively.

Examinees with a person parameter of zero showed decreases in the overall magnitude of bias using the 3PL with small differences between CMP and SCIP. For all examinees, MCAB was between -0.005 and -0.007 for the two respective algorithms. For the shifted examinees, MCAB was -0.113 using CMP and -0.152 using SCIP. Using CMP, MCSE was 0.069 for the all-examinees group and 0.302 for the shifted examinees compared to MCSE of 0.075 and 0.409 using SCIP, showing that changes in bias tended to the positive direction, especially when focusing on the group with shift errors. The NRM performed similarly but not quite as well as the 3PL in reducing the MCAB and in its tendency to raise MCSE.

Examinees with a person parameter of one had negative MCAB across all models and algorithms, showing the greatest improvements when using the 3PL. The all-examinees group saw MCAB of around -0.050 using the 3PL and closer to -0.030 using the NRM. The shifted examinees had MCAB of -0.580 with 3PL/CMP, -0.663 using 3PL/SCIP, -0.551 with NRM/CMP, and -0.657 with NRM/SCIP. MCSE for the all-examinees group had opposite tendencies for the 3PL and NRM models but with little difference between algorithms. The 3PL had an MCSE that tended to the negative with values near -0.040 while the NRM tended toward the positive with MCSE near 0.040. For shifted examinees, MCSE had a positive tendency across conditions, ranging from 0.719 for 3PL/CMP to 0.871 for NRM/SCIP.

Tables 46 and 47 give the MCAB and MCSE for shifts of length 7 when shift errors are corrected for all model/algorithm combinations for examinees of all three ability levels at a false discovery rate of .05. These tables show that at this more permissive false discovery rate, there is a uniform trend across all conditions toward reduction in absolute bias and increase in MSE.

4.4.3 Long Shifts

Long shift errors were of length 10. Figures 46 to 49 show ROC curves at the three person parameter levels for each of the four scoring-algorithm/probability-model calculations while Figures 50 to 52 show how the algorithm/model combinations differ at each of the three levels. Once again, the ROC curves shows that for these shift errors, ability level was a great factor in the effectiveness of all four of the algorithm/model combinations. For person parameter of negative one, detection rates were better than for the short and medium shifts, though still well below the middle- and high-ability groups. For the CMP algorithm, detection rates at low false discovery rates weren't much better than for the medium-length shifts, though the peak at the other end of the false discovery range were higher. For the SCIP algorithm, detection rates were higher, starting at over 20% when combined with the NRM even at the lowest false discovery rates. When the person parameter was one, all four algorithm/model combinations started very close to 100% detection at the bottom of the false discovery range and reached 100% very quickly. When person parameters were zero, the SCIP algorithm started detection close to 90% and eventually reached nearly 100% while the CMP method did not perform as well.

Table 48 shows detection rates of all method combinations for the medium shifts at the false discovery rate of .00 and Table 49 shows detection rates with a false discovery rate of .05. These again reflect what the ROC curves tell us. When allowing no false positives, shift error detection rates for people with low person parameters were 3.3% and 3.7% for CMP using the 3PL and NRM, respectively and did not improve when false discovery rate was raised to .05. Using SCIP, rates were better, starting at 11.3% and 23.5% using 3PL and NRM, respectively, increasing to 14.1% and 33.0% when the false discovery rate was raised to .05. For person parameter of zero, CMP had detection rates of 64.3% and 71.7% using 3PL and NRM without allowing false positives and climbed to 75.3% and 82.6% at the false discovery rate of .05. SCIP performed even better, starting at 90.4% and 92.9% without allowing false positives and climbed to 96.3% and 97.7%. At the person parameter of one, CMP using 3PL and NRM, respectively,

had rates of 96.6% and 97.7% without false positives, climbing to 98.7% and 99.4% at false discovery rate of .05. SCIP using 3PL and NRM had detection rates of 99.05 and 99.2% without allowing false positives and, at the false discovery rate of .05, reached 99.8% and 99.9%.

Tables 50 and 51 give the MCAB and MCSE for shifts of length 10 using estimated person parameters when shift errors are corrected for all model/algorithm combinations for each of the three person parameter groups at false discovery rates of zero. For examinees with a person parameter of negative one, using the 3PL resulted in positive MCAB for the all-examinees group and little difference between algorithms, both algorithms having mean MCAB of approximately 0.063 for the all examinees group and 0.085 or 0.095 for the shifted-only examinees group. MCSE using the 3PL showed a trend for bias changing toward the positive for the all examinees group with a MCSE of 0.031 while MCSE for the shifted examinees trended negatively with rates of -0.035 using CMP and -0.031 using SCIP. Using the NRM on these low-ability examinees again produced low detection rates with little to no impact on bias, inflating it rather than reducing it where it was impactful. NRM using CMP showed no change in absolute bias for the all-examinees group while using SCIP had an MCAB of only 0.001. For the shifted examinees, NRM with CMP and SCIP had MCAB of 0.001 and 0.017, respectively. MCSE when using NRM trended upward as well with values of 0.001 and 0.003 for the all-examinees group using CMP and SCIP, respectively, and 0.014 and 0.048 for CMP and SCIP with the shifted examinees.

Examinees with a person parameter of zero showed decreases in the overall magnitude of bias for all model/algorithm combinations. For all examinees, 3PL/CMP produced an MCAB of -0.010, 3PL/SCIP produced an MCAB of -0.016, NRM/CMP had an MCAB of -0.008, and NRM/SCIP's MCAB was -0.016. For the shifted examinees, MCAB for 3PL/CMP was -0.185, for 3PL/SCIP was -0.291, for NRM/CMP was -0.173, and for NRM/SCIP was -0.321. Across all conditions, MCSE tended toward the positive, with values between 0.016 and 0.087, NRM/CMP being the lowest and 3PL/SCIP being the highest. For the shifted examinees, positive change in MCSE was quite large, ranging from 0.315 for NRM/CMP to 0.624 for 3PL/SCIP.

Examinees with a person parameter of one also had negative MCAB across all models and algorithms, showing the greatest improvements when using SCIP and with 3PL slightly outperforming the NRM. The all-examinees group saw an MCAB of between -0.038 for NRM/CMP and -0.065 using 3PL/SCIP. The shifted examinees had their absolute bias reduced, with MCAB of -0.767 with NRM/CMP to -0.973 using 3PL/SCIP. MCSE for the all examinees group again had opposite tendencies for the 3PL and NRM models. The 3PL had an MSE that tended to the negative with values near -0.030 while the NRM tended toward the positive with MCSE near 0.050. For shifted examinees, MCSE had a positive tendency across conditions, ranging from 0.915 for 3PL/CMP to 1.151 for NRM/SCIP.

Tables 52 and 53 give the MCAB and MCSE for shifts of length 10 when shift errors are corrected for all model/algorithm combinations for examinees of all three ability levels at a false discovery rate of .05. These tables show that at this more permissive false discovery rate, there is a uniform trend across all conditions toward reduction in absolute bias and increase in MSE.

4.4.4 Mixed Length Shifts

Mixed length shifts included an equal number of shifts of every length from 3 to 10. Figures 53 to 56 show ROC curves at the three person parameter levels for each of the four scoring-algorithm/probability-model calculations while Figures 57 to 59 show how the algorithm/model combinations differ at each of the three levels. Ability level was a huge determinant factor in the effectiveness of all four of the algorithm/model combinations. When the person parameter was negative one, shift errors were between 1% and 10% at low false discovery rates for all algorithm/model combinations with the CMP algorithm showing little separation between the NRM and 3PL until false discovery rates reached .7. The SCIP algorithm consistently outdetected the CMP and SCIP using the NRM consistently outperformed SCIP using the 3PL. Differences between models and algorithms were minor for examinees with a person parameter of one, all of them having detection rates between 75% and 85% when false discovery rates were below .05 and all of them plateauing very close to 90%. For examinees with

person parameters of zero, differences between models and algorithms were fairly evenly spaced, showing preference for the SCIP algorithm over CMP and for the NRM over the 3PL. Detection rates for person parameter zero was consistently between the detection rates for negative one and one, but was closer to the higher ability group and grew even closer at the higher end of the false discovery range.

Table 54 shows detection rates of all method combinations for the short shifts at the false discovery rate of .00 and Table 55 shows detection rates with a false discovery rate of .05. The trends here are similar to those of the medium-length shift errors. When allowing no false positives, shift error detection rates are low for examinees with person parameter negative one, ranging from 1.5% using CMP with the 3PL up to 9.0% with SCIP using the NRM model. AT false discovery rate of .05, only the SCIP/NRM combination improved and only slightly, to 10.1%. When person parameter was zero, detection rates without false positives ranged from 38.6% for CMP using the 3PL up to 56.9% for the SCIP algorithm using the NRM and increased when the false discovery rate was raised to .05, with the CMP/3PL combination detecting 45.9% and SCIP/NRM detecting 68.1%. For person parameters set to one, differences between algorithm/model combinations were smaller and the order changed slightly when no false positives were allowed, the SCIP/3PL combination being the lowest at 71.5% of shift errors detected while the SCIP/NRM combination remained the best, detecting 80.5%. At the false discovery rate of .05, detection rates all improved slightly and the more typical order was restored, with CMP/3PL having the lowest detection rate at 81.4% and SCIP using the NRM detecting 86.9% of the shift errors.

Tables 56 and 57 give the MCAB and MCSE for shifts of length 10 using estimated person parameters when shift errors are corrected for all model/algorithm combinations for each of the three person parameter groups at false discovery rates of zero. For examinees with a person parameter of negative one, using the 3PL resulted in increased magnitudes of bias for the all examinees group and little difference between algorithms, both algorithms having MCAB of

approximately 0.061 for the all-examinees group and 0.076 or 0.079 for the shifted-only examinees group. MCSE using the 3PL showed a trend for bias changing toward the positive for the all examinees group with an MCSE of 0.032 while MCSE for the shifted examinees trended negatively with rates of -0.006. Using the NRM on these low-ability examinees again produced low detection rates with little impact on bias, inflating it rather than reducing it where it was impactful. NRM/CMP showed no MCAB for the all-examinees group while using SCIP had an MCAB of only 0.001. For the shifted examinees, NRM with CMP and SCIP had MCAB of 0.004 and 0.012. MCSE when using NRM trended upward as well with values of 0.001 for the all-examinees group using both algorithms and 0.010 and 0.016 using CMP and SCIP for the shifted examinees.

Examinees with a person parameter of zero showed decreases in the overall magnitude of bias for all model/algorithm combinations. For all examinees, 3PL/CMP produced an MCAB of -0.004, 3PL/SCIP had an MCAB of -0.007, NRM/CMP's MCAB was of -0.003, and NRM/SCIP produced an MCAB of -0.006. For the shifted examinees, MCAB using 3PL/CMP was -0.092, using 3PL/SCIP was -0.128, for NRM/CMP was -0.056, and for NRM/SCIP was -0.121. Across all conditions, MCSE tended toward the positive, with values between 0.010 and 0.068, NRM/CMP being the lowest and 3PL/SCIP being the highest. For the shifted examinees, positive change in MCSE was larger, ranging from 0.194 for NRM/CMP to 0.364 for 3PL/SCIP.

Examinees with a person parameter of one also showed decreased bias magnitudes across all models and algorithms, showing the greatest improvements when using SCIP and with 3PL slightly outperforming the NRM. The all-examinees group saw an MCAB of between -0.025 for NRM/CMP and -0.045 using 3PL/SCIP. The shifted examinees had their absolute bias reduced, with an MCAB of -0.526 with NRM/CMP to -0.602 using 3PL/SCIP. MCSE for the all examinees group again had opposite tendencies for the 3PL and NRM models. The 3PL had an MSE that tended to the negative with values near -0.050 while the NRM tended toward the

positive with MCSE's near 0.035. For shifted examinees, MCSE had a positive tendency across conditions, ranging from 0.633 for 3PL/CMP to 0.765 for NRM/SCIP.

Tables 58 and 59 give the MCAB and MCSE for shifts of length 10 when shift errors are corrected for all model/algorithm combinations for examinees of all three ability levels at a false discovery rate of .05. These tables show that at this more permissive false discovery rate, there is a uniform trend across all conditions toward reduction in absolute bias and increase in MSE.

4.4.5 Other Results

Though not elaborated upon here, estimated and bias-corrected person parameters exhibited the same trends in shift error detection rates as the true person parameters and the results from those simulations are represented in Figures 60 to 115 and Tables 60 to 75.

4.5 Study 4: Comparison of Shift Error Detection Methods to the H^T Person-Fit Statistic

In order to evaluate the relative performances of SCIP, CMP, and H^T, study four compared the SCIP and CMP results from study one to results obtained by calculating true positive and false discovery rates using H^T. Figure 116 shows the ROC curves for all shift length scenarios when using H^T as a threshold for detecting examinees with shift errors. H^T performs similarly for all shift error length scenarios, with little to no detection power at false discovery rates below .7, showing some detection power between .7 and .8, then spiking sharply at .95. True positive rates were zero or nearly non-zero at the thresholds where false discovery rates are zero or .05.

CHAPTER 5

DISCUSSION

5.1 Overview

The purpose of this dissertation was to provide practical knowledge for improving upon the validity of test score interpretations through detection and classification of shift errors on paper-and-pencil tests. To that end, a small three-dimensional matrix of detection algorithms, IRT models, and person parameter techniques were developed and implemented in a series of four studies, testing permutations of these elements for their sensitivity and selectivity relative to one another and to a more traditional person-fit statistic under different simulated conditions and for their power to detect shift errors within empirical data. This section will begin with a discussion of each of the four studies, the specific questions each was designed to answer, and what information the results of each study provide in helping to answer those questions. Following that will be a summary of overarching conclusions obtained from the studies, a look at some of the limitations in the scope of the dissertations and ideas for improvement, ideas for applying and broadening the concepts explored within it, and more specific ideas for continuing with this line of research.

5.2 Study 1: Simulation Study Based on Empirical Data

The first study was designed to evaluate the accuracy and effectiveness of the proposed shift error detection methods as they might be applied under empirical conditions. By using parameter estimates from the administration of a representative paper-and-pencil K-12 proficiency exam, a realistic scenario on which to base simulation was provided and simulation results would, if methods were effective, be calibrated for application back to the empirical data set. Effectiveness of the detection techniques, measured by true positive detection rates, false discovery rates, and the change in bias when corrective measures were implemented, was dependent on all four of the investigated factors: the model, algorithm, and person parameter estimation method employed in detecting the shift errors as well as the length of the shift errors

simulated in each scenario. A look at how each of these elements impacted shift error detection will be discussed as will any interactions between methods as each is introduced into the discussion.

5.2.1 Shift Error Length

The relationship between shift error length and shift error detection was both clear and predictable: longer shift errors are easier to detect. Tables 74 and 75, representing differences between algorithms at different shift lengths based on estimated person parameters, illustrate the increasing power of all of the shift error detection methods as shift error length increases. While short shifts are detectable in only low single digit rates with any strictness in false discovery rates, over half of long shifts can be detected with the same error levels. Mixed length shifts had a mean shift length of 6.5 and its detection rates fell in just below those of shift length 7. This phenomenon, detection rates of shifts of mean length being similar to detection rates of shifts of a similar fixed length, may or may not hold depending on how shift error lengths are distributed. Tables 9, 15, and 21 show MCAB at lengths 3, 7, and 10 at false discovery rate .05. It can be seen that correcting larger shifts results in greater reductions in bias. This is in part because the larger shifts introduce more bias to be corrected but it is reassuring that the methods are capable of correcting the bias proportionally to how much is introduced by the shifts. Tables 10, 16, and 22 show MCSE at lengths 3, 7, and 10 at false discovery rate .05. When looking at MCAB and MCSE together, as shift error length increases, it can be seen that correcting the errors results in larger MCSE and increasingly negative MCAB for the shift-error group, meaning that shift error correction tends to raise scores for those examinees and, in so doing, more accurately reflect their ability.

5.2.2 Shift Detection Algorithms

Of the two shift error algorithms, the most probable correction method, which produces the SCIP index, consistently, though not exclusively, outperformed the misaligned response detection method, which produced the CMP index. Table 76 shows the differences in

performance of the two algorithms using each IRT model at all lengths using estimated person parameters and allowing no false positives. For shift errors of lengths 7 and 10, and for the mixed-length scenarios, SCIP outperformed CMP, with the gap growing as the shifts get longer. Conversely, CMP caught up to SCIP as the shift errors got shorter and, for the shift errors of length 3, CMP outperformed SCIP when no false positives were allowed. The shift error length at which this reversal is likely to happen was not determined by this study, but the differences are so small at shift error length 3, SCIP may already be superior to CMP by shift length 4 in this data set. Table 77 shows the differences in performance of the two algorithms using each IRT model at all lengths using estimated person parameters with a false discovery rate of .05. At this error rate, SCIP provided higher detection rates for all shift error lengths studied. Again, the general trend was that the difference between methods grows as shift error length grows, but shift lengths 7 and 10 were nearly identical using the NRM, even slightly favoring the shorter errors.

Figures 8 to 10, 15 to 17, 22 to 24, and 29 to 31 illustrate the differences between algorithm/model combinations across the entire range of false discovery rates. Focusing on the differences in algorithm, you can see that the difference between them is established at fairly low false discovery rates and remains nearly uniform throughout the entire false discovery rate range. Figures 8 to 10 highlight the exception to this tendency, showing the SCIP methods to be inferior to the CMP methods when false discovery rates are very low, but surpassing it fairly quickly and moving parallel to it, making another leap in true positive detections near the middle of the false discovery range, then continuing in parallel to CMP.

MCAB and MCSE showed the same tendencies in detection ability for the two detection algorithms. Tables 7 to 10, 13 to 16, and 19 to 22 show MCAB and MCSE for shift error lengths 3, 7 and 10. For short shifts, CMP has slightly better MCAB compared to SCIP but this reverses slightly for the medium and long shifts. Whether or not a high MCSE is favorable depends on whether one is moving toward the mean or away from it. Under the premise that shifted examinees have under-representative scores (i.e., negative bias) and that correcting the shift

errors should raise those scores, higher changes in MSE that coincide with reductions in absolute bias in the shifted group provide some evidence that shift error correction is doing its job. The trends displayed by MCSE were completely consistent with all of the other evidence, being higher for CMP than for SCIP when shifts are of length 3 but higher for SCIP than for CMP at shift lengths 7 and 10. While some of the gain may have been inflation beyond examinees' true ability, most of the positive MCSE coincided with negative MCAB.

5.2.3 Probability Models

Of the two IRT models employed to make probability calculations within these methods, the NRM, with minimal exceptions, outperformed the 3PL model in detecting shift errors. Table 78 shows the differences in performance of the two models used within each detection algorithm at all lengths using estimated person parameters and allowing no false positives. Differences between the models tended to be smaller for CMP than for SCIP and at shorter shift error lengths. 3PL outperformed NRM using CMP and performed almost as well as NRM using SCIP when no false positives were allowed, suggesting that the NRM may be more likely to promote false positives at these shorter lengths. Table 79 shows the differences in performance of the two models used within each detection algorithm at all lengths using estimated person parameters with a false discovery rate of .05. At this error rate, the trends of between-model differences were the same as at the no-error rate, but with more favoritism toward the NRM at all levels. Though 3PL still outperformed the NRM at short error lengths using CMP, the difference was smaller at this higher error rate.

Figures 8 to 10, 15 to 17, 22 to 24, and 29 to 31 illustrate the differences between algorithm/model combinations across the entire range of false discovery rates. Focusing on the differences between models, you can see that the differences between them were smallest at the lowest false discovery rates, separate fairly quickly, then kept a fairly uniform distance as they false discovery rates increased before moving slightly closer together at the highest false discovery rates, once all detectable true positives had been found. For the shift errors of length 3,

looking at Figures 8 to 10 reveals that it was only the estimated person parameters that were producing superior detection rates using the 3PL. True person parameters favored the NRM across the whole range of false discovery rates as did the bias-controlled estimates. The estimated person parameters only favored the 3PL at very low false discovery rates. Given that only the estimated person parameters exhibited this tendency, a possible explanation is that it is estimation bias having a more profound effect on the more precise NRM model, causing it to produce more false positives at these lower false discovery rates than the 3PL, which doesn't differentiate between wrong answers. Consider that very few false positives are necessary to produce those low false discovery rates. If a decrease in ability for one examinee provides a profound difference in the probability of one incorrect response, a very improbable false positive could result, weakening the threshold at which error-free would take place. Only a few such incidents would need to occur for this to affect other low-error detection levels. If such occurrences were consistently present but at low frequencies, this would explain the curves obtained under these circumstances.

MCAB and MCSE were consistent with the conclusion that NRM outperforms 3PL only more so. When it comes to MCAB, at no point did 3PL outperform NRM. Tables 7 to 10, 13 to 16, and 19 to 22 show MCAB and MCSE at lengths 3, 7, and 10. NRM showed consistently lower MCAB compared to the 3PL when looking at all examinees. When shifted examinees were evaluated, MCAB showed the same tendencies. MCSE had higher positive tendencies for the 3PL compared to the NRM when shifts were length 3 and at all lengths using SCIP but since this outpaced the gains in reducing absolute bias, it is unclear whether how much of this additional gain represented improvement.

5.2.4 Person Parameter Estimation Methods

Person parameters were obtained in three ways: using the estimates obtained from the empirical data and treating them as true parameters, taking the estimates obtained from calibration after shift errors were introduced into the data, and taking estimates obtained after a

second calibration that treated shift error candidates as missing. Shift error candidates were defined liberally such that only false positives at a threshold worse than the worst true positive failed to meet this classification. Differences between shift error detection rates using true person parameters and estimated provide a sense of how the bias in shift error estimation caused by the shift errors themselves impairs their detection. Differences between detection rates using estimated and bias-corrected person parameters give an indication of the effectiveness of the bias control method. Tables 80 to 83 show the differences in performance between these estimation methods for each algorithm/model combination while allowing no false positives. Independent of model and algorithm, the difference between true and estimated parameters in detection rate got worse as shift error length increased. This makes sense as a shift of 10 is going to severely impact person parameter estimate and make a string of incorrect answers more probable, whereas a shift of 3, while more difficult to detect, has less of an impact on the person parameter estimate and so using an estimate doesn't hurt the detection as much. Additionally, when detection rates are smaller overall, there are fewer detections to lose to parameter bias. Dropping from 50% detected to 40% and dropping from 10% detected to 8% due to bias would both represent losing 10% of your detections to bias – the same relative impact though one is much greater in magnitude.

Looking at detection algorithms, SCIP was more robust to estimation bias than CMP when no false positives were allowed, with estimated parameters even outperforming the true parameters for the very short shift errors. Situations in which biased estimates could improve accuracy of detection over the true person parameters seem counterintuitive, but reasonable explanations are possible. Two of the features of the SCIP algorithm are that it allows incorrect answers within the shift error candidate and that it looks at changes in probability. When ability is high, an incorrect answer is highly improbable and, thus, will lead to a smaller change in probability. Underestimating ability, then, softens the impact of that incorrect answer, leading to easier identification of shift errors, true or false, when they are short and contain an unlikely

incorrect response. If the impact of the underestimation is stronger on true positive identification than false positive identification, a result as seen here would make sense.

For probability models, 3PL was more robust to estimation bias than the NRM. This is both logical and misleading. It is logical because the 3PL does not treat incorrect responses as precisely as the NRM. The NRM is going to lose more of its power when using less accurate person parameters. On the other hand, one should not be misled into thinking this means that the 3PL was better than the NRM when using estimated parameters. While the NRM lost more power, it had more power to lose and still maintained higher detection rates when using estimated person parameters.

Turning to differences between detection rates using estimated parameters and bias-corrected person parameters, a broad analysis would be that the bias-correction did not work. When no false positives were allowed, only six of sixteen algorithm/model/shift length scenarios showed improvement under bias correction. Specifically, bias correction was effective with the SCIP algorithm only at shift length 10. Bias correction of CMP using the NRM was effective at all lengths, though least so at shift length 10. One might expect bias correction to be most effective when shift error lengths are long, and thus most impactful on the person parameter estimates, and looking at SCIP, this would appear to be the case. Looking at CMP, however, one might conclude the opposite. The reality of what is happening using CMP may be that, because it is dependent on a long string of misaligned correct answers, incorrect answers shorten the detected string, something perhaps more impactful on detectability than the associated probability, but also something of potentially great impact on the bias correction, which can only use a portion of the shift error in correcting the bias. It could also simply be that one test with one answer key pattern may yield unpredictable results and that the bias control was generally ineffective.

Tables 84 to 87 show the differences in performance between these estimation methods for each algorithm/model combination at a false discovery rate of .05. Trends in the difference

between true parameters and estimated parameters are generally the same as when no false positives are allowed, except larger. A few of the results are notable, either in confirming these trends or as exceptions. The most notable exceptions were when using SCIP with the NRM. At shift error length 3, estimates no longer outperformed true parameters. Differences between true and estimated parameters at longer shift error lengths shrunk rather than grew at the higher false discovery rate. The most notable reinforcement of the trends was the estimated person parameters outperforming the true parameters with SCIP with the 3PL at shift error length 3, detecting nearly 5% more of the shift errors using estimated instead of true persom parameters. Bias correction appeared to be no more effective in improving detection rates at the false discovery rate of .05 as it was at the rate of .00.

The only clear story that MCAB tells in regard to the person parameter estimation methods is that the estimated and bias-corrected methods were not doing as good a job as when truth was known. Figures 7 to 10, 13 to 16, and 19 to 22 show MCAB and MCSE at lengths 3, 7, and 10. Across all length, model, and algorithm combinations, at no point did bias-controlled or estimated person parameters produce favorable bias results relative to the true parameters, an unsurprising result since true person parameters ignored some of the bias created by the shift errors. Bias control did not consistently or predictably improve the bias results suggesting that it may have been inflating the bias as often as it was reducing it but close examination of the change in bias results provides no evidence of a particular trend in when it inflated and when it reduced the bias.

5.3 Study 2: Empirical Application of Simulation Study Results

The purposes of the second study were to determine, through the application of shift error methods at thresholds determined in the simulation study, the effectiveness of the shift error methods in an empirical situation and the extent to which shift errors may be present within the empirical data set on which the simulation study was based. Two different approaches were taken. In the first, the short, medium, and long shift error detection thresholds were applied

progressively, each only allowed to detect shift errors up to the length at which those thresholds were determined to function in simulation at the desired false discovery rate. In the second, the thresholds from the mixed application were applied to detect shift errors of all lengths from 3 to 10 simultaneously.

5.3.1 The Progressive Approach

The progressive approach counted shifts of specific lengths at appropriate thresholds, first using the thresholds obtained for shifts of length 3 to detect shifts of that length or shorter, then used the thresholds obtained for shifts of length 7 to detect shifts of that length or shorter, then used the thresholds obtained the thresholds obtained for shifts of length 10 to detect shifts of that length or shorter. Tables 29 to 31 give the counts and projected total shift errors in the empirical data based on the simulation thresholds and true positive rates from the simulation of the progressively longer shift error detections. When no false positives were allowed, detections of all lengths were minimal, with the SCIP methods never finding more than 1, CMP using NRM finding 2 of length 10 or shorter, and CMP detecting 3 of length 3 or shorter and 6 of length 10 or shorter. In simulation, CMP with the 3PL was shown to be superior in detection of short-length shifts when no false positives were allowed, but not for longer shift errors. Its higher detection levels at all lengths suggest that either CMP with 3PL is more prone to type I errors than was discovered in simulation, the other methods are more prone to type II errors, or both of those issues were present simultaneously. When the thresholds associated with a false discovery rate of .05 were used, again CMP using 3PL detected more shift errors than the other methods, especially relative to its percentages in simulation, which suggest it should have found fewer than the other methods. Projected total positives show good consistency between the other three methods, though the inconsistency as false discovery rate increases points to a different problem with the simulation method employed in these studies.

5.3.2 The Mixed-Length Approach

The mixed-length approach took the thresholds obtained in the mixed-length simulation scenario and applied them to detect all shifts between 3 and 10 in length simultaneously. Table 32 gives the counts and projected total shift errors in the empirical data based on the simulation thresholds and true positive rates from the mixed-length simulation. Results are similar to the progressive scenario, with CMP using the 3PL giving higher counts and projections. Again, the other methods are consistent in their projected total shift errors at the two false discovery rates. Further evidence of consistency between methods can be obtained by looking at agreement rates between the methods. Table 33 gives those agreement rates for the mixed length shift thresholds applied at the false discovery rate of .05. The two SCIP algorithms agreed on 77.2% of the shift errors they detected at that rate while the CMP algorithms only agreed on 34.9%. Agreement across algorithms was low, indicating that they are looking for and finding different things. MAD and MSD were calculated after correcting shift errors for the mixed-length approach and showed that the 3PL was tending to move examinee scores but with no direction tendency, MSD being very close to zero. Shift error correction based on NRM showed no real movement in examinee scores.

5.3.3 Lack of Simulation Applicability

As can be seen in Tables 29 to 32, the projected total positives are not consistent as the false discovery threshold is raised from zero to .05. Looking only at those thresholds, one might be able to conclude that this is a byproduct of the error-free detection level not really working on empirical data so the projections at that level cannot be trusted. Table 88 shows the simulated true positive rate and empirical detections and totals as the false discovery rate is increased from .05 to .94 for the mixed-length scenario using SCIP and the NRM. It can be seen that projected shift error totals continue to grow, approaching the number that were simulated in the first study. Whether that is coincident or systematic, the inconsistency in projections points to a need for a different approach to better understand the nature of shift errors within empirical data.

5.4 Study 3: Simulation Study Based on Stratified Ability Levels

In order to determine whether the proposed shift error methods function differentially dependent on examinee ability, the third study replicated the procedure from study one using smaller samples and stratified person parameters fixed at negative one, zero, and one, each then subjected to the same procedure as in the first study. Tables 36 to 75 illustrate the differences in detection rates and changes in bias dependent on person parameter for all algorithms, probability models, person parameter estimation methods, and shift error lengths. The results were striking, showing that the accuracy of all of the shift error methods is highly dependent on examinee ability. This holds for all shift error lengths, detection algorithms, and scoring models. For shift errors of length 10, for instance, one might expect them to be so easy to detect that the person parameter might not have a meaningful impact on detection, but the results do not bear this out. Using the best of the methods, SCIP with NRM, less than 15% of the shift errors in the low-ability group were found compared to 99.9% in the high-ability group. That is a huge disparity. For short shifts, a good number can be detected when the person parameter is one while detection is nearly useless when it is negative one. For the middle group, with person parameter zero, short shift errors were difficult to find but not altogether undetectable while longer shifts had detection rates that started to approach those for the high-ability group. An investigation of the MCAB and MCSE tables demonstrated a striking improvement in results as person parameters increase across all scenarios. They also suggested some implications on fairness, discussed in section 5.7.4.

5.5 Study 4: Comparison of Shift Error Detection Methods to the H^T Person-Fit Statistic

In order to determine the relative effectiveness of SCIP and CMP compared to H^T, this study involved calculation of H^T and using it as an index for detecting examinees who committed shift errors. Use of a person fit statistic has a couple of disadvantages independent of the results, the first being that it is not designed to target a specific form of misfit and could detect any form of misfit, not just shift errors in the data. The second disadvantage is that, even if the person-fit

statistic does an excellent job of detecting the people who have committed shift errors, it provides no mechanism for pinpointing the shift error within an examinee's response string. However, if person-fit proved more capable of detecting the individuals who committed the shift error than a shift-error, it would likely serve as the best screener at which point the nature of the misfit could be determined. Results from this study showed a complete failure of H^T in detecting examinees with shift errors. The near inability to detect any shift errors before the .95 false discovery rate, at which point detections spike to 100% is indicative of the fact that H^T is not differentiating at all between the positives and negatives. With 5% of the data shifted, a random drawing would have a false discovery rate of .95, indicating that H^T is really operating no better than chance.

5.6 Summary of Findings

Results from the four studies reveal some interesting findings that can inform operational practice and future studies into better detection of shift errors. Firstly and foremostly, these studies demonstrated that shift errors were detectable with methods designed specifically for their identification. Even with low false discovery rates, some portion of shift errors as short as three in length were detectable at rates as high as 20%. For longer shifts, detection rates as high as 75% were attainable. While some methods inflated bias, especially at short lengths, for the most part, decisions to correct shift errors resulted in decreased magnitudes of bias. The method that stood out above the others was the SCIP algorithm making use of the NRM for its probability calculations. With the exception of short shifts with no false positives allowed, it consistently provided the best rates of detection. Examinee ability was a great determinant of shift error detection. Using SCIP with the NRM, which was most robust to differences in person parameter, at the false discovery rate of .05, short shift errors were nearly undetectable in low ability examinees while reaching a detection rate of over 50% for high-ability examinees. For long shift errors, roughly one-third were detected in low ability examinees while nearly 100% were detected in high ability examinees. Application to empirical data proved tricky. While all of the methods proved capable of detecting shift-error behavior within the empirical responses, CMP using the

3PL did so at the highest rates, a result that contradicted the simulation findings. The other three methods were in much better agreement as to detection levels and in terms of agreement on which candidates were flagged as shift errors. CMP using the 3PL showed little agreement with the other methods, agreeing only 34% of the time with CMP using NRM. At the same time, SCIP using the 3PL and NRM had agreement between 77% of the candidates that they flagged for shift errors. Attempting to project total shift error counts based on the true positive percentages yielded inconsistent results, with projected counts rising as false discovery rate rose. Using H^T as a shift detection index proved completely ineffective, having a detection rate little better than would be achieved by selecting examinees at random.

5.7 Implications

The results of these studies have the potential to meaningfully inform measurement practices. This dissertation was not designed to solve the problem of shift errors on paper-and-pencil tests all at once but to explore methods that improve upon current practices, both in shift error detection and in subject areas with similar applications. This section will examine how the findings may be applied to empirical shift error detection, further simulations for improving on shift error detection methods, and person-fit research, both general and as targeted toward specific sources of aberrance.

5.7.1 Empirical Application of Shift Error Detection Methods

The methods explored in this study have some limited but important application empirically. Primarily, they are capable of discovering shift errors within empirical data. While the methods lacked the consistency one might want if one wishes to correct shift errors with certainty that false positives were not also being falsely corrected, as a flagging tool for further examination or retest, they proved capable of detection within the empirical data set. The attempt to understand how shift errors occur within the empirical data fell short. To better understand this and to increase the certainty of the results of the shift error detection, the algorithms and models themselves may need no adjustment, but the simulation that is performed for calibration of

thresholds to the empirical data needs to better reflect the shift errors within the empirical data, challenging given that the nature of such errors is unknown prior to detection.

5.7.2 Simulation Studies of Shift Error Detection Methods

Some shortcomings of the simulation study were made clear in the empirical application of the thresholds obtained through those simulations. While the simulation demonstrated the effectiveness of the shift error detection methods that were developed, it was unclear how replicable the results would be for those detection methods given different shift error lengths or levels of saturation within the data set. Given the lack of consistency in results in the empirical data , an attempt to better reflect the nature of the empirical shift errors in the simulation may be desirable. This presents a Catch-22 for shift error research, one in which simulation must reflect empirical data for its accurate application while accurate application requires that the simulation reflect the empirical data. One possible solution to this conundrum may be found by addressing the lack of consistent projected shift error counts in the empirical data. Perhaps by varying the nature of the shift errors in simulation, finding a scenario that produces thresholds that, in turn, lead to consistent projections in the empirical data, this could produce evidence that the simulated shift errors match the empirical shift errors. Even if it does not accomplish this, it would answer an important question regarding the extent to which the indices and false discovery rate thresholds are dependent on the lengths and saturation levels of the shift errors. Another approach would be to introduce a shift error into each examinee response string in simulation and do a before/after detection in order to see what detection rates are like with no shift errors and all shift errors. As the threshold is moved, true positive and false discovery rates would be determined by the counts in the shifted and unshifted data.

Other logical extensions for simulation studies would be to simulate shifts of varying distance from their correct location and simulating different shift error lengths. Especially of interest are those lengths between 3 and 7. Shift error length 3 had quite different characteristics than the longer shifts in terms of which methods were most accurate and how the bias control

measure worked. Finding the shift error length at which detection behaves the same as at the longer shift error lengths would be of interest.

5.7.3 Person Fit Research

In the fourth study, the H^T person fit statistic proved incapable of differentiating between examinees with shift errors and examinees simulated without misfit. It could be that other person fit statistics may be better suited to detection of shift errors, but as is suggested by Drasgow and Levine's (1986) optimal detection methods, modeling specific forms of person-misfit may prove a more suitable approach for any type of misfit that may be present within a test. Alternately, given that response order is inherent in the shift error problem, Armstrong and Shi's (2009) CUSUM approach, summing differences between expected and observed responses as they occur sequentially, may be more suitable than the preponderance of person fit statistics that rely on Guttman ordering rather than the actual item sequence. In some ways, the SCIP method for shift detection is a variant on the CUSUM approach, summing changes in probability and finding the largest sum within a response string in the order the test was administered. That a person-fit statistic relying on Guttman ordering proved ineffective at detecting shift errors is unsurprising, because shift errors do not exclusively occur in an area of specific item difficulty. Cheating behavior, which may largely occur on more difficult items, or creative behavior, which may lead to underperformance on easy items, would be more susceptible to detection by a Guttman-based person-fit statistic. Shift errors fall into blocks of sequential items that could be of any difficulty level, most likely a mix of difficulties. As such, failure to align to a Guttman ordering of the items may fall within normal variance and not appear as misfit.

One thing made clear in these studies is that the way in which most person-fit research is conceptualizing these indices is not adequate for empirical application. A reliance on false positive rates rather than false discovery rates says little about how useful a person-fit statistic may be in application. In the simulations within this dissertation, 95% of examinees had no shift errors introduced. At a false positive rate of .05, that would allow as many false positives as there

were actual shift errors simulated into the data. Similarly, when investigating the effectiveness of a person-fit statistic, misfit is something that, by its nature, will not affect most examinees. If all examinees are misfitting, a test can't really be measuring anything. When one expects a small minority of examinees to exhibit misfitting behavior, most of the population being made up of negatives requires a miniscule false positive rate in order to distinguish true positives from false positives. Much of the analysis of person-fit statistics, including Karabatsos' (2003) comparison of 36 person-fit statistics, based their analyses on false positive rates, obscuring the likelihood that a detection is a true positive rather than a false positive. By focusing instead on false discovery rates, one can instantly recognize the ratio of true to false positives, which would lead to clearer decision-making when applying person-fit indices. This is likely also true in many areas that require statistical analysis.

5.7.4 Fairness of Shift Error Detection and Correction

These shift error detection methods proved much more effective in accurately detecting shift errors as examinee ability increased. It is worth considering whether implementing methods that will differentially help those who need the help the least is fair. The flipside of the fairness coin would be to consider whether a procedure that improves the validity of some scores should not be performed because it will not improve the validity of other scores. It may be that despite the large difference in detection rates between high ability and low ability examinees, the fairness issue is not as extreme as it seems. Shift error detection is weakened by the existence of wrong answers within the shifted response substrings. With CMP, the shift error string is broken up and only part of it is detected. With SCIP, the probability difference of a wrong-to-wrong shift is generally smaller than a wrong-to-right, the exception being examinees of low enough ability that specific incorrect responses are more probable than correct responses.

In the case of both methods, but much more when SCIP is used, the same thing that makes the shift errors more difficult to detect – incorrect responses within the substrings – minimizes the value of being able to detect them. Consider shifts of length 10. In a high ability

examinee, if nine of ten responses were correct but because of a shift error, only three were marked as correct, the benefit of shifting to the correct location would be an improvement of six more correct items. Take a lower-ability examinee who only responded correctly to five of the ten items in the shifted substring. It could be that the shift error still results in five of the ten being correct, making a correction of minimal value. Even if the correction would be beneficial to the examinee's score, it is unlikely to be as meaningful as for the high-ability examinee.

Interestingly, an examination of mean changes in bias suggest that the 3PL may unfairly reward low-ability examinees rather than those of high-ability. When using the 3PL, at all lengths the all examinee group shows both positive changes in MCAB and in MCSE, suggesting that scores are becoming more biased and in the positive direction, artificially inflating scores for those low ability examinees. In other words, the 3PL is more likely to produce false positives in low ability examinees. Meanwhile, the 3PL tends to reduce the magnitude of bias in the high-ability examinees, a good thing, but does so by tending to reduce their scores as well, even while profoundly improving the scores of those high-ability examinees who committed shift errors. Effects on examinees who did not commit shift errors is small, whether trending positive or negative, but it does call into question the fairness of the 3PL in a minor way.

The issue of fairness or at least the differential performance of shift error detection methods that is dependent on ability presents an interesting avenue for further study. Beyond investigating the differences in detection rates as performed in study three of this dissertation, an understanding of how this affects the scores at these different ability levels and how it affects scores in empirical data near cut scores would be especially interesting.

5.7.5 Other Applications

The effectiveness of the proposed methods in detecting shift errors, especially as compared to more general person-fit statistics, suggests that type-specific person-misfit indices may be better suited to dealing with the problem of person misfit. Whether it is cheating behavior, lack of effort, or a sudden streak of spuriously low or high behavior in an adaptive situation,

modeling the expected behavior and comparing it probabilistically to the observed as was done with the SCIP algorithm could detect any form of misfit that can be posited and modeled. Additionally, some simple shift error detection may eliminate other sources of misfit. Consider cheating behavior that is detectable through erasure analysis. Large blocks of erasures resulting in wrong-to-right corrections are typically regarded as evidence of cheating behaviors. However, using shift error detection on the erasure pattern and the corrected pattern may determine that the change was not a cheat, but instead a correction of a shift error.

Shift error detection may also be applicable in detecting cheating behavior. Methods developed in this series of studies focus on finding improbable substrings and evaluating a proposed alternative in which the improbable substring is shifted. Success of the methods depend, at least in part, on recognition that the proposed alternative is more probable than leaving the substring in place. But, in some cases, it could be that correcting a substring to better align to the answer key results in a highly improbable solution given the examinee's unshifted responses. An examinee who is performing poorly, for instance, suddenly has several correct answers in a row, but in misalignment with the answer key. This could provide evidence of cheating behavior and such a response string could be flagged for comparison to neighboring examinees' test forms.

5.8 Limitations

Limitations of the studies within this dissertation have been mentioned within specific parts of the discussion, but they are worth enumerating in one place. This section will underscore the limitations of the study, their impact, and how they might be addressed in the future.

5.8.1 Empirical Data

These studies centered around only one empirical data set, simulating item and person parameters based on this data set and applying the thresholds that were obtained only to that empirical data. It stands to reason that different data sets, based on different answer sheet types, taken by examinees of different backgrounds and experience levels, with different numbers of responses and other test characteristics will contain different types and amounts of shift errors.

Whether or not the methods developed in this series of studies will prove equally effective under different conditions requires that those conditions specifically be investigated. Additionally, given the lack of consistency in projecting the true number of shift errors, this underscores the need for simulation methods that can determine and properly reflect the nature of the shifts within any given data set. Once this is done for any one data set, the validation of the methods on other, different sets of data will be critical to their viability as generally applicable methods.

5.8.2 Shift Error Lengths

Outside of the mixed-length scenario, simulated shift error lengths within the study were limited to only three lengths, 3, 7 and 10. While it is reasonable to surmise that shifts of the intermediate lengths will behave similarly to the lengths studied, the exact nature of how detection changes at each length is of interest and may prove valuable in application to empirical data. Additionally, shift error detection methods behaved differently at the shortest length in the studies, suggesting that at some point between length 3 and length 7, these methods stabilize and behave more predictably. Choosing the best methods for certain lengths will depend on understanding at exactly which lengths those methods perform at their best. Lastly, shift errors may be longer than 10 or shorter than 3 and, while shift errors of length 3 are already taxing the shift error detection methods, understanding how well they detect those shorter lengths would still be of interest and refining methods to perform better on all shift error lengths would be of interest. Presumably, longer shift errors will be easier to detect to the point that method and ability level become less important in detecting them. Understanding the nature of shift errors of all lengths and their prevalence within empirical data would be of value.

5.8.3 Bias Correction

Only one method for correcting bias caused by the shift errors was implemented. The method involved setting a permissive threshold for shift error classification and treating all candidates that met the threshold as missing data for recalibration. This method proved too permissive, and resulted in worse detection rates under many of the scenarios. Based on the

successful bias reduction of actual corrections in simulation, something that less permissively targets shift error candidates for correction might prove better suited in getting person parameter estimates closer to truth, thereby getting shift error detection rates closer to those obtainable with the true person parameters.

5.8.4 Person Fit Comparison

Only one person-fit statistic was looked at: H^T . None of the more advanced person-fit methods, such as statistically optimal measures of person fit or CUSUM, were used. Although some aspects of the SCIP index were similar to aspects of statistically optimal detection and CUSUM, there were also distinct differences in the methods. It would be interesting to see how these methods compare to the shift error methods developed in this dissertation. Additionally, though the nature of shift errors not locating in specific item difficulty areas may make traditional person-fit statistics insufficient for their detection, assuming this based on testing only one person-fit statistic is unsafe. Comparison to other person-fit statistics would serve to rule them out or determine which ones are effective in detecting shift errors.

5.8.5 Unfairness

The methods in this study all used methods that were IRT-based, which make probabilistic determinations based on examinee ability. Given that the methods clearly favored higher ability examinees who commited shift errors and the 3PL falsely rewarded low ability examinees who did not commit shift errors, it may be that IRT models are unfair for application to shift error detection. Comparison to methods that do not depend on examinee ability may prove that more fair methods are available. It seems unlikely, given that shift error detection requires some form of alignment to the answer key and that such alignment is unlikely to be recognizable without some correct responses, that even a method that doesn't depend directly on examinee ability will be doing so indirectly and any method will favor higher ability examinees in detecting their shift errors. Still, methods within this dissertation exclusively used probabilities based on examinee ability and the alternative is worth consideration.

5.8.6 Simulation Methods

As previously discussed, shift error saturation levels within the simulations were not set to reflect a known empirical situation. The result was that, while the shift error detection methods proved effective and capable of finding shift errors within both simulated and empirical data, the true positive rates in simulation did not accurately reflect the true positive rates in the empirical data, as evidenced by the fact that projected shift error totals were not consistent as false discovery rates were increased. Simulation conditions that vary saturation levels or that calculate true and false positive rates from shifted and unshifted data, respectively, may stabilize projected shift error totals and better pinpoint the nature of the shift errors within the empirical data.

5.9 Future Directions

While some aspects of these studies are, as highlighted in the implications section, extensible beyond their application to shift error detection, this section will focus on the future directions to be taken in improving upon shift error detection methods. Within this section, future studies based on ideas for building off of the current studies or that address some of the implications and limitations previously discussed will be suggested.

5.9.1 The Multiple Choice Model

The 3PL is a good dichotomous model for multiple choice items because it incorporates a guessing parameter. The NRM is a good polytomous model for multiple choice items because it provides different probabilities for the different response options without making prior assumptions as to ability levels associated with the distractors. The Multiple Choice Model (MCM; Thissen, 1989) model incorporates both of these elements, parameterizing the relative difficulties and discriminations of the different response options like the NRM while incorporating a guessing parameter like the 3PL. As such, the MCM may provide improvements over the two models employed in this study. Its implementation would require no modification to any of the detection algorithms except for the use of the MCM model for probability calculations, making it a straightforward extension of the studies contained herein.

5.9.2 Shift Error Lengths

In order to determine how shift error detection methods work in detecting shift errors of lengths not specifically investigated within these studies, simulations with shift errors greater than 10 in length, shorter than 3, and at lengths in between those covered in the studies should be performed. Due to a difference in how the methods performed with the shifts of length 3 and the longer shift lengths in the study, shift error detection rates on shifts of length 4, 5 and 6 would all be of interest.

5.9.3 Shift Error Saturation Analysis

As discussed in the study limitations and the discussion of study two, the saturation of shift errors into the simulated data was unlikely to have accurately reflected the quantity and lengths of shift errors in the empirical data set. Additionally, those are factors that are unknowable ahead of time without an arduous manual identification process that would render automatic detection unnecessary. Studies that vary the shift saturations or simulate and compare unshifted data sets and data sets with 100% of examinees committing a shift error may prove better in finding shift error rates more consistent with the empirical data and therefore would be more appropriate for detecting shift errors within the empirical data and projecting total shift error counts.

5.9.4 Shift Quantities and Distances

This series of studies assumed shift errors were limited to one per examinee and were never committed more than one item away from their correct location. Neither of these are safe assumptions, but were suitable for testing the proposed methods. It seems unlikely that shift errors of different distances from their correct locations would be handled differently by shift error detection methods, but a study that varies these distances would confirm that assumption. It may also be unlikely that examinees are committing multiple shift errors within one test form, but it is possible that this could occur or that some patterns for correcting shift errors when the examinees catch them compound the problem in an equally destructive way. Simulating multiple

shifts within an examinee could reflect some of these patterns and successful detection of them in simulation could lead to their detection in empirical data.

5.9.5 Bias Control Measures

As discussed, the bias control measure within this series of studies was generally no better than using estimated person parameters for detecting shift errors. Given that true person parameters provided the best detection rates in simulation, a bias control method that brings estimates as close to truth would be desirable. Treating shift error candidates as missing was largely ineffective in these studies because it was too permissive, treating too many false positives as true positives, omitting them in recalibration, and biasing person parameters to be higher than the true parameters. A more restrictive criterion for treating a candidate as missing could prove more effective in improving person parameter estimates and the resultant shift error detections. A study that picks several thresholds of varying degrees of permissiveness to determine if they reduce person parameter bias and improve shift error detection rates would be a logical next step along these lines.

5.9.6 Other Person Fit Measures

H^T was used as the measure of person-fit within this study because it was determined to be the most effective in detecting all types of misfit under most testing conditions (Karabatsos, 2003). It proved completely ineffective in finding shift errors within these studies. A study that analyzes other measures of person-fit to see if any are more suitable for detecting shift errors would be worthwhile, especially should one prove more effective than the shift-error-specific methods proposed within these studies.

Additionally, CUSUM and statistically optimal person-fit detection show more potential as detectors of shift error than the more traditional person-fit measures. The methods examined within this series of studies were, in some ways, hybrids of these two person-fit methodologies. Application of them specifically may prove as or more effective than the shift error methods that were studied here.

5.9.7 Fairness Analysis

The study that investigated shift error detection rates in examinees of different ability levels reached a clear conclusion that though shift error detection was far more effective on high-ability examinees than low-ability examinees, this is not inherently unfair. It may be that shift errors are far more damaging to high-ability examinees and only minimally damaging or having no overall effect on lower-ability examinees. That is to say, because shift errors are most detectable when they are moving the most right answers into wrong positions, their detectability may be proportional to the damage they cause. As such, it may be that the most damaging shift errors are found for examinees of all abilities and the methods may not discriminate based on ability except in finding shift errors that are of less concern because they contained few correct responses when located properly. On the other hand, low-ability examinees may have their shift errors go undetected purely because the probabilities associated with correct answers is not sufficiently higher than probabilities associated with incorrect answers and their scores are under-representing their ability. If shift errors are more prominent in low ability examinees and if this is affecting scores near cut points, even small under-representations of ability could be quite harmful. An extension of the current studies could include closer scrutiny of the relative benefits of shift error detection for different ability groups, looking not just at detection rates but on the gains provided by their correction, not just correcting found errors but also undetected errors to see what is lost through their non-detection.

5.9.8 Application to Other Empirical Data

Once successful refinements are implemented in the simulation studies, replication of the studies on empirical data sets with different characteristics is essential to the generalizability of the methods. Empirical data sets for examinees of different age groups, given the studies that determined younger examinees struggle more with separated forms, would be an obvious target, but how these methods perform when there are different numbers of response options, numbers of items, subject matter, answer form types, or really any variation in test administration where shift

errors can occur would be of interest. Ultimately, any operational implementation of these shift error detection methods will require calibration of thresholds to the specific operational data set. There may be some generalizability across tests, but it may very well be that any operational application will first require a simulation study based on the operational parameters.

5.10 Conclusion

This series of studies set out to determine the effectiveness of a matrix of detection algorithms, probability models, and person parameter estimation techniques in detecting shift errors, a potentially serious threat to the validity of test score interpretations. The proposed methods, particularly that which used the SCIP algorithm and the NRM, proved to be particularly effective at finding shift errors within simulated data. While concerns remain regarding differences in detection for different ability examinees and application to empirical data pointed out need for refinement in simulation methods, the initial evidence within these studies is that shift error detection with SCIP and, to a lesser extent, with CMP, is effective, whereas traditional person-fit statistics are not. Correcting shift errors detected at conservative false discovery rates resulted in large reductions in bias among those who committed shift errors and removed most of the bias created by the simulation of these shift errors. With further refinements in probability model, simulation technique, and bias control method, the methods developed and tested within this series of studies show great potential for removing a source of person misfit and improving validity of interpretations based on paper-and-pencil tests.

APPENDIX I

TABLES

Table 1: Quotes on Invalidity of Aberrant Response Patterns (Petridou & Williams, 2010)

Reise	2000	All person-fit indices are based on the premise that inconsistencies between IRT model and observed data are a sign that an individual's model derived trait score is likely invalid because factors beyond the individual's standing on the latent trait may be influencing their responses. (p. 552)
Nering & Meijer	1998	... several methods have been proposed to detect item score patterns that are not in agreement with the item score pattern expected based on a particular test model. These item score patterns should be detected because scores of such persons may not be adequate descriptions of their trait level (p. 53).
Reise & Flannery	1996	Among educational psychologists, interest in person-fit assessment and associated scalability indexes developed out of three major concerns. The first was and remains the general need to identify invalid test protocols (p. 10).
Drasgow, Levine, & Zickar	1996	Optimal appropriateness measurement statistically provides the most powerful methods for identifying individuals who are mismeasured by a standardized psychological test or scale (p. 47).
Wright	1995	... if the fit statistic for a person's performance is acceptable, then that person's test performances are interpreted as a "valid" basis for inferring a measure of that person's ability. To the extent that a person's test performances do not approximate the model, the validity of that person's ability is in doubt (p. 96).
Meijer, Molenaar, & Sijtsma	1994	For people who respond aberrantly to a test, it is questionable whether the test score is an appropriate measure of the trait that is being measured (p. 111).
Meijer & deLeeuw	1993	For persons detected as aberrant the total score does not adequately reflect the attribute that is being measured... (p. 235).
Drasgow & Guertler	1987	Item response theory provides a model-based approach to the identification of individuals for whom total test scores are not representative measures of their abilities (p. 11).
Drasgow, Levine, & Williams	1985	The test scores of some examinees on a multiple choice test may not provide satisfactory measures of their abilities. The goal of appropriateness measurement is to identify such individuals (p. 67).
Levine & Rubin	1979	A student can be so unlike other examinees that the resulting test score cannot be regarded as an appropriate ability measure (p. 269).

Table 2: Person Fit Statistics (Meijer & Sijtsma, 2001)

Group-Based (CTT)
- r_{pbis}, r_{bis} (Donlon & Fischer, 1968)
- *C* (Sato, 1975)
- *U* (van der Flier, 1980, Meijer, 1994)
- A_i, D_i, E_i (Kane & Brennan, 1980)
- *C** (Harnisch & Linn, 1981)
- *ZU3* (van der Flier, 1982)
- *NCI, ICI* (Tatsuoka & Tatsuoka, 1983)
- H_i^T (Sijtsma, 1986; Sijtsma & Meijer, 1992)

Rasch Model
- *U* (Wright & Stone, 1979)
- *W* (Wright & Masters, 1982)
- *UB, UW* (Smith, 1985)
- *M* (Molenaar & Hoijtink, 1990)
- X^2_{sc} (Klauer & Rettig, 1990)
- *T(X)* (Klauer, 1991, 1995)

2PL and 3PL
- l_0 (Levine & Rubin, 1979)
- *D(θ)* (Weiss, 1973; Trabin & Weiss, 1983)
- *ECI* statistics (Tatsuoka, 1984)
- l_z (Drasgow, Levine & Williams, 1985)
- *JK, O/E* (Drasgow, Levine & McLaughlin, 1987)
- l_{zm} (Drasgow, Levine & McLaughlin, 1991)
- *c* (Levine & Drasgow, 1988)

Computer Adaptive
- *K* (Bradlow, Weiss & Cho, 1998)
- *T* statistics(van Krimpen-Stoop & Meijer, 2000)
- Z_c (McLeod & Lewis, 1999)

Table 3: Results for Single Scan Detection of Shift Errors (Skiena & Sumazin, 2000a)

Shift Length	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5
0	.019	.000	.010	.015	.019
3	.429	.171	.209	.127	.128
4	.638	.321	.342	.258	.233
5	.743	.453	.463	.376	.353
6	.836	.611	.584	.498	.470
7	.883	.696	.674	.587	.566
8	.911	.766	.739	.702	.650
9	.929	.809	.789	.752	.713
10	.943	.842	.817	.797	.750

Table 4: Thresholds for 100% shift classification accuracy (Cook & Foster, 2012)

Length	3PL		Equal Probability	
	threshold	percent detected	threshold	percent detected
3	1E-08	1.0%	n/a	n/a
4	0.0000001	6.8%	n/a	n/a
5	0.0000001	13.0%	n/a	n/a
6	0.0000001	19.4%	n/a	n/a
7	0.0000001	24.2%	0.000001	0.1%
8	0.0000001	28.0%	0.00001	4.4%
9	0.0000001	31.0%	0.00001	16.1%
10	0.000001	40.7%	0.00001	18.0%
mixed	1E-08	15.3%	0.0000001	0.1%

Table 5: Shift error detection rates and thresholds, false discovery rate = .00, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
True	7.8%	$1.0*10^{-8}$	1.6%	3.82	10.0%	$1.3*10^{-6}$	2.2%	3.37
Estimated	5.0%	$1.6*10^{-7}$	2.1%	3.26	2.4%	$2.8*10^{-6}$	2.3%	3.03
Corrected	4.4%	$3.3*10^{-11}$	0.1%	4.44	7.9%	$4.6*10^{-8}$	0.2%	4.00

Table 6: Shift error detection rates and thresholds, false discovery rate = .05, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
True	15.8%	$5.2*10^{-7}$	13.5%	2.93	19.8%	$2.9*10^{-5}$	21.9%	2.50
Estimated	11.1%	$2.4*10^{-6}$	18.1%	2.60	9.5%	$2.6*10^{-5}$	19.6%	2.35
Corrected	8.3%	$1.6*10^{-9}$	0.2%	4.20	16.7%	$1.8*10^{-6}$	1.0%	3.66

Table 7: Mean Change in Absolute Bias, false discovery rate = .00, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.020	-0.041	0.021	-0.017	-0.003	-0.048	-0.001	-0.010
Estimated	0.021	-0.021	0.022	-0.014	-0.001	-0.010	0.000	-0.007
Corrected	0.021	-0.020	0.022	-0.009	-0.002	-0.028	0.000	0.000

Table 8: Mean Change in Signed Error, false discovery rate = .00, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.000	0.083	0.000	0.054	0.000	0.067	0.000	0.016
Estimated	0.000	0.069	0.000	0.053	0.000	0.021	0.000	0.015
Corrected	0.000	0.067	0.000	0.041	0.000	0.058	0.000	0.000

Table 9: Mean Change in Absolute Bias, false discovery rate = .05, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.019	-0.052	0.020	-0.041	-0.004	-0.063	-0.003	-0.043
Estimated	0.020	-0.028	0.020	-0.029	-0.002	-0.026	-0.002	-0.029
Corrected	0.020	-0.025	0.022	-0.009	-0.003	-0.041	0.000	-0.002

Table 10: Mean Change in Signed Error, false discovery rate = .05, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.000	0.103	0.000	0.100	0.000	0.098	0.000	0.074
Estimated	0.000	0.087	0.000	0.091	0.000	0.050	0.000	0.068
Corrected	0.000	0.079	0.000	0.042	0.000	0.088	0.000	0.005

Table 11: Shift error detection rates and thresholds, false discovery rate = .00, shift length 7

	3PL				NRM			
Person	CMP		SCIP		CMP		SCIP	
Parameter	percent	threshold	percent	threshold	percent	threshold	percent	threshold
True	34.2%	$9.4*10^{-9}$	43.2%	3.82	39.0%	$1.3*10^{-6}$	49.2%	3.36
Estimated	26.7%	$2.1*10^{-7}$	40.0%	3.28	28.6%	$3.4*10^{-6}$	44.6%	3.04
Corrected	25.0%	$3.6*10^{-11}$	33.4%	4.33	33.1%	$6.7*10^{-8}$	37.9%	3.93

Table 12: Shift error detection rates and thresholds, false discovery rate = .05, shift length 7

	3PL				NRM			
Person	CMP		SCIP		CMP		SCIP	
Parameter	percent	threshold	percent	threshold	percent	threshold	percent	threshold
True	48.5%	$2.7*10^{-6}$	64.7%	2.53	54.4%	$7.8*10^{-5}$	70.0%	2.23
Estimated	42.3%	$1.4*10^{-5}$	59.7%	2.33	45.8%	$1.1*10^{-4}$	66.7%	2.10
Corrected	38.2%	$4.3*10^{-8}$	60.5%	2.82	49.3%	$9.1*10^{-6}$	65.8%	2.55

Table 13: Mean Change in Absolute Bias, false discovery rate = .00, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.010	-0.214	0.008	-0.249	-0.013	-0.213	-0.015	-0.255
Estimated	0.012	-0.172	0.010	-0.210	-0.010	-0.164	-0.014	-0.225
Corrected	0.012	-0.166	0.011	-0.197	-0.011	-0.181	-0.012	-0.205

Table 14: Mean Change in Signed Error, false discovery rate = .00, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.000	0.311	0.000	0.378	0.000	0.318	0.000	0.372
Estimated	0.000	0.273	0.000	0.333	0.000	0.266	0.000	0.348
Corrected	0.000	0.262	0.000	0.304	0.000	0.292	0.000	0.315

Table 15: Mean Change in Absolute Bias, false discovery rate = .05, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.008	-0.237	0.007	-0.261	-0.014	-0.240	-0.016	-0.273
Estimated	0.010	-0.202	0.009	-0.235	-0.012	-0.202	-0.015	-0.253
Corrected	0.011	-0.195	0.009	-0.232	-0.013	-0.215	-0.014	-0.243

Table 16: Mean Change in Signed Error, false discovery rate = .05, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.000	0.353	0.000	0.421	0.000	0.359	0.000	0.413
Estimated	0.000	0.321	0.000	0.384	0.000	0.317	0.000	0.397
Corrected	0.000	0.308	0.000	0.374	0.000	0.336	0.000	0.381

Table 17: Shift detection rates and thresholds, false discovery rate = .00, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
True	42.9%	$9.0*10^{-9}$	60.2%	3.81	49.3%	$1.4*10^{-6}$	65.7%	3.34
Estimated	31.2%	$1.9*10^{-7}$	51.6%	3.31	36.5%	$4.2*10^{-6}$	58.2%	3.08
Corrected	30.6%	$5.8*10^{-11}$	53.6%	4.33	38.9%	$7.7*10^{-8}$	58.3%	3.95

Table 8: Shift detection rates and thresholds, false discovery rate = .05, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
True	57.8%	$3.5*10^{-6}$	74.9%	2.48	64.3%	$9.2*10^{-5}$	79.6%	2.19
Estimated	47.2%	$1.3*10^{-5}$	67.8%	2.29	55.0%	$1.4*10^{-4}$	75.7%	2.08
Corrected	43.7%	$7.0*10^{-8}$	72.1%	2.78	55.6%	$1.2*10^{-5}$	77.1%	2.52

Table 19: Mean Change in Absolute Bias, false discovery rate = .00, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.005	-0.300	0.000	-0.392	-0.018	-0.313	-0.023	-0.410
Estimated	0.008	-0.239	0.003	-0.337	-0.015	-0.256	-0.022	-0.376
Corrected	0.008	-0.236	0.002	-0.348	-0.016	-0.266	-0.022	-0.378

Table 20: Mean Change in Signed Error, false discovery rate = .00, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.000	0.393	0.000	0.555	0.000	0.407	0.000	0.558
Estimated	0.000	0.331	0.000	0.477	0.000	0.350	0.000	0.521
Corrected	0.000	0.325	0.000	0.495	0.000	0.362	0.000	0.523

Table 21: Mean Change in Absolute Bias, false discovery rate = .05, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.003	-0.330	-0.001	-0.406	-0.019	-0.339	-0.024	-0.425
Estimated	0.006	-0.279	0.001	-0.366	-0.017	-0.298	-0.023	-0.401
Corrected	0.006	-0.271	0.001	-0.377	-0.017	-0.302	-0.023	-0.404

Table 22: Mean Change in Signed Error, false discovery rate = .05, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.000	0.435	0.000	0.590	0.000	0.441	0.000	0.587
Estimated	0.000	0.383	0.000	0.527	0.000	0.399	0.000	0.562
Corrected	0.000	0.371	0.000	0.548	0.000	0.405	0.000	0.567

Table 23: Shift detection rates and thresholds, false discovery rate = .00, mixed length shifts

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
True	29.1%	$1.1*10^{-8}$	35.2%	3.78	33.6%	$1.5*10^{-6}$	39.9%	3.32
Estimated	21.4%	$1.8*10^{-7}$	32.3%	3.27	22.3%	$2.9*10^{-6}$	35.7%	3.03
Corrected	21.1%	$6.9*10^{-11}$	26.7%	4.32	27.1%	$5.5*10^{-8}$	30.0%	3.93

Table 24: Shift detection rates and thresholds, false discovery rate = .05, mixed length shifts

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
True	41.1%	$1.5*10^{-6}$	56.0%	2.57	47.9%	$6.8*10^{-5}$	60.7%	2.26
Estimated	35.0%	$9.3*10^{-6}$	51.5%	2.37	39.2%	$9.6*10^{-5}$	57.4%	2.14
Corrected	32.1%	$3.2*10^{-8}$	50.6%	2.88	41.2%	$7.4*10^{-6}$	55.1%	2.62

Table 25: Mean Change in Absolute Bias, false discovery rate = .00, mixed length shifts

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.011	-0.178	0.010	-0.204	-0.011	-0.184	-0.012	-0.209
Estimated	0.014	-0.135	0.012	-0.172	-0.008	-0.129	-0.011	-0.185
Corrected	0.013	-0.135	0.013	-0.156	-0.009	-0.148	-0.010	-0.165

Table 26: Mean Change in Signed Error, false discovery rate = .00, mixed length shifts

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.000	0.265	0.000	0.323	0.000	0.269	0.000	0.310
Estimated	0.000	0.224	0.000	0.285	0.000	0.205	0.000	0.289
Corrected	0.000	0.222	0.000	0.256	0.000	0.239	0.000	0.255

Table 27: Mean Change in Absolute Bias, false discovery rate = .05, mixed length shifts

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.010	-0.197	0.009	-0.218	-0.012	-0.205	-0.013	-0.231
Estimated	0.012	-0.161	0.011	-0.195	-0.010	-0.167	-0.012	-0.213
Corrected	0.013	-0.157	0.011	-0.190	-0.010	-0.178	-0.012	-0.203

Table 28: Mean Change in Signed Error, false discovery rate = .05, mixed length shifts

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
True	0.000	0.302	0.000	0.368	0.000	0.307	0.000	0.358
Estimated	0.000	0.269	0.000	0.335	0.000	0.263	0.000	0.343
Corrected	0.000	0.260	0.000	0.324	0.000	0.285	0.000	0.324

Table 29: Empirical true positives at simulation thresholds, shift lengths of 3 or less

Model/FDR	CMP				SCIP			
	threshold	percent	count	projected	threshold	percent	count	projected
3PL/.00	$1.6*10^{-7}$	5.0%	3	60	3.26	2.1%	0	0
3PL/.05	$2.4*10^{-6}$	11.1%	17	145	2.60	18.1%	11	58
NRM/.00	$2.8*10^{-6}$	2.4%	0	0	3.03	2.3%	0	0
NRM/.05	$2.6*10^{-5}$	9.5%	2	20	2.35	19.6%	7	34

Table 30: Empirical true positives at simulation thresholds, shift lengths of 7 or less

Model/FDR	CMP				SCIP			
	threshold	percent	count	projected	threshold	percent	count	projected
3PL/.00	$2.1*10^{-7}$	26.7%	6	22	3.28	40.0%	1	3
3PL/.05	$1.4*10^{-5}$	42.3%	118	265	2.33	59.7%	81	129
NRM/.00	$3.4*10^{-6}$	28.6%	2	7	3.04	44.6%	1	2
NRM/.05	$1.1*10^{-4}$	45.8%	70	145	2.10	66.7%	86	122

Table 31: Empirical true positives at simulation thresholds, shift lengths of 10 or less

Model/FDR	CMP				SCIP			
	threshold	percent	count	projected	threshold	percent	count	projected
3PL/.00	$1.9*10^{-7}$	31.2%	6	19	3.31	51.6%	1	2
3PL/.05	$1.3*10^{-5}$	47.2%	115	231	2.29	67.8%	106	149
NRM/.00	$4.2*10^{-6}$	36.5%	2	5	3.08	58.2%	1	2
NRM/.05	$1.4*10^{-4}$	55.0%	101	174	2.08	75.7%	108	136

Table 32: Empirical true positives at simulation thresholds, mixed shift lengths

Model/FDR	CMP				SCIP			
	threshold	percent	count	projected	threshold	percent	count	projected
3PL/.00	$1.8*10^{-7}$	21.4%	6	28	3.27	32.3%	1	3
3PL/.05	$9.3*10^{-6}$	35.0%	86	233	2.37	51.5%	79	146
NRM/.00	$2.9*10^{-6}$	22.3%	1	4	3.03	35.7%	1	3
NRM/.05	$9.6*10^{-5}$	39.2%	63	153	2.14	57.4%	88	146

Table 33: Agreement rates between methods, false discovery rate = .05, mixed length shifts

Model/Algorithm Combination	count	percent
SCIP	61	77.2%
CMP	22	34.9%
NRM	13	16.5%
3PL	9	14.3%
All	5	7.9%
Any 3	14	22.2%

Table 34: Mean Absolute Difference, empirical data, mixed length shifts

False Discovery Rate	3PL		NRM	
	CMP	SCIP	CMP	SCIP
0%	0.112	0.112	0.000	0.000
5%	0.112	0.112	0.001	0.000

Table 35: Mean Signed Difference, empirical data, mixed length shifts

False Discovery Rate	3PL		NRM	
	CMP	SCIP	CMP	SCIP
0%	0.000	0.000	0.000	0.000
5%	0.000	0.000	0.000	0.000

Table 36: Shift error detection rates with true person parameters, FDR = .00, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.2%	$1.7*10^{-5}$	0.5%	2.63	0.1%	$2.6*10^{-5}$	0.4%	2.55
0	6.4%	$2.2*10^{-6}$	3.6%	2.98	12.8%	$2.3*10^{-5}$	5.7%	2.81
1	33.9%	$2.6*10^{-7}$	7.6%	3.10	45.9%	$3.4*10^{-5}$	36.2%	2.67

Table 37: Shift error detection rates with true person parameters, FDR = .05, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.2%	$1.7*10^{-5}$	0.5%	2.63	0.1%	$2.6*10^{-5}$	0.4%	2.55
0	7.8%	$3.0*10^{-6}$	5.8%	2.88	13.8%	$3.3*10^{-5}$	10.0%	2.67
1	45.2%	$1.2*10^{-6}$	45.0%	2.77	55.7%	$1.2*10^{-4}$	52.6%	2.28

Table 38: Mean Change in Absolute Bias, estimated parameters, FDR = .00, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.059	0.052	0.059	0.054	0.000	0.000	0.000	0.000
0	-0.001	-0.029	-0.001	-0.026	0.001	0.015	0.001	0.009
1	-0.022	-0.137	-0.023	-0.152	-0.007	-0.144	-0.006	-0.125

Table 39: Mean Change in Signed Error, estimated parameters, FDR = .00, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.033	0.018	0.033	0.020	0.000	0.000	0.000	0.000
0	0.061	0.126	0.061	0.128	0.002	0.035	0.002	0.031
1	-0.072	0.212	-0.072	0.225	0.015	0.295	0.012	0.244

Table 40: Mean Change in Absolute Bias, estimated parameters, FDR =.05, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.059	0.052	0.059	0.054	0.000	0.000	0.001	0.000
0	-0.005	-0.029	-0.001	-0.025	0.001	0.016	0.001	0.012
1	-0.023	-0.154	-0.025	-0.180	-0.008	-0.156	-0.008	-0.152

Table 41: Mean Change in Signed Error, estimated parameters, FDR =.05, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.033	0.018	0.033	0.020	0.000	0.000	0.001	0.000
0	0.061	0.126	0.061	0.131	0.002	0.037	0.002	0.044
1	-0.071	0.239	-0.070	0.254	0.017	0.325	0.014	0.281

Table 42: Shift error detection rates with true person parameters, FDR = .00, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	2.3%	$6.0*10^{-6}$	5.0%	2.79	2.4%	$1.5*10^{-5}$	8.5%	2.68
0	47.3%	$2.5*10^{-6}$	66.0%	2.97	54.9%	$2.5*10^{-5}$	73.3%	2.78
1	86.7%	$1.7*10^{-7}$	93.1%	3.14	90.3%	$2.8*10^{-5}$	96.1%	2.77

Table 43: Shift error detection rates with true person parameters, FDR = .05, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	2.3%	$6.0*10^{-6}$	5.1%	2.77	2.4%	$1.5*10^{-5}$	10.3%	2.60
0	56.1%	$1.1*10^{-5}$	80.1%	2.49	64.0%	$1.0*10^{-4}$	86.2%	2.31
1	93.4%	$3.0*10^{-6}$	98.1%	2.46	95.4%	$2.0*10^{-4}$	98.8%	2.11

Table 44: Mean Change in Absolute Bias, estimated parameters, FDR =.00, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.060	0.071	0.061	0.074	0.000	0.004	0.001	0.011
0	-0.005	-0.113	-0.007	-0.152	-0.004	-0.090	-0.007	-0.148
1	-0.045	-0.580	-0.050	-0.663	-0.027	-0.551	-0.033	-0.657

Table 45: Mean Change in Signed Error, estimated parameters, FDR =.00, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.032	-0.014	0.032	-0.010	0.001	0.010	0.001	0.018
0	0.069	0.302	0.075	0.409	0.013	0.261	0.019	0.379
1	-0.044	0.719	-0.040	0.802	0.038	0.759	0.043	0.871

Table 46: Mean Change in Absolute Bias, estimated parameters, FDR =.05, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.060	0.072	0.061	0.074	0.000	0.004	0.001	0.011
0	-0.005	-0.118	-0.008	-0.166	-0.005	-0.107	-0.008	-0.162
1	-0.047	-0.604	-0.051	-0.676	-0.029	-0.572	-0.033	-0.660

Table 47: Mean Change in Signed Error, estimated parameters, FDR =.05, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.032	-0.014	0.032	-0.010	0.000	0.010	0.001	0.019
0	0.070	0.320	0.077	0.445	0.015	0.288	0.021	0.408
1	-0.043	0.743	-0.039	0.814	0.040	0.782	0.045	0.874

Table 48: Shift detection rates with true person parameters, FDR = .00, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	3.3%	$4.2*10^{-6}$	11.3%	2.79	3.7%	$1.2*10^{-5}$	23.5%	2.62
0	64.3%	$2.8*10^{-6}$	90.4%	2.97	71.7%	$2.4*10^{-5}$	92.9%	2.83
1	96.6%	$1.4*10^{-7}$	99.0%	2.11	97.7%	$1.7*10^{-5}$	99.2%	3.34

Table 49: Shift detection rates with true person parameters, FDR = .05, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	3.3%	$4.2*10^{-6}$	14.1%	2.62	3.7%	$1.2*10^{-5}$	33.0%	2.38
0	75.3%	$1.6*10^{-5}$	96.3%	2.42	82.6%	$1.3*10^{-4}$	97.7%	2.24
1	98.7%	$3.7*10^{-6}$	99.8%	2.47	99.4%	$2.0*10^{-4}$	99.9%	2.12

Table 50: Mean Change in Absolute Bias, estimated parameters, FDR =.00, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.063	0.085	0.063	0.095	0.000	0.001	0.001	0.017
0	-0.010	-0.185	-0.016	-0.291	-0.008	-0.173	-0.016	-0.321
1	-0.056	-0.789	-0.065	-0.973	-0.038	-0.767	-0.048	-0.961

Table 51: Mean Change in Signed Error, estimated parameters, FDR =.00, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.031	-0.035	0.031	-0.031	0.001	0.014	0.003	0.048
0	0.073	0.353	0.087	0.624	0.016	0.315	0.031	0.609
1	-0.036	0.915	-0.028	1.085	0.048	0.954	0.058	1.151

Table 52: Mean Change in Absolute Bias, estimated parameters, FDR =.05, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.063	0.085	0.063	0.095	0.000	0.001	0.001	0.015
0	-0.011	-0.202	-0.017	-0.319	-0.010	-0.202	-0.016	-0.338
1	-0.058	-0.818	-0.066	-0.977	-0.039	-0.774	-0.048	-0.963

Table 53: Mean Change in Signed Error, estimated parameters, FDR =.05, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.031	-0.035	0.031	-0.031	0.001	0.014	0.003	0.059
0	0.075	0.377	0.089	0.658	0.018	0.001	0.033	0.629
1	-0.034	0.943	-0.027	1.089	0.049	0.961	0.058	1.152

Table 9: Shift detection rates with true person parameters, FDR = .00, mixed-length shifts

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	1.5%	$4.9*10^{-6}$	4.6%	2.81	1.9%	$1.4*10^{-5}$	9.0%	2.65
0	38.6%	$2.4*10^{-6}$	51.2%	2.99	46.1%	$2.5*10^{-5}$	56.9%	2.77
1	74.4%	$2.3*10^{-7}$	71.5%	3.10	78.1%	$2.8*10^{-5}$	80.5%	2.69

Table 55: Shift detection rates with true person parameters, FDR = .05, mixed-length shifts

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	1.5%	$4.9*10^{-6}$	4.6%	2.81	1.9%	$1.4*10^{-5}$	10.1%	2.60
0	45.9%	$8.8*10^{-6}$	63.9%	2.56	52.6%	$8.4*10^{-5}$	68.1%	2.36
1	81.4%	$2.8*10^{-6}$	85.8%	2.50	84.3%	$1.7*10^{-4}$	86.9%	2.14

Table 56: Mean Change in Absolute Bias, estimated parameters, FDR =.00, mixed lengths

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.061	0.076	0.061	0.079	0.000	0.004	0.001	0.012
0	-0.004	-0.092	-0.007	-0.128	-0.003	-0.056	-0.006	-0.121
1	-0.041	-0.527	-0.045	-0.602	-0.025	-0.526	-0.028	-0.586

Table 57: Mean Change in Signed Error, estimated parameters, FDR =.00, mixed lengths

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.032	-0.006	0.032	-0.006	0.001	0.010	0.001	0.016
0	0.068	0.264	0.072	0.364	0.010	0.194	0.016	0.325
1	-0.052	0.633	-0.048	0.714	0.033	0.693	0.037	0.765

Table 58: Mean Change in Absolute Bias, estimated parameters, FDR =.05, mixed lengths

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.061	0.076	0.061	0.079	0.000	0.004	0.001	0.012
0	-0.005	-0.097	-0.007	-0.141	-0.003	-0.068	-0.006	-0.135
1	-0.042	-0.553	-0.046	-0.618	-0.026	-0.550	-0.029	-0.600

Table 59: Mean Change in Signed Error, estimated parameters, FDR =.05, mixed lengths

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	All	Shifted	All	Shifted	All	Shifted	All	Shifted
-1	0.032	-0.006	0.032	-0.006	0.001	0.010	0.001	0.016
0	0.068	0.275	0.074	0.399	0.011	0.217	0.018	0.356
1	-0.050	0.658	-0.047	0.731	0.036	0.719	0.038	0.781

Table 60: Shift detection rates with estimated person parameters, FDR = .00, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.1%	$4.6*10^{-6}$	0.3%	2.63	0.1%	$1.9*10^{-5}$	0.4%	2.52
0	5.9%	$1.8*10^{-6}$	7.9%	2.81	7.8%	$4.1*10^{-5}$	7.4%	2.59
1	27.3%	$2.1*10^{-6}$	37.4%	2.61	41.7%	$1.2*10^{-4}$	41.3%	2.33

Table 61: Shift detection rates with estimated person parameters, FDR = .05, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.1%	$4.6*10^{-6}$	0.3%	2.63	0.1%	$1.9*10^{-5}$	0.4%	2.52
0	6.0%	$1.9*10^{-6}$	9.6%	2.77	8.3%	$4.4*10^{-5}$	12.7%	2.46
1	36.1%	$6.9*10^{-6}$	49.3%	2.25	50.6%	$3.2*10^{-4}$	53.8%	2.01

Table 62: Shift detection rates with estimated person parameters, FDR = .00, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.9%	$5.1*10^{-6}$	2.5%	2.64	1.9%	$1.6*10^{-5}$	5.6%	2.56
0	30.4%	$2.0*10^{-6}$	53.5%	2.86	42.0%	$3.9*10^{-5}$	64.6%	2.61
1	74.3%	$1.8*10^{-6}$	93.1%	2.67	84.6%	$1.0*10^{-4}$	95.5%	2.42

Table 63: Shift detection rates with estimated person parameters, FDR = .05, shift length 7.

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.9%	$5.1*10^{-6}$	2.5%	2.64	1.9%	$1.6*10^{-5}$	5.8%	2.55
0	37.4%	$6.0*10^{-6}$	69.6%	2.44	51.2%	$1.2*10^{-4}$	80.9%	2.17
1	82.5%	$1.4*10^{-5}$	97.7%	2.05	90.5%	$4.5*10^{-4}$	98.6%	1.88

Table 64: Shift detection rates, estimated person parameters, FDR = .00, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	1.6%	$4.9*10^{-6}$	3.2%	2.65	2.7%	$1.4*10^{-5}$	10.9%	2.54
0	37.2%	$2.1*10^{-6}$	74.3%	2.89	54.5%	$4.3*10^{-5}$	86.5%	2.65
1	83.5%	$2.4*10^{-6}$	98.4%	2.90	93.7%	$9.8*10^{-5}$	98.8%	2.77

Table 65: Shift detection rates, estimated person parameters, FDR = .05, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	1.6%	$4.9*10^{-6}$	3.2%	2.65	2.7%	$1.4*10^{-5}$	14.2%	2.43
0	45.0%	$7.7*10^{-6}$	88.6%	2.35	66.6%	$1.6*10^{-4}$	95.3%	2.12
1	91.4%	$1.5*10^{-5}$	99.8%	2.06	97.2%	$5.0*10^{-4}$	99.9%	1.88

Table 66: Shift detection rates with estimated person parameters, FDR = .00, mixed-lengths

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.8%	$3.9*10^{-6}$	1.7%	2.72	1.3%	$1.6*10^{-5}$	4.7%	2.59
0	26.4%	$5.8*10^{-6}$	43.8%	2.84	34.7%	$3.9*10^{-5}$	52.3%	2.58
1	62.8%	$2.2*10^{-6}$	78.1%	2.64	72.9%	$9.6*10^{-5}$	80.6%	2.36

Table 67: Shift detection rates with estimated person parameters, FDR = .05, mixed-lengths

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.8%	$3.9*10^{-6}$	1.7%	2.72	1.3%	$1.6*10^{-5}$	4.7%	2.59
0	31.1%	$5.8*10^{-6}$	56.0%	2.50	42.8%	$1.1*10^{-4}$	64.5%	2.24
1	71.0%	$4.1*10^{-5}$	85.1%	2.06	80.1%	$4.5*10^{-5}$	87.4%	1.89

Table 68: Shift detection rates with bias-corrected parameters, FDR = .00, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.2%	$4*10^{-7}$	0.3%	3.27	0.1%	$1.2*10^{-5}$	0.1%	5.03
0	7.0%	$2.6*10^{-7}$	2.6%	3.48	9.7%	$9.6*10^{-6}$	1.8%	3.70
1	26.4%	$1.2*10^{-7}$	7.2%	3.07	42.2%	$2.6*10^{-5}$	32.5%	2.76

Table 69: Shift detection rates with bias-corrected parameters, FDR = .05, shift length 3

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	0.2%	$4*10^{-7}$	0.3%	3.27	0.1%	$1.2*10^{-5}$	0.1%	5.03
0	7.0%	$2.7*10^{-7}$	2.6%	3.48	11.1%	$1.3*10^{-5}$	1.8%	3.70
1	33.9%	$5.1*10^{-7}$	45.1%	2.69	51.5%	$9.8*10^{-5}$	51.4%	2.36

Table 70: Shift detection rates with bias-corrected parameters, FDR = .00, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	1.3%	$2.7*10^{-7}$	2.7%	3.58	2.3%	$6.8*10^{-6}$	3.0%	5.05
0	34.7%	$2.3*10^{-7}$	53.0%	3.52	43.3%	$8.8*10^{-6}$	62.3%	3.65
1	75.7%	$9.2*10^{-8}$	92.2%	3.10	84.5%	$2.0*10^{-5}$	95.3%	2.86

Table 71: Shift detection rates with bias-corrected parameters, FDR = .05, shift length 7

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	1.3%	$2.7*10^{-7}$	2.7%	3.58	2.3%	$6.8*10^{-6}$	3.0%	5.05
0	42.4%	$1.1*10^{-6}$	71.1%	2.87	53.4%	$1.0*10^{-4}$	77.1%	3.03
1	83.3%	$1.4*10^{-6}$	97.9%	2.29	90.3%	$1.5*10^{-4}$	98.7%	2.14

Table 72: Shift detection rates with bias-corrected parameters, FDR = .00, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	2.1%	$2.8*10^{-7}$	7.7%	3.50	3.4%	$6.9*10^{-6}$	10.1%	5.03
0	43.6%	$3.0*10^{-7}$	79.9%	3.60	55.5%	$1.1*10^{-5}$	87.4%	3.69
1	85.1%	$1.5*10^{-7}$	98.5%	3.20	93.7%	$2.3*10^{-5}$	99.2%	3.35

Table 73: Shift detection rates with bias-corrected parameters, FDR = .05, shift length 10

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	2.1%	$2.8*10^{-7}$	8.0%	3.46	3.4%	$6.9*10^{-6}$	12.6%	4.84
0	50.5%	$1.3*10^{-6}$	91.6%	2.81	67.0%	$5.5*10^{-5}$	94.7%	2.97
1	92.0%	$1.5*10^{-6}$	99.8%	2.11	96.9%	$1.6*10^{-4}$	99.9%	2.10

Table 74: Shift detection rates with bias-corrected parameters, FDR = .00, mixed-lengths

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	1.2%	$2.4*10^{-7}$	2.5%	3.63	1.8%	$6.5*10^{-6}$	3.3%	5.05
0	30.2%	$3.4*10^{-7}$	43.0%	3.48	35.8%	$8.9*10^{-6}$	48.2%	3.64
1	64.5%	$1.3*10^{-7}$	71.0%	3.04	73.4%	$2.2*10^{-5}$	79.5%	2.76

Table 75: Shift detection rates with bias-corrected parameters, FDR = .05, mixed-lengths

Person Parameter	3PL				NRM			
	CMP		SCIP		CMP		SCIP	
	percent	threshold	percent	threshold	percent	threshold	percent	threshold
-1	1.2%	$2.4*10^{-7}$	2.5%	3.63	1.8%	$6.5*10^{-6}$	3.3%	5.05
0	34.1%	$9.4*10^{-7}$	53.8%	2.96	44.0%	$3.4*10^{-5}$	58.2%	3.12
1	71.9%	$1.2*10^{-6}$	84.9%	2.36	80.5%	$1.5*10^{-5}$	86.5%	2.19

Table 76: Detection rate differences between algorithms, estimated parameters, FDR = .00

Shift Length	3PL			NRM		
	CMP	SCIP	Difference	CMP	SCIP	Difference
3	5.0%	2.1%	-2.9%	2.4%	2.3%	-0.1%
7	26.7%	40.0%	13.3%	28.6%	44.6%	16.0%
10	31.2%	51.6%	20.4%	36.5%	58.2%	21.7%
mixed	21.4%	32.3%	10.9%	22.3%	35.7%	13.4%

Table 77: Detection rate differences between algorithms, estimated parameters, FDR = .05

Shift Length	3PL			NRM		
	CMP	SCIP	Difference	CMP	SCIP	Difference
3	11.1%	18.1%	7.0%	9.5%	19.6%	10.1%
7	42.3%	59.7%	17.4%	45.8%	66.7%	20.9%
10	47.2%	67.8%	20.6%	55.0%	75.7%	20.7%
mixed	35.0%	51.5%	16.5%	39.2%	57.4%	18.2%

Table 78: Detection rate differences between IRT models, estimated parameters, FDR = .00

Shift	CMP			SCIP		
Length	3PL	NRM	Difference	3PL	NRM	Difference
3	5.0%	2.4%	-2.6%	2.1%	2.3%	0.2%
7	26.7%	28.6%	1.9%	40.0%	44.6%	4.6%
10	31.2%	36.5%	5.3%	51.6%	58.2%	6.6%
mixed	21.4%	22.3%	0.9%	32.3%	35.7%	3.4%

Table 79: Detection rate differences between IRT models, estimated parameters, FDR = .05

Shift	CMP			SCIP		
Length	3PL	NRM	Difference	3PL	NRM	Difference
3	11.1%	9.5%	-1.6%	18.1%	19.6%	1.5%
7	42.3%	45.8%	3.5%	59.7%	66.7%	7.0%
10	47.2%	55.0%	7.8%	67.8%	75.7%	7.9%
mixed	35.0%	39.2%	4.2%	51.5%	57.4%	5.9%

Table 80: Differences between parameter estimation methods, CMP/3PL, FDR = .00

Shift Length	True	Estimated	Corrected	E-T Diff	C-E Diff
3	7.8%	5.0%	4.4%	-2.8%	-0.6%
7	34.2%	26.7%	25.0%	-7.5%	-1.7%
10	42.9%	31.2%	30.6%	-11.7%	-0.6%
mixed	29.1%	21.4%	21.1%	-7.7%	-0.3%

Table 81: Differences between parameter estimation methods, CMP/NRM, FDR = .00

Shift Length	True	Estimated	Corrected	E-T Diff	C-E Diff
3	10.0%	2.4%	7.9%	-7.6%	5.5%
7	39.0%	28.6%	33.1%	-10.4%	4.5%
10	49.3%	36.5%	38.9%	-12.8%	2.4%
mixed	33.6%	22.3%	27.1%	-11.3%	4.8%

Table 82: Differences between parameter estimation methods, SCIP/3PL, FDR = .00

Shift Length	True	Estimated	Corrected	E-T Diff	C-E Diff
3	1.6%	2.1%	0.1%	0.5%	-2.0%
7	43.2%	40.0%	33.4%	-3.2%	-6.6%
10	60.2%	51.6%	53.6%	-8.6%	2.0%
mixed	35.2%	32.3%	26.7%	-2.9%	-5.6%

Table 83: Differences between parameter estimation methods, SCIP/NRM, FDR = .00

Shift Length	True	Estimated	Corrected	E-T Diff	C-E Diff
3	2.2%	2.3%	0.2%	0.1%	-2.1%
7	49.2%	44.6%	37.9%	-4.6%	-6.7%
10	65.7%	58.2%	58.3%	-7.5%	0.1%
mixed	39.9%	35.7%	30.0%	-4.2%	-5.7%

Table 84: Differences between parameter estimation methods, CMP/3PL, FDR = .05

Shift Length	True	Estimated	Corrected	E-T Diff	C-E Diff
3	15.8%	11.1%	8.3%	-4.7%	-2.8%
7	48.5%	42.3%	38.2%	-6.2%	-4.1%
10	57.8%	47.2%	43.7%	-10.6%	-3.5%
mixed	41.1%	35.0%	32.1%	-6.1%	-2.9%

Table 85: Differences between parameter estimation methods, CMP/NRM, FDR = .05

Shift Length	True	Estimated	Corrected	E-T Diff	C-E Diff
3	19.8%	9.5%	16.7%	-10.3%	7.2%
7	54.4%	45.8%	49.3%	-8.6%	3.5%
10	64.3%	55.0%	55.6%	-9.3%	0.6%
mixed	47.9%	39.2%	41.2%	-8.7%	2.0%

Table 86: Differences between parameter estimation methods, SCIP/3PL, FDR = .05

Shift Length	True	Estimated	Corrected	E-T Diff	C-E Diff
3	13.5%	18.1%	0.2%	4.6%	-17.9%
7	64.7%	59.7%	60.5%	-5.0%	0.8%
10	74.9%	67.8%	72.1%	-7.1%	4.3%
mixed	56.0%	51.5%	50.6%	-4.5%	-0.9%

Table 87: Differences between parameter estimation methods, SCIP/NRM, FDR = .05

Shift Length	True	Estimated	Corrected	E-T Diff	C-E Diff
3	21.9%	19.6%	1.0%	-2.3%	-18.6%
7	70.0%	66.7%	65.8%	-3.3%	-0.9%
10	79.6%	75.7%	77.1%	-3.9%	1.4%
mixed	60.7%	57.4%	55.1%	-3.3%	-2.3%

Table 88: Counts and projected total shift errors in empirical data, mixed length shifts

Simulated FDR	Threshold	Simulated TPR	Count	Projected Total
0.05	2.14	57.4%	88	146
0.25	1.65	67.3%	530	591
0.5	1.37	75.9%	1554	1024
0.75	1.01	83.7%	5234	1563
0.94	0.11	92.3%	29021	1887

APPENDIX II

FIGURES

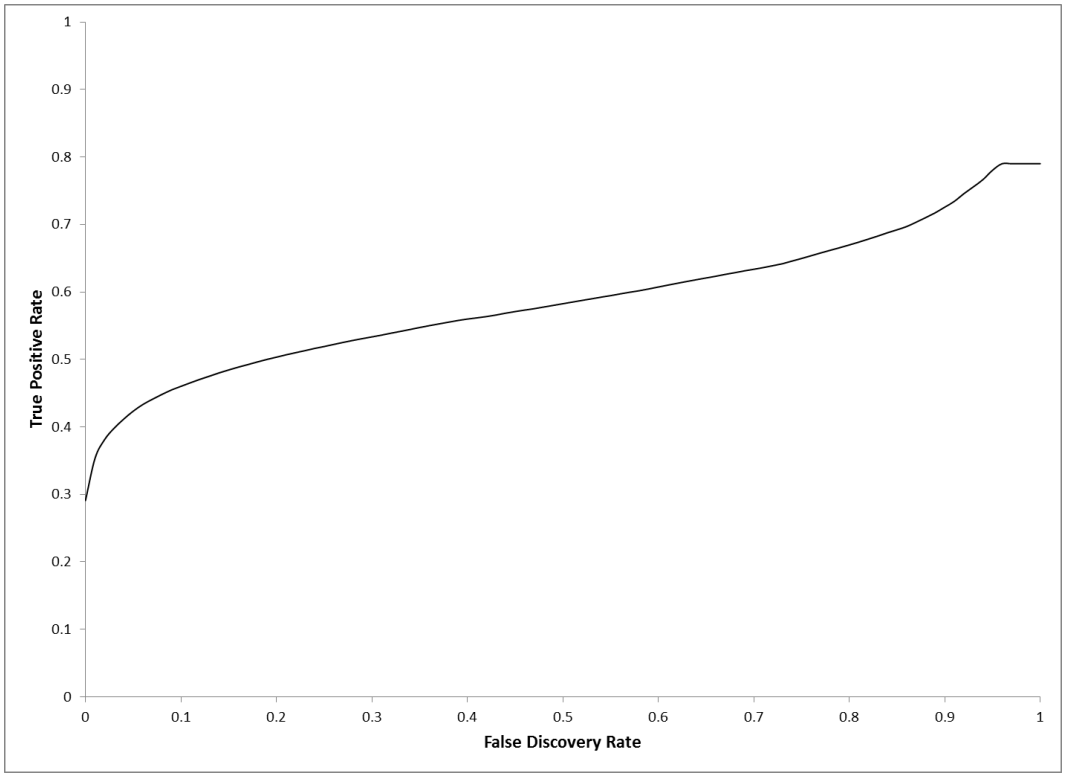

Figure 1: ROC curve using false discovery rate

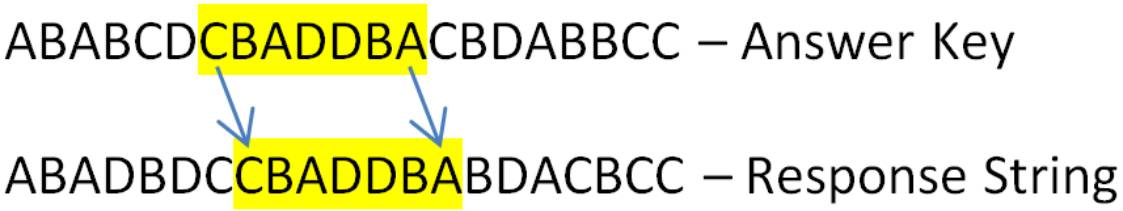

Figure 2: Misaligned response string, misaligned 1 forward starting at item 7

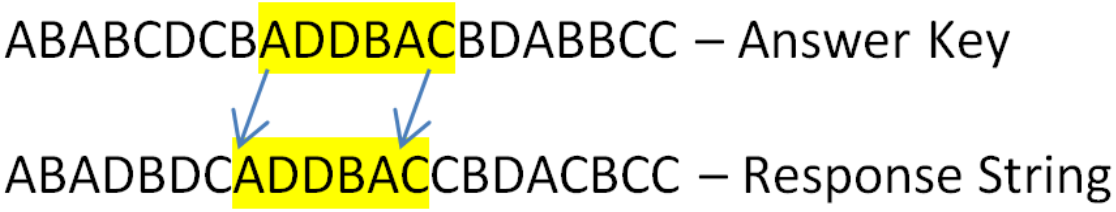

Figure 3: Misaligned response string, misaligned 1 backward starting at item 8

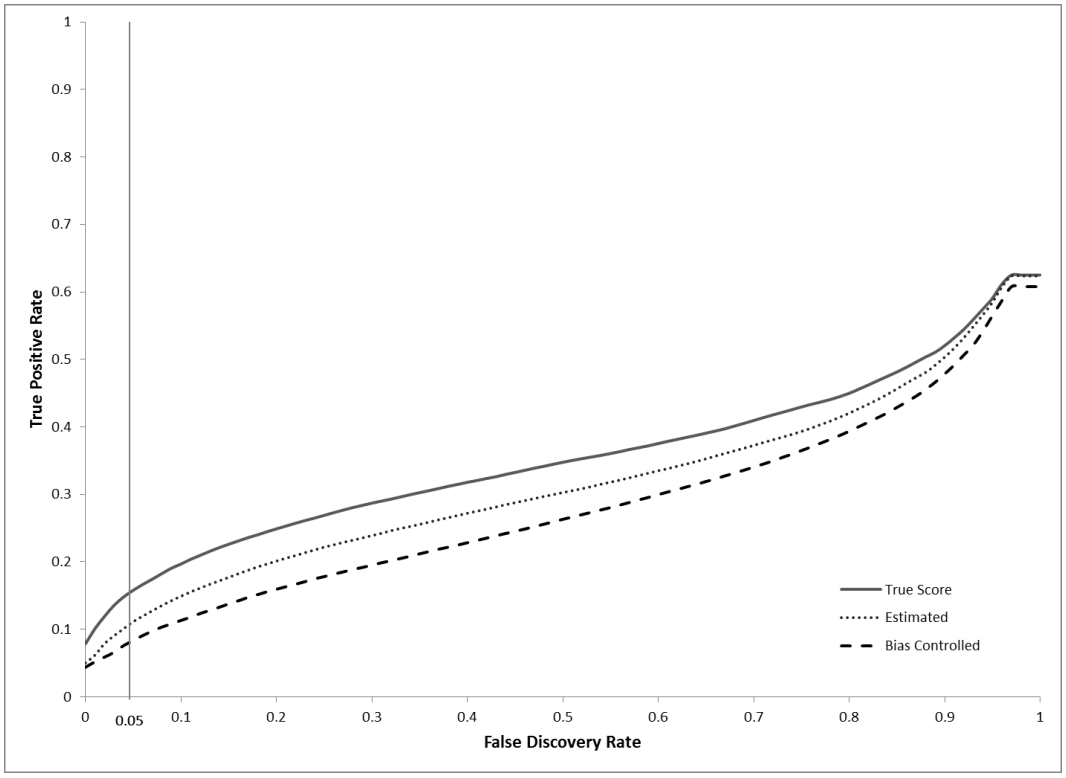

Figure 4: ROC Curves, CMP/3PL, all person parameter methods, shift length 3

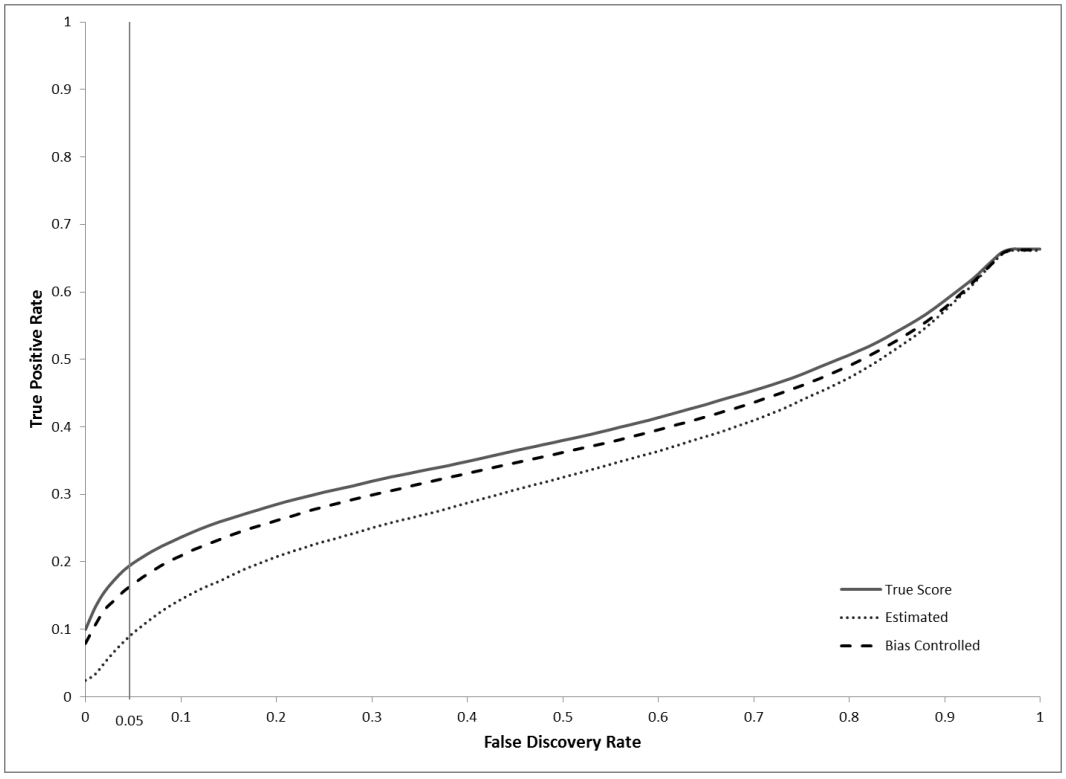

Figure 5: ROC Curves, CMP/NRM, all person parameter methods, shift length 3

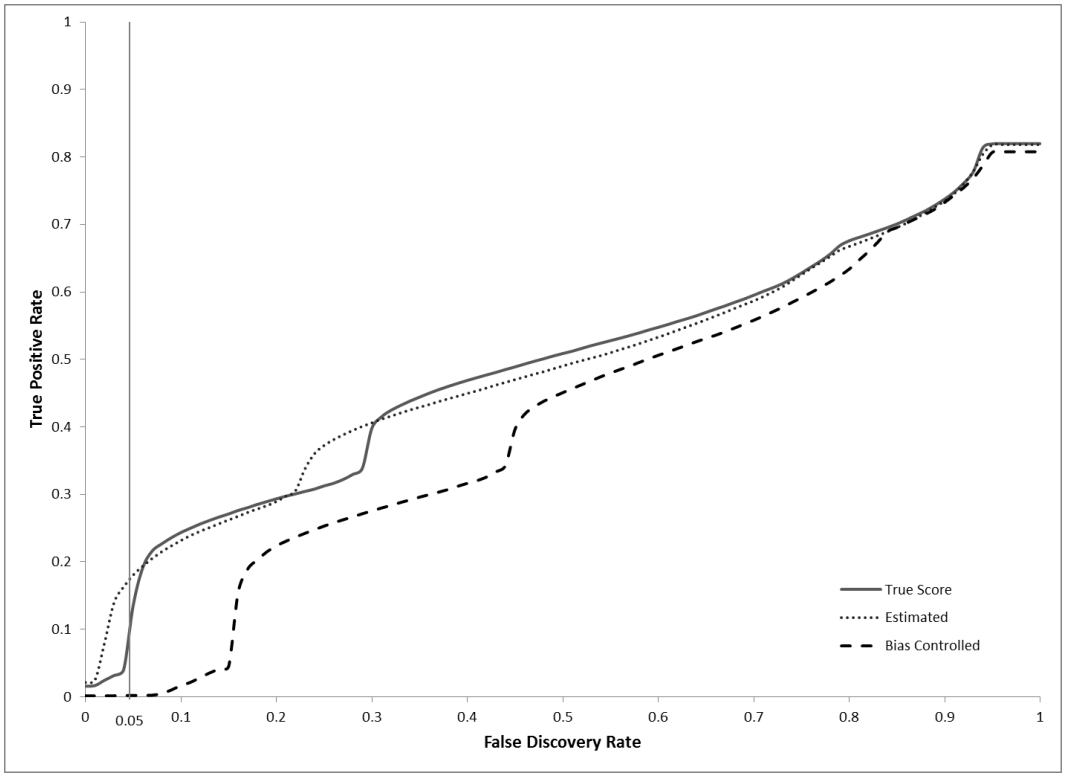

Figure 6: ROC Curves, SCIP/3PL, all person parameter methods, shift length 3

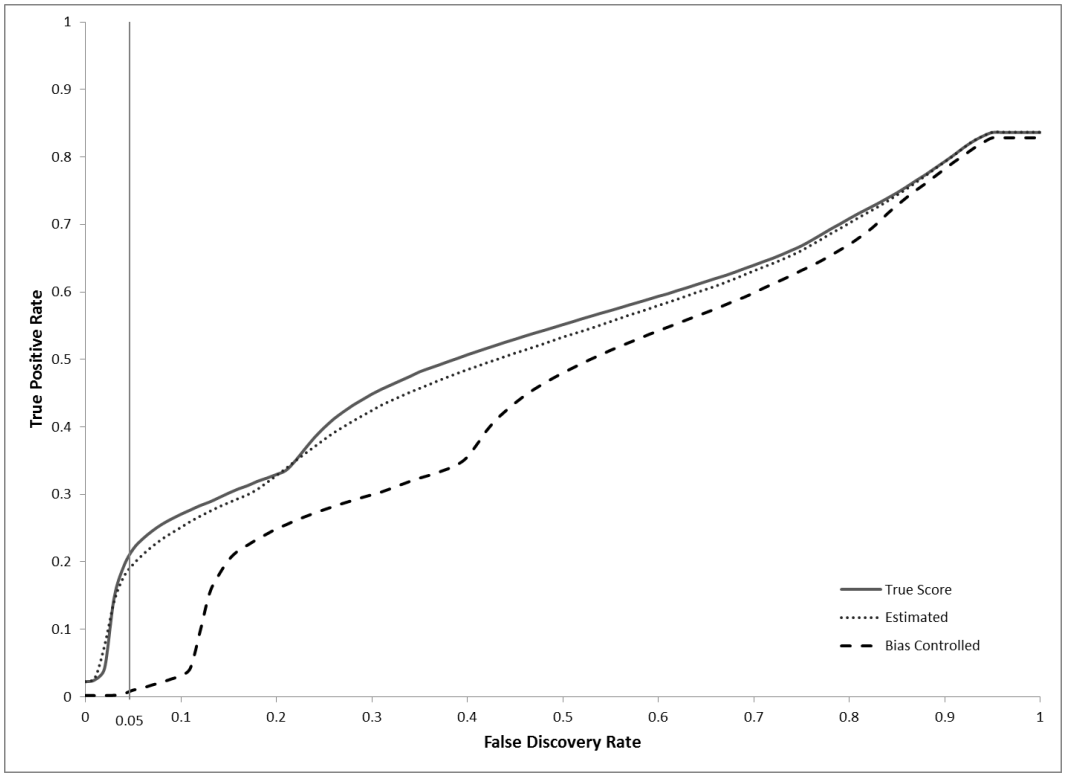

Figure 7: ROC Curves, SCIP/NRM, all person parameter methods, shift length 3

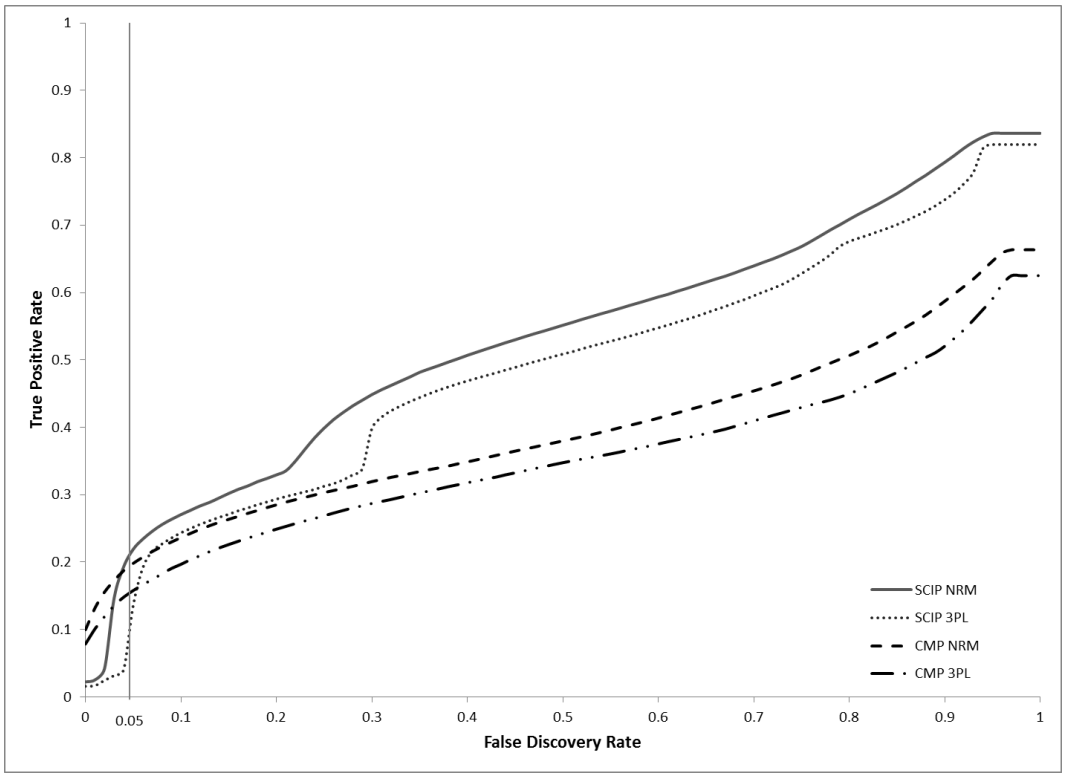

Figure 8: ROC Curves, all methods, true person parameters, shift length 3

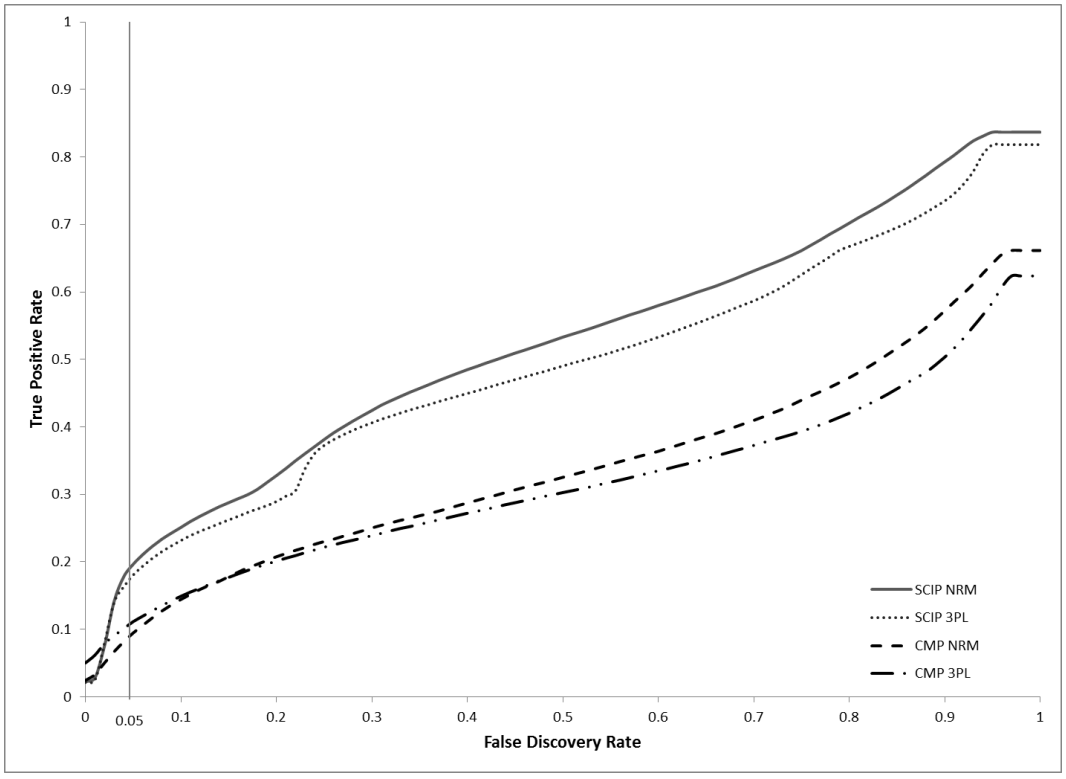

Figure 9: ROC Curves, all methods, estimated person parameters, shift length 3

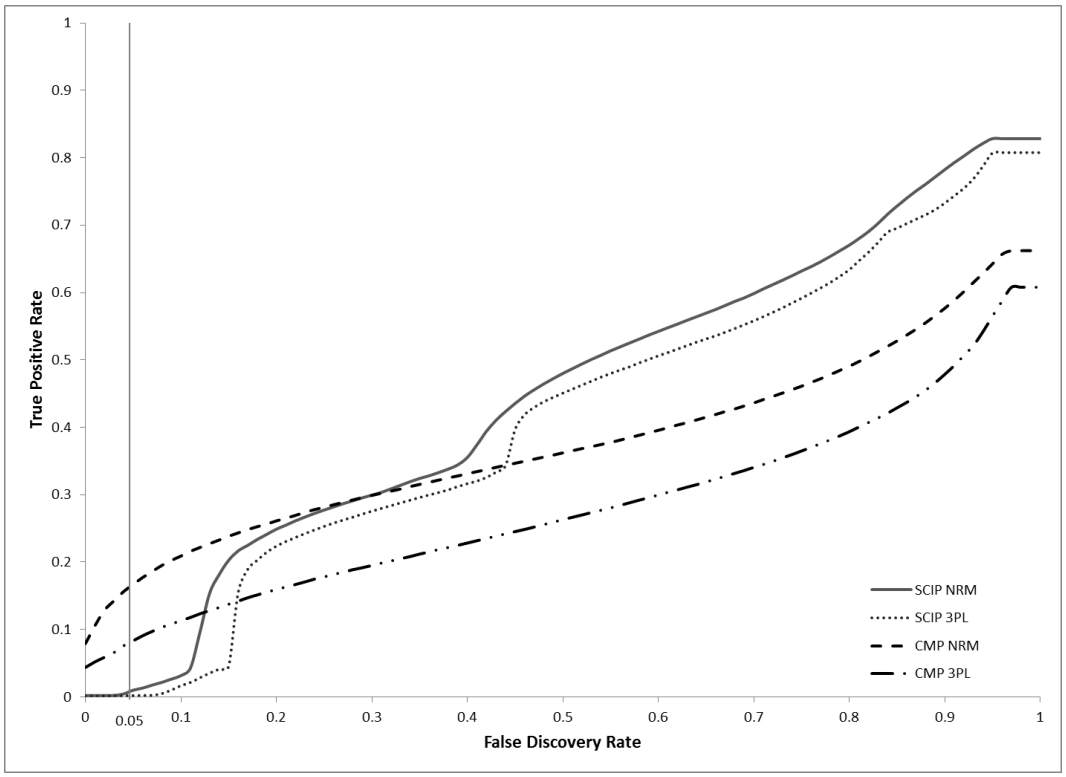

Figure 10: ROC Curves, all methods , bias-corrected person parameters, shift length 3

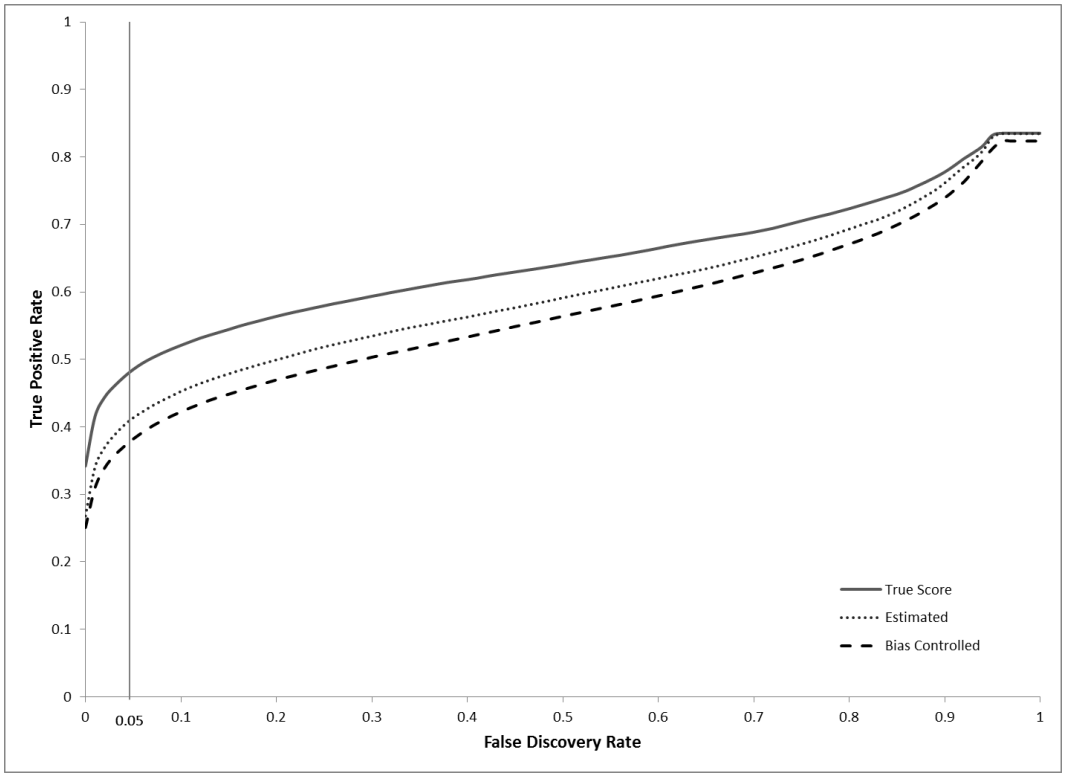

Figure 11: ROC Curves, CMP/3PL, all person parameter methods, shift length 7

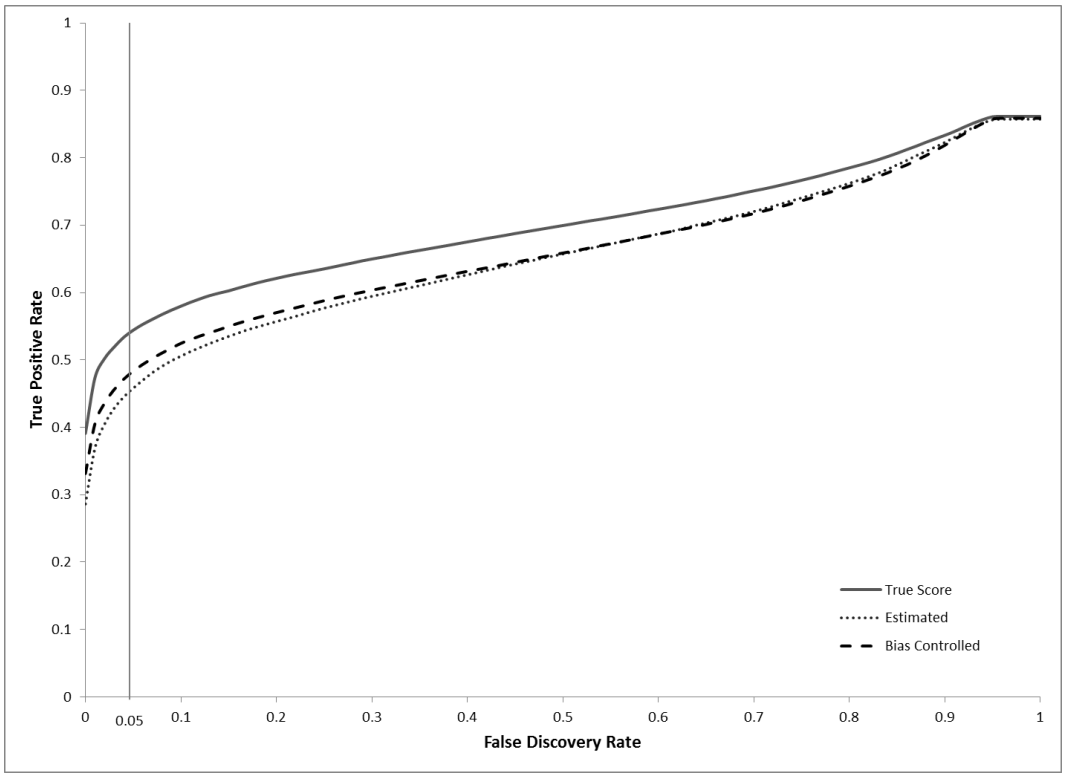

Figure 12: ROC Curves, CMP/NRM, all person parameter methods, shift length 7

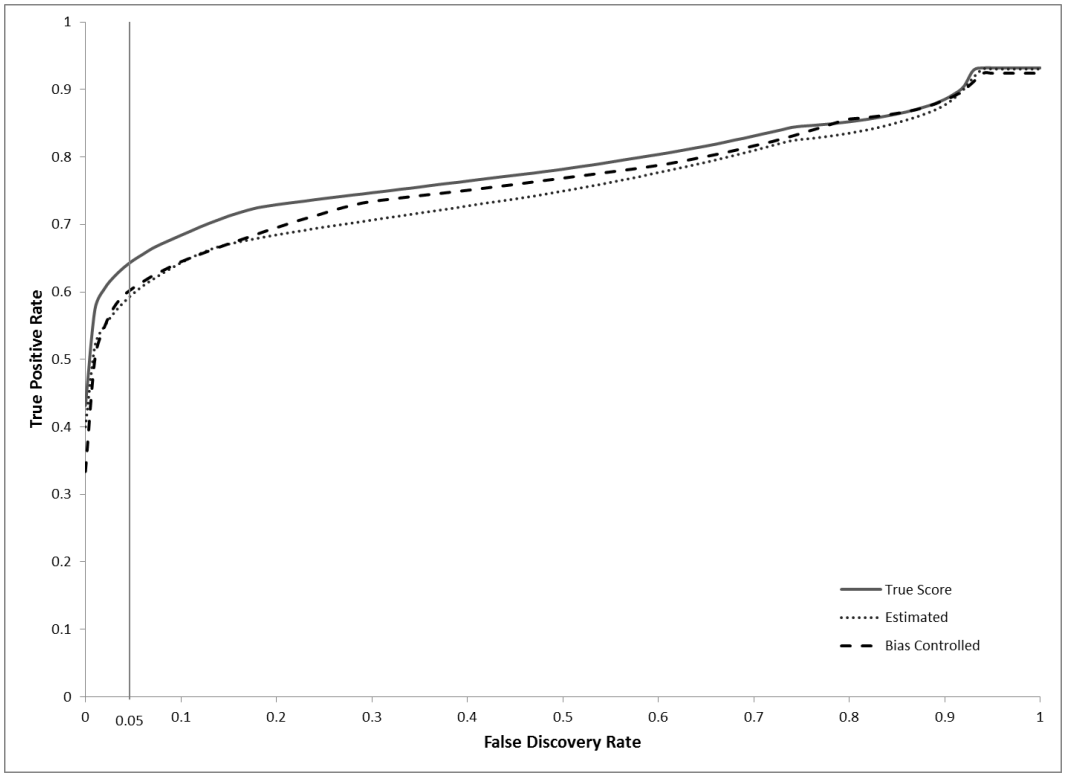

Figure 13: ROC Curves, SCIP/3PL, all person parameter methods, shift length 7

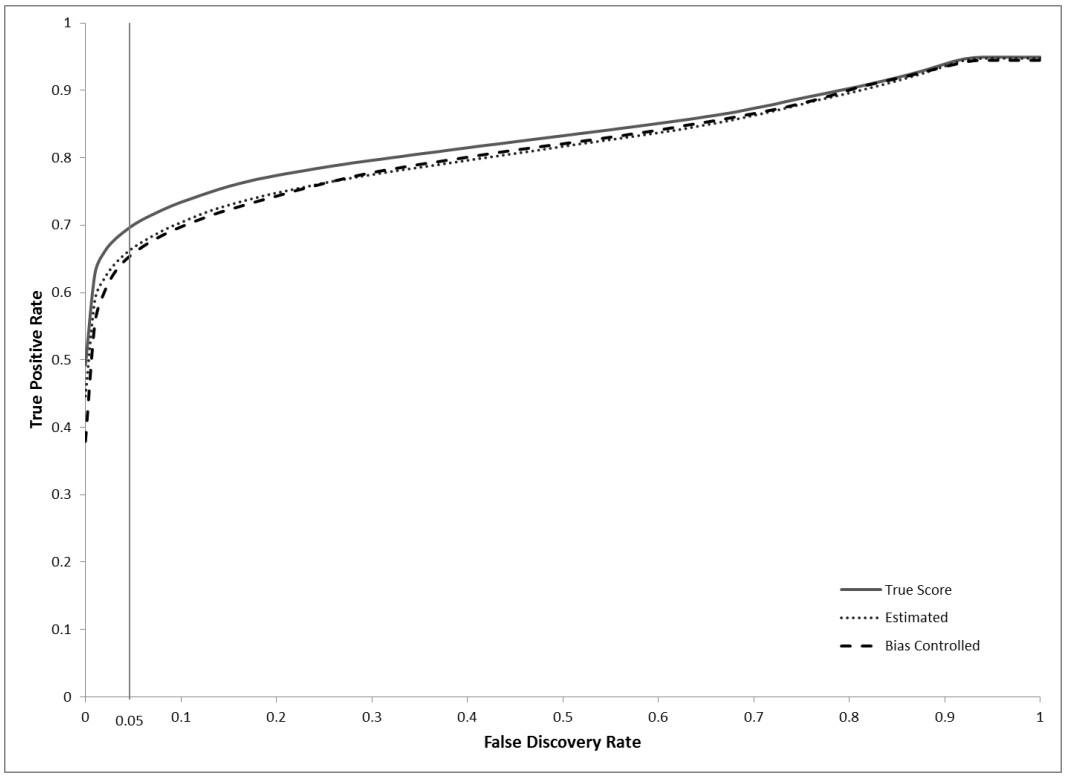

Figure 14: ROC Curves, SCIP/NRM, all person parameter methods, shift length 7

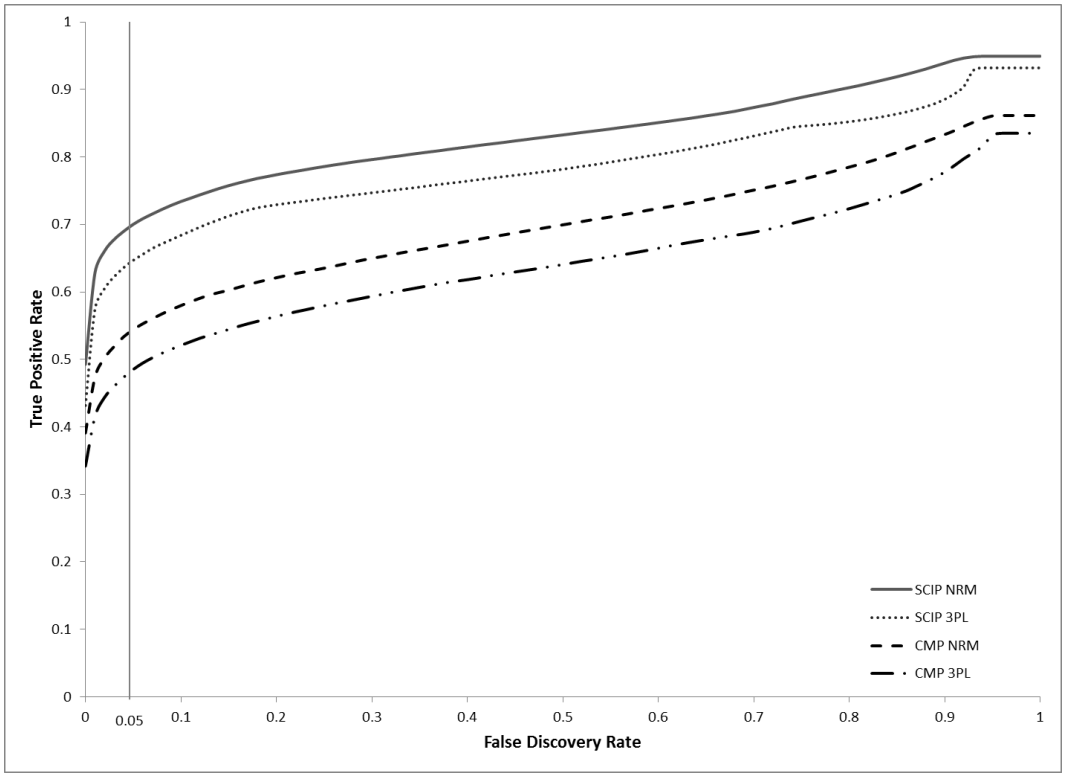

Figure 15: ROC Curves, all methods, true person parameters, shift length 7

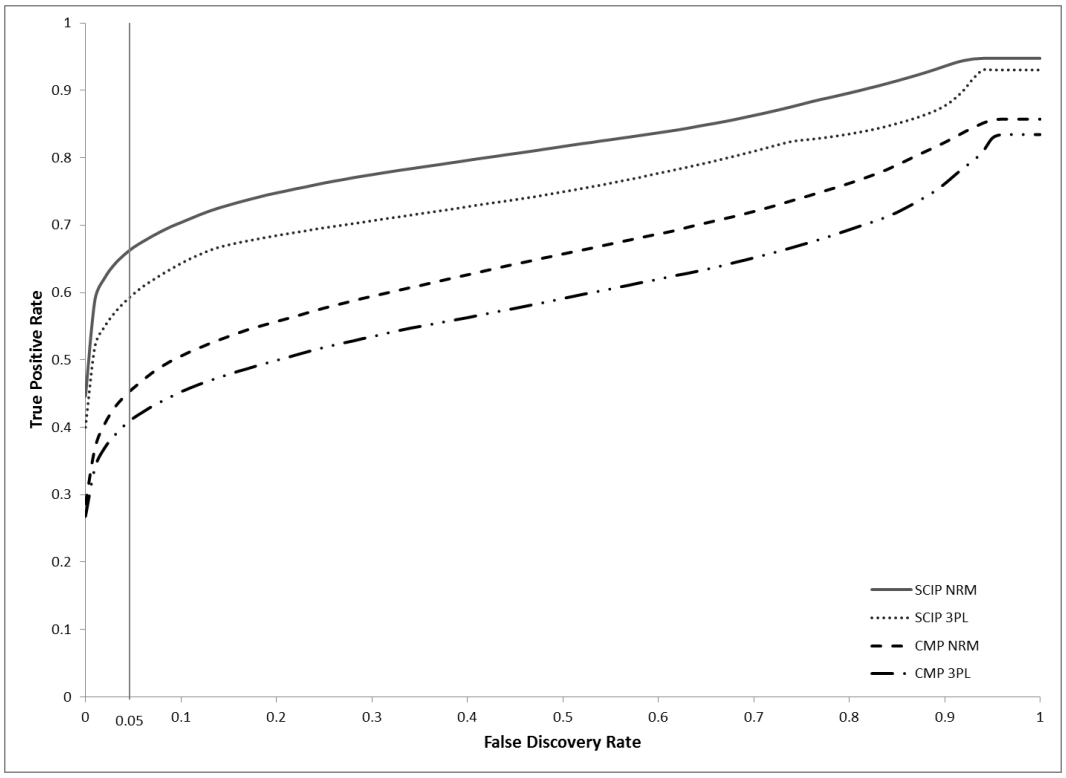

Figure 16: ROC Curves, all methods, estimated person parameters, shift length 7

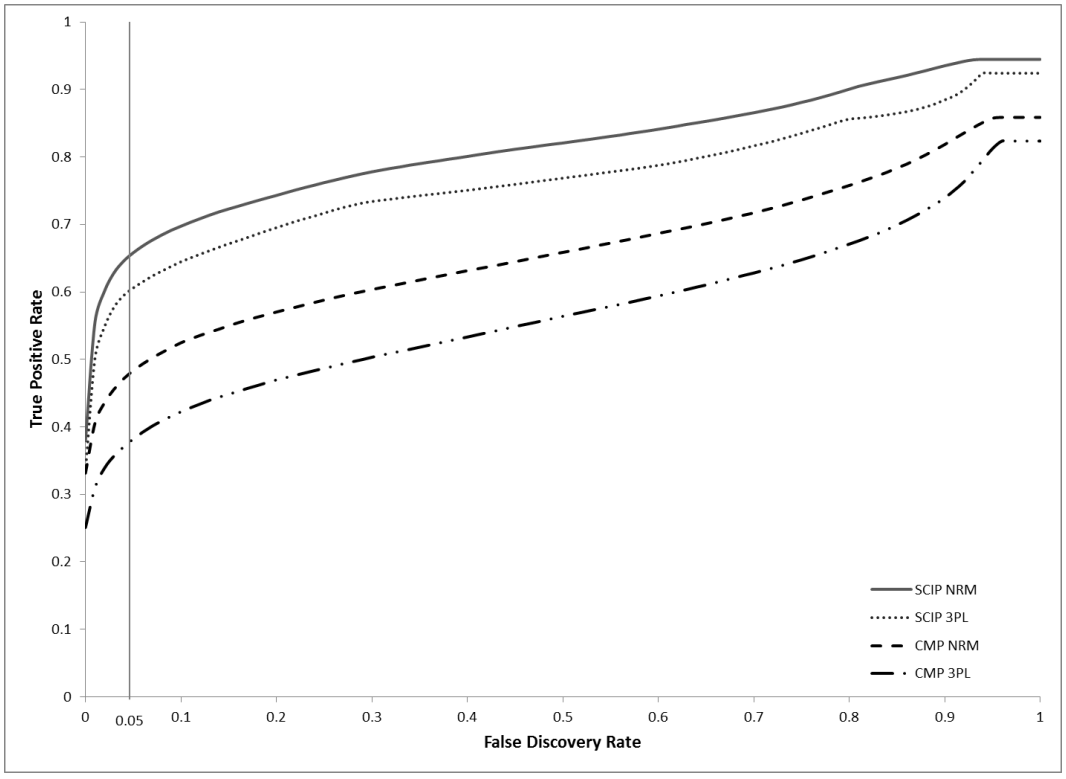

Figure 17: ROC Curves, all methods , bias-corrected person parameters, shift length 7

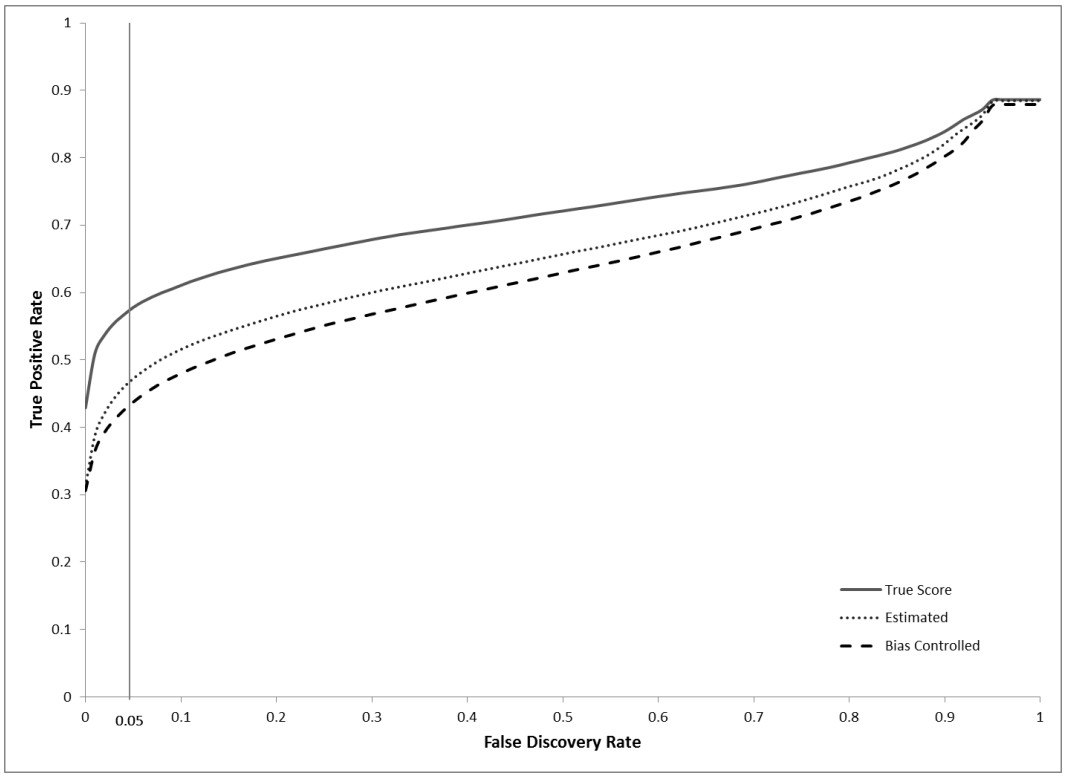

Figure 18: ROC Curves, CMP/3PL, all person parameter methods, shift length 10

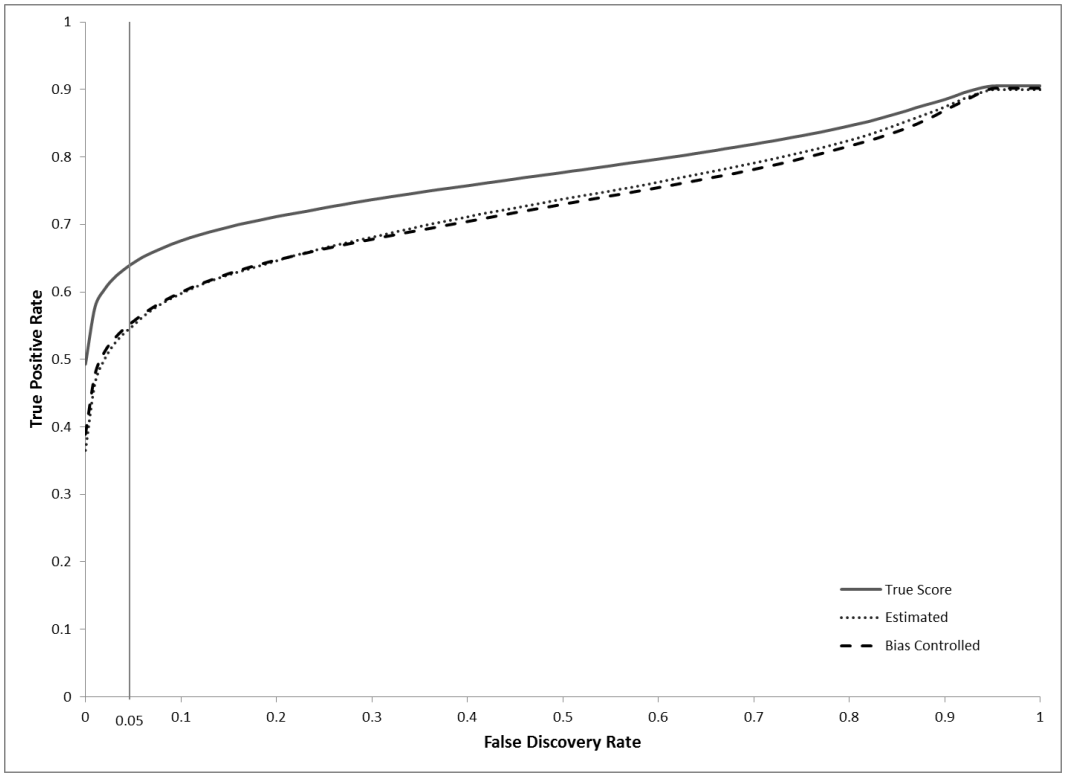

Figure 19: ROC Curves, CMP/NRM, all person parameter methods, shift length 10

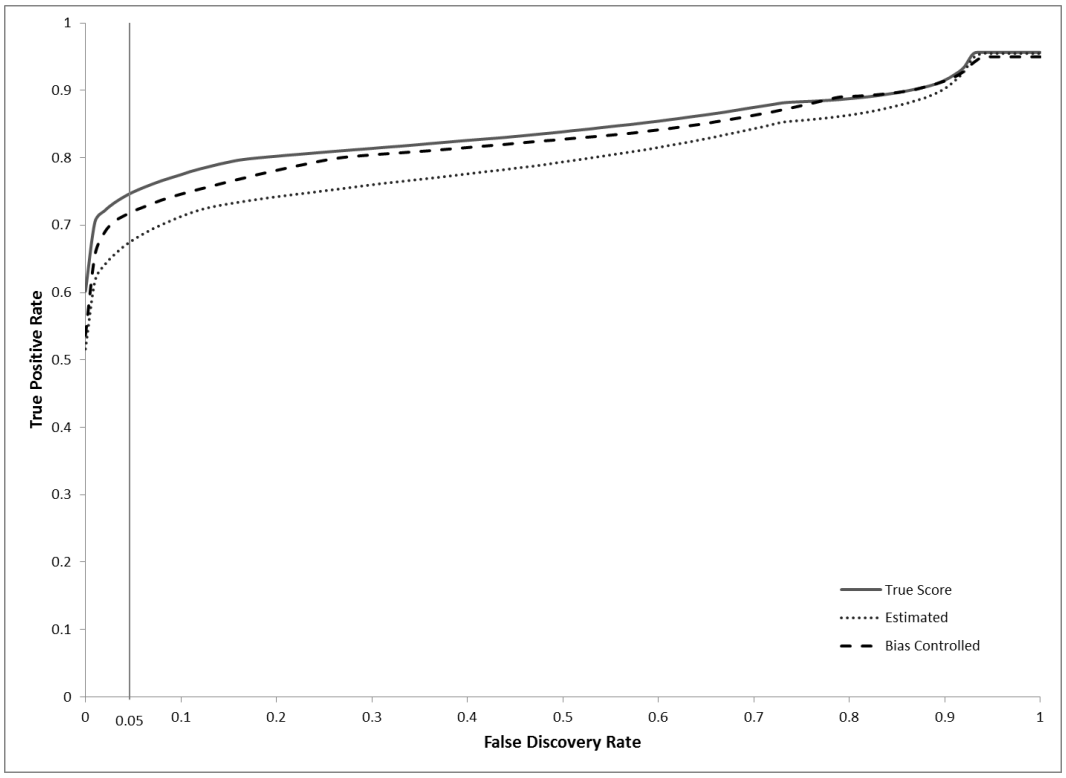

Figure 20: ROC Curves, SCIP/3PL, all person parameter methods, shift length 10

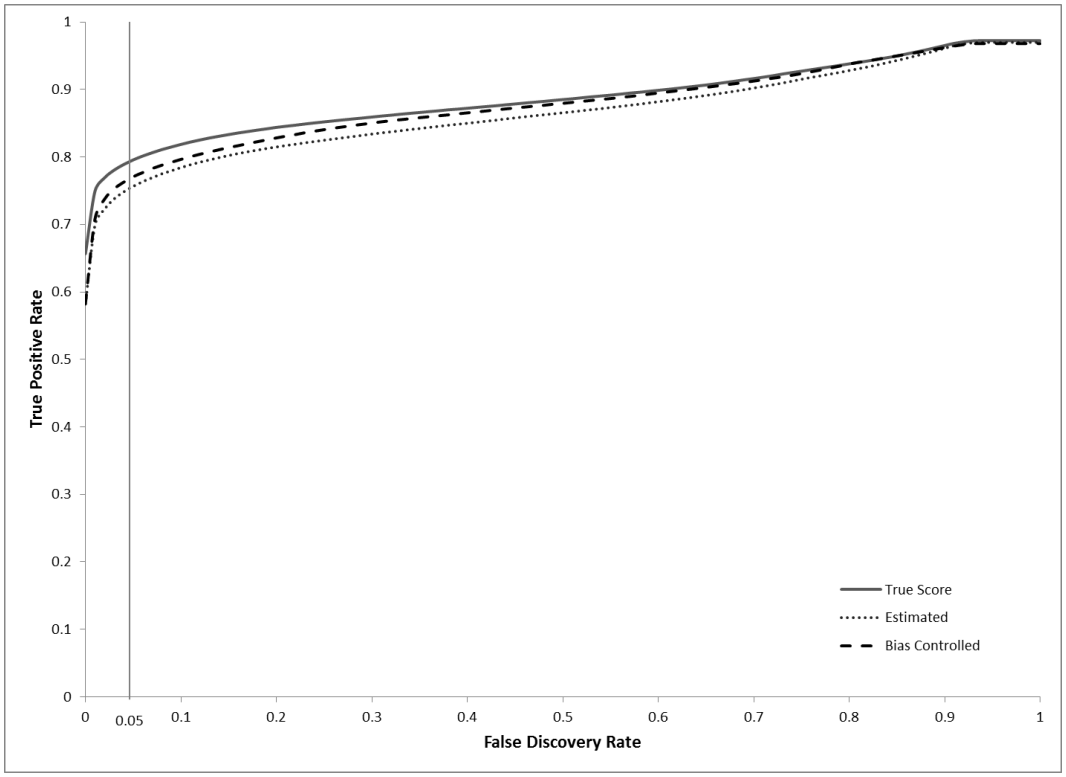

Figure 21: ROC Curves, SCIP/NRM, all person parameter methods, shift length 10

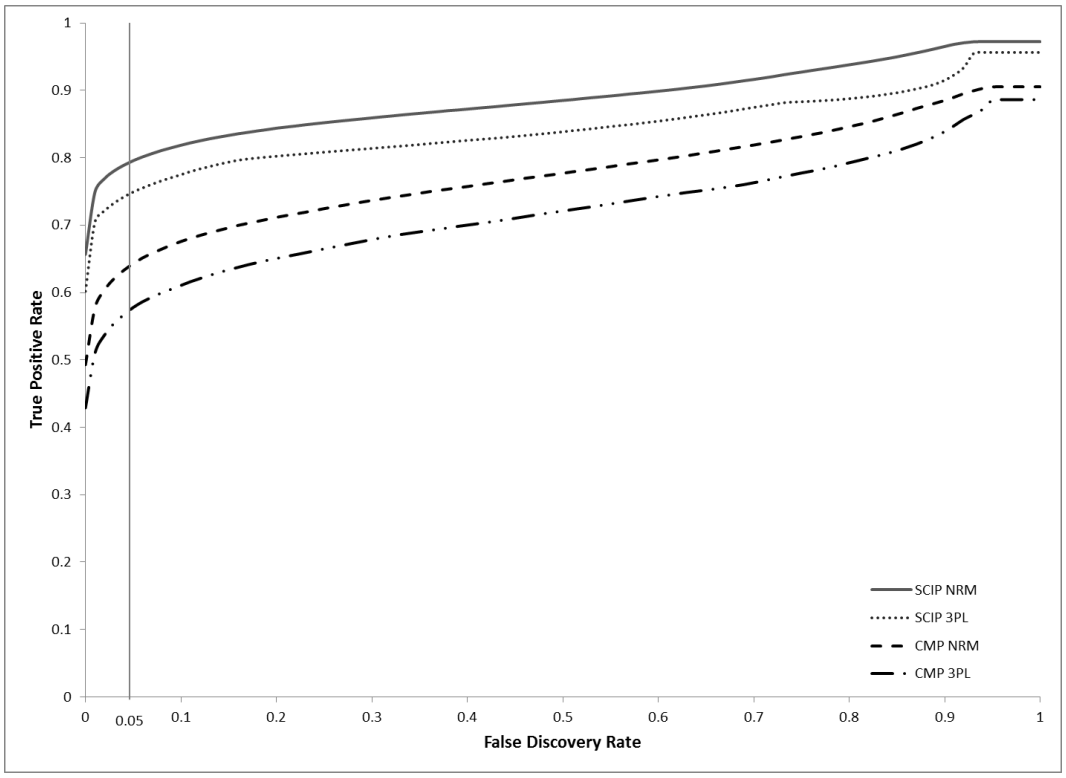

Figure 22: ROC Curves, all methods, true person parameters, shift length 10

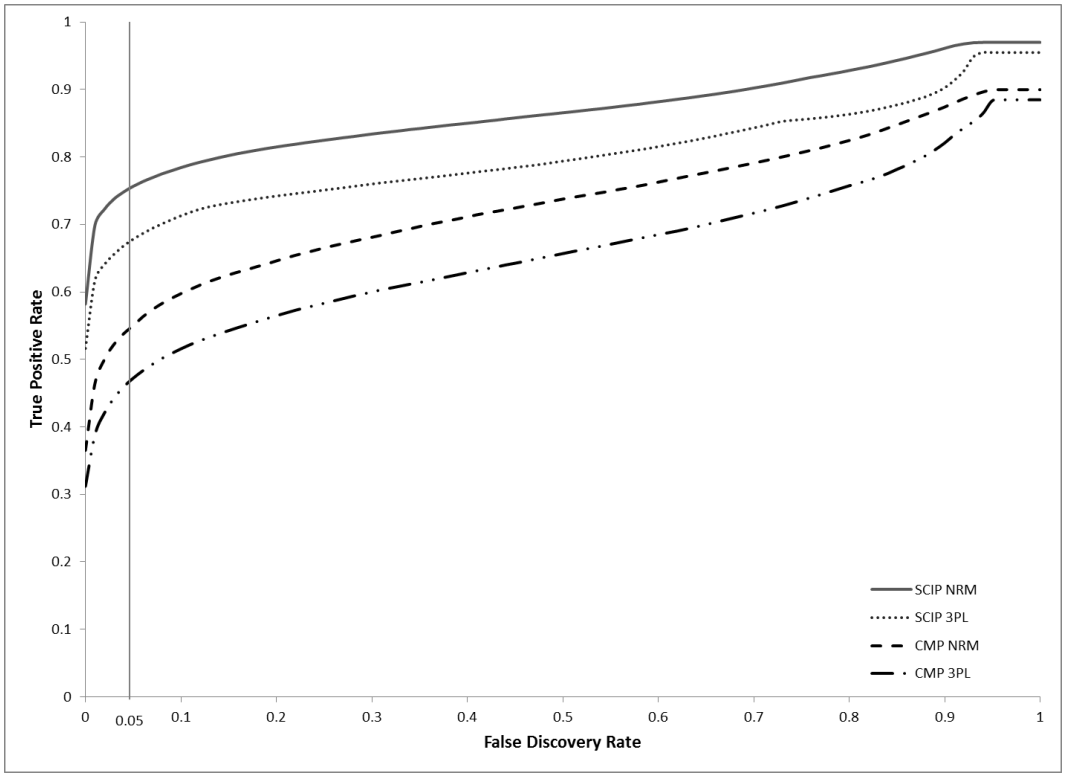

Figure 23: ROC Curves, all methods, estimated person parameters, shift length 10

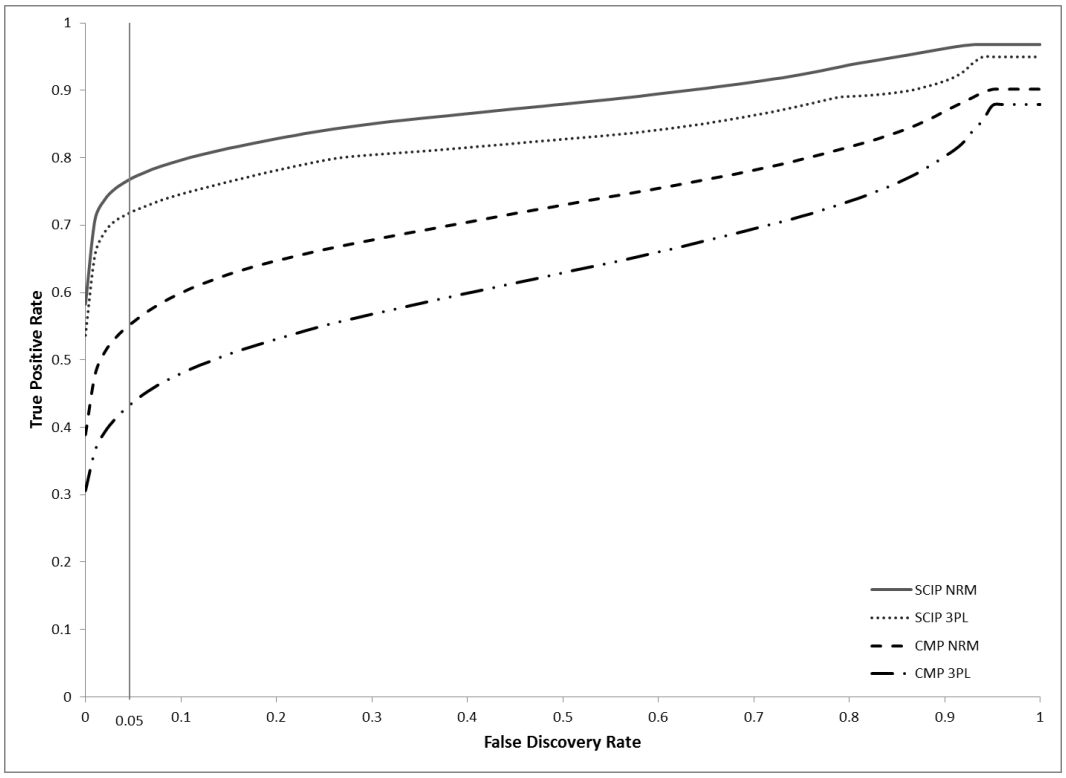

Figure 24: ROC Curves, all methods , bias-corrected person parameters, shift length 10

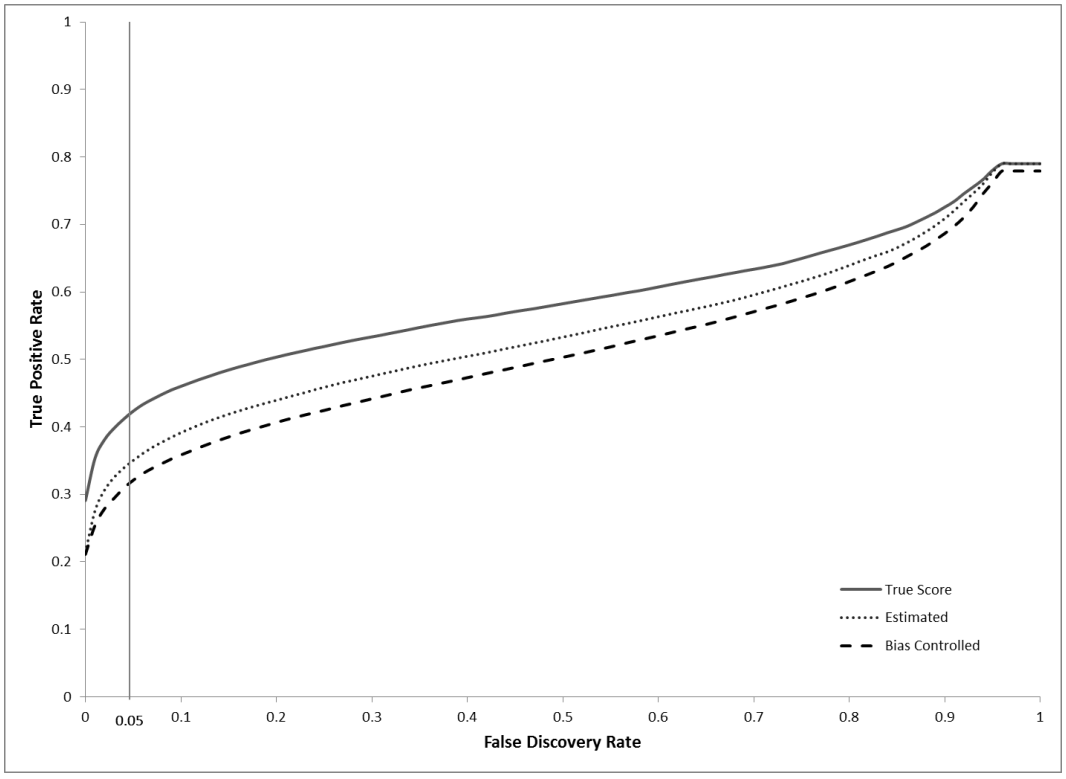

Figure 25: ROC Curves, CMP/3PL, all person parameter methods, mixed-length shifts

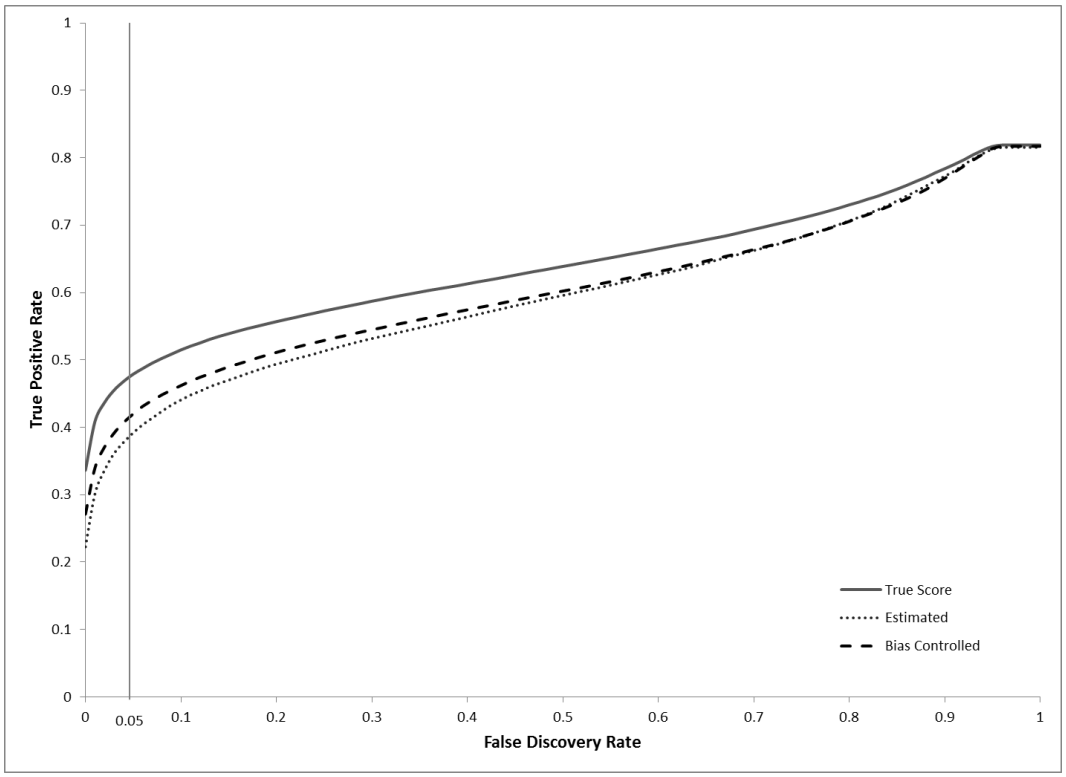

Figure 26: ROC Curves, CMP/NRM, all person parameter methods, mixed-length shifts

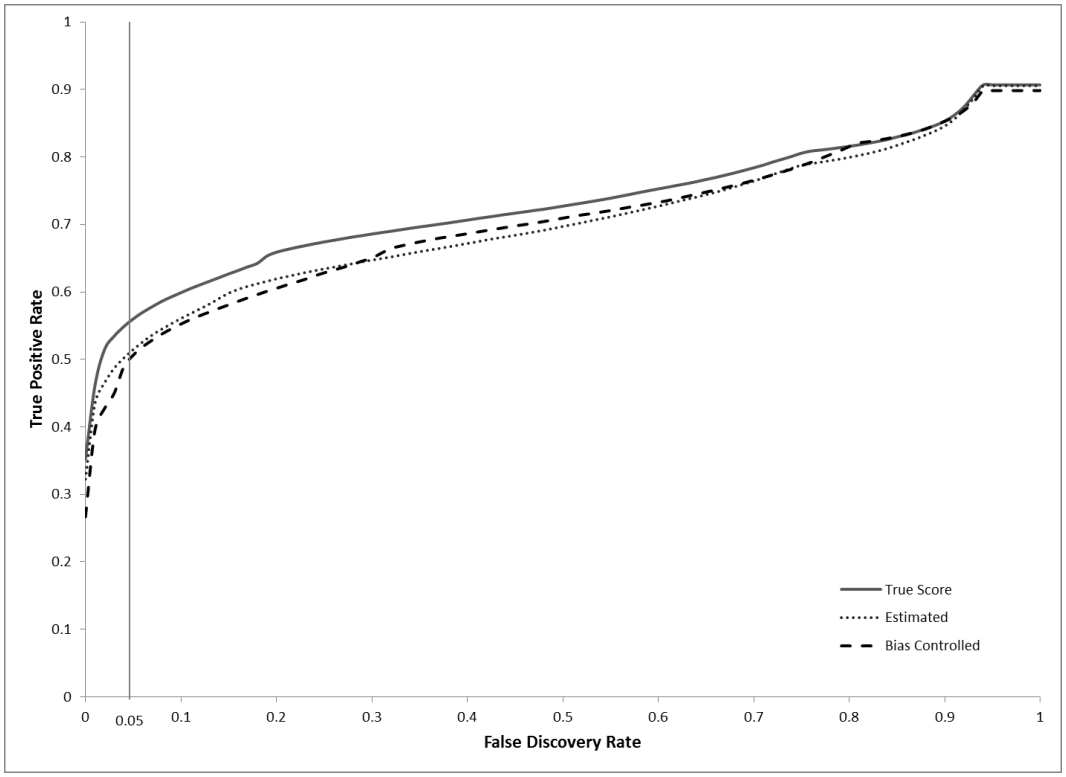

Figure 27: ROC Curves, SCIP/3PL, all person parameter methods, mixed-length shifts

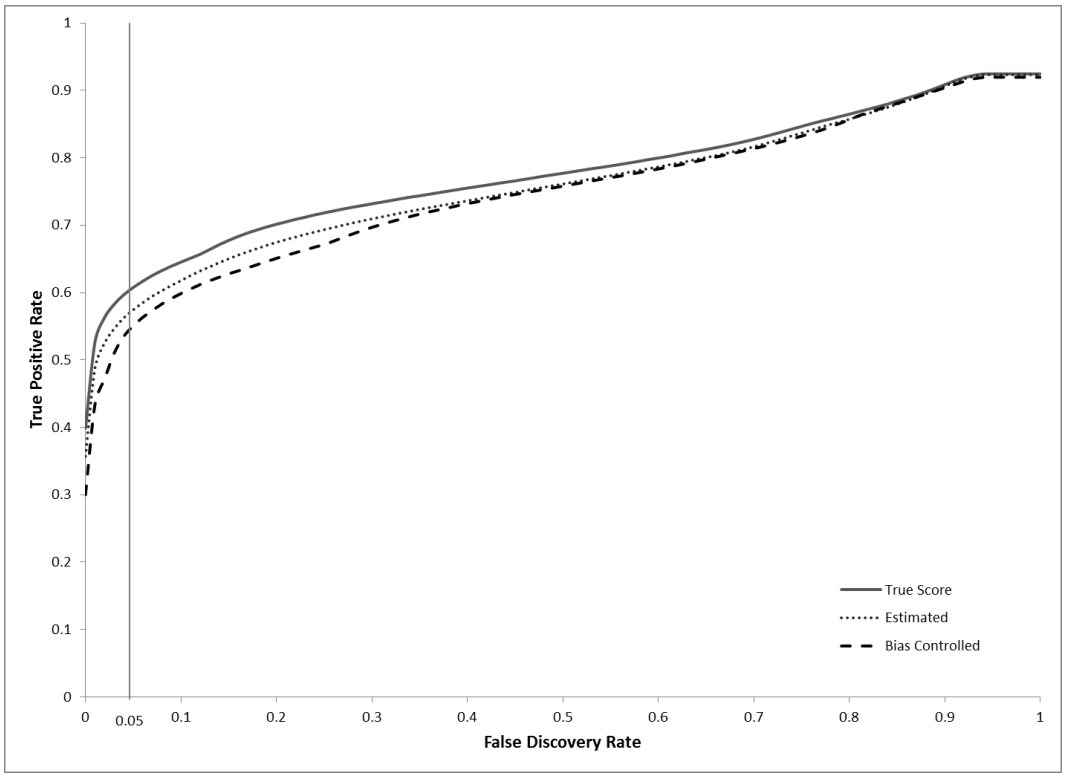

Figure 28: ROC Curves, SCIP/NRM, all person parameter methods, mixed-length shifts

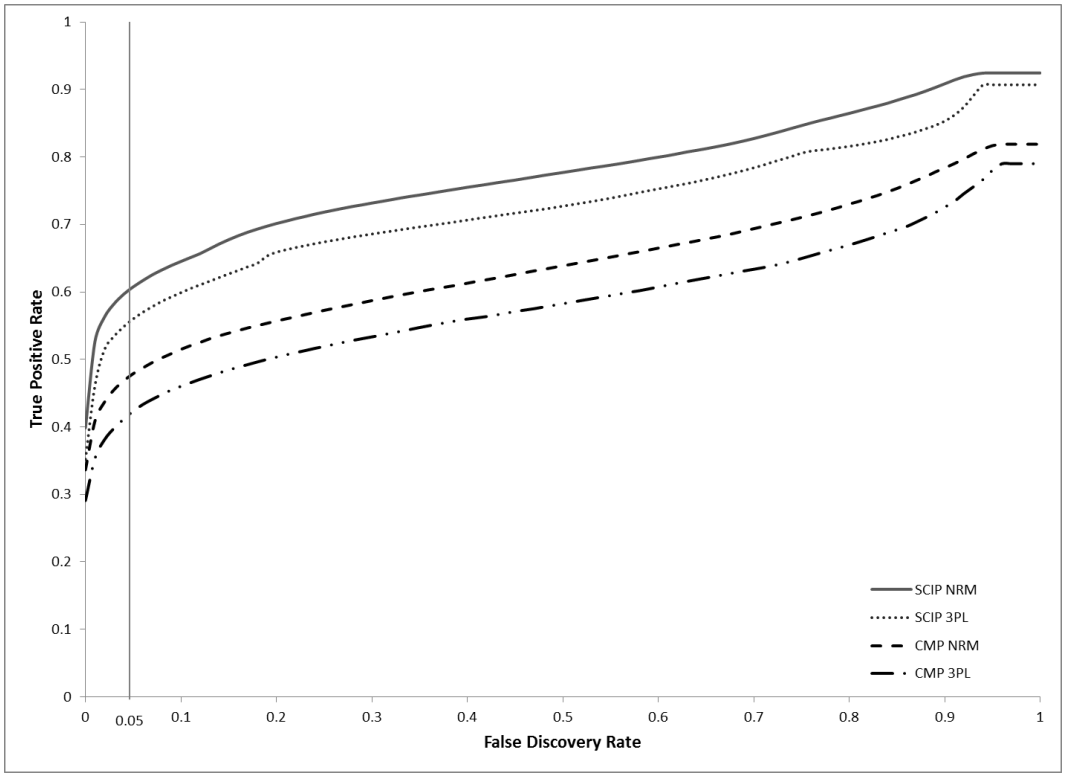

Figure 29: ROC Curves, all methods, true person parameters, mixed-length shifts

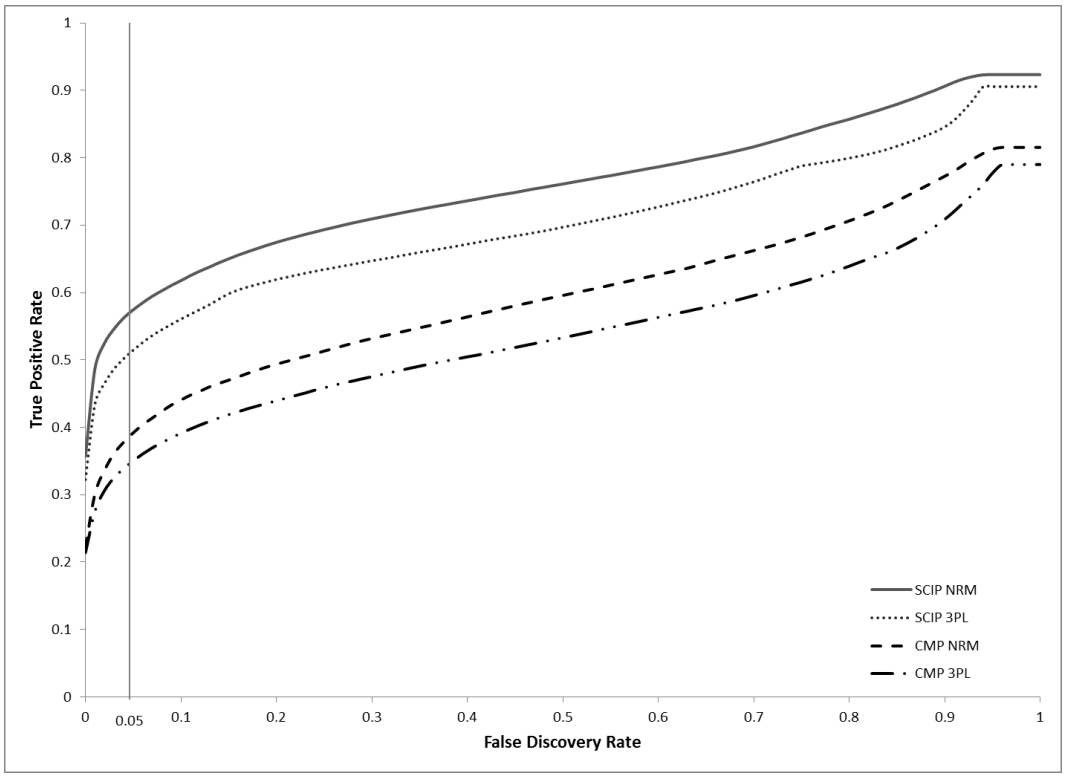

Figure 30: ROC Curves, all methods, estimated person parameters, mixed-length shifts

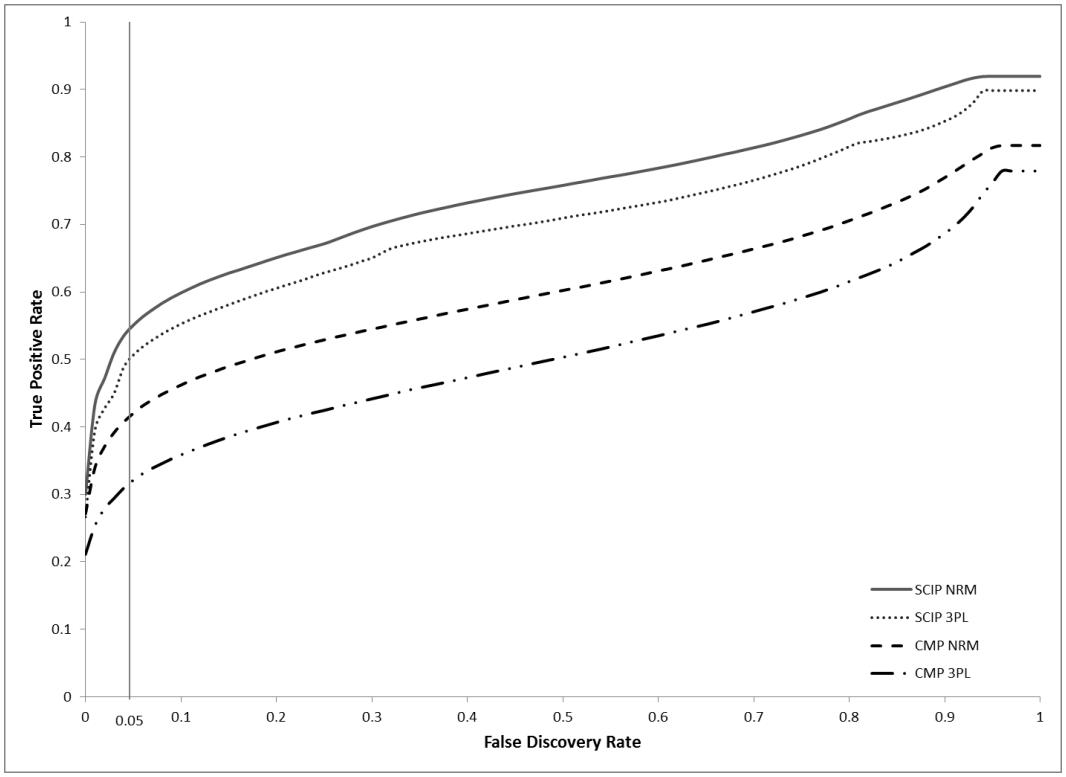

Figure 31: ROC Curves, all methods , bias-corrected person parameters, mixed shifts

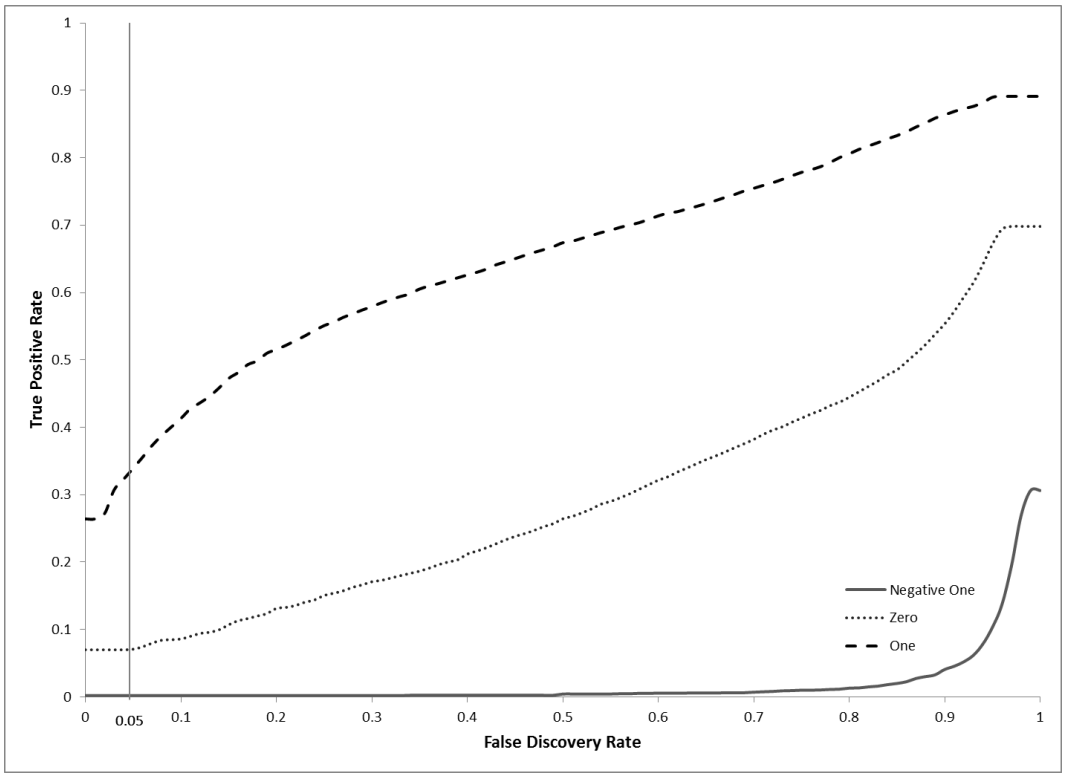

Figure 32: ROC Curves, CMP/3PL, true person parameters, shift error length 3

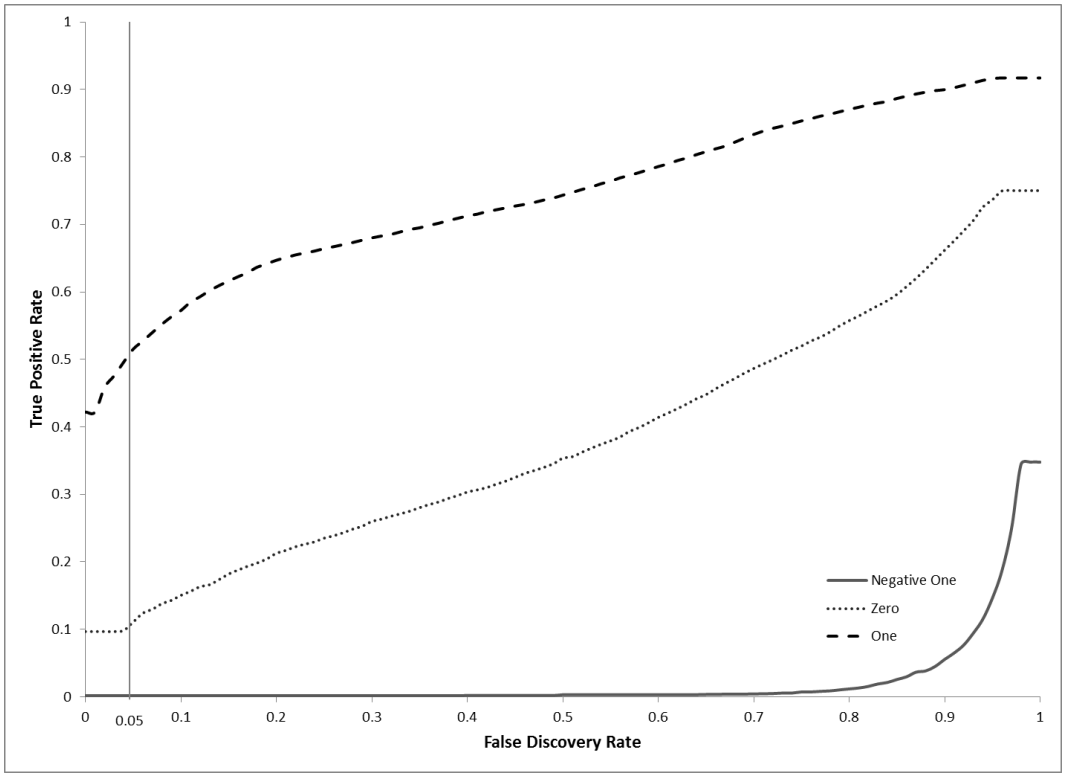

Figure 33: ROC Curves, CMP/NRM, true person parameters, shift error length 3

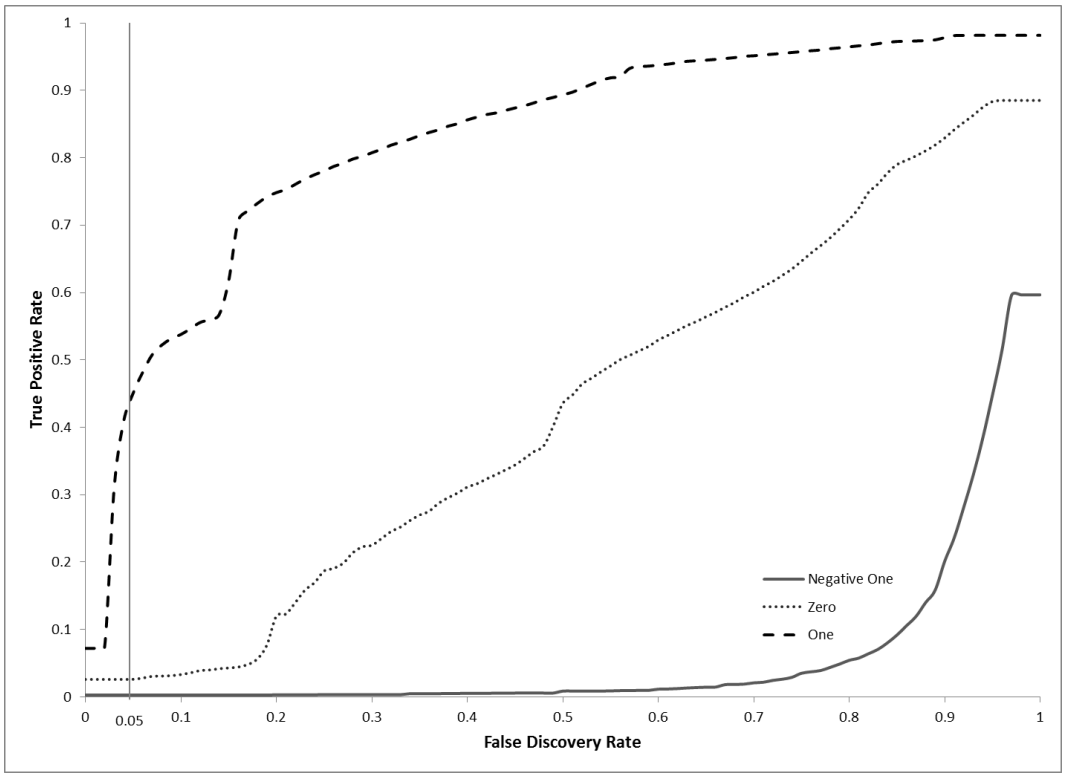

Figure 34: ROC Curves, SCIP/3PL, true person parameters, shift error length 3

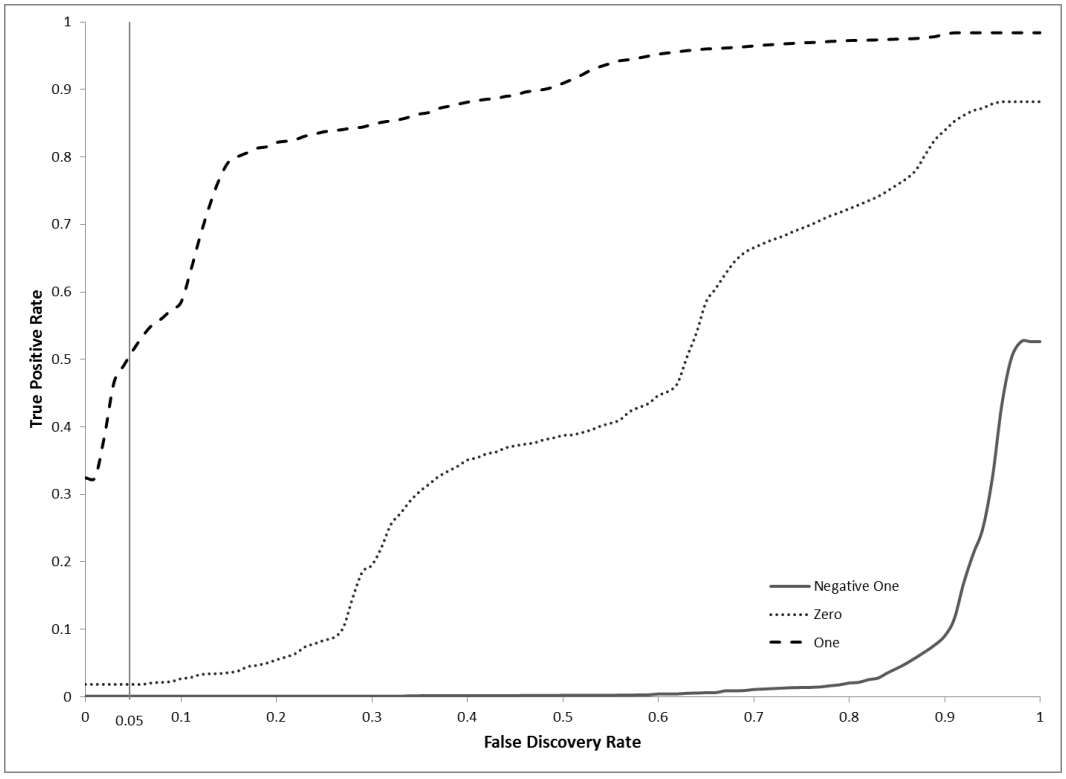

Figure 35: ROC Curves, SCIP/NRM, true person parameters, shift error length 3

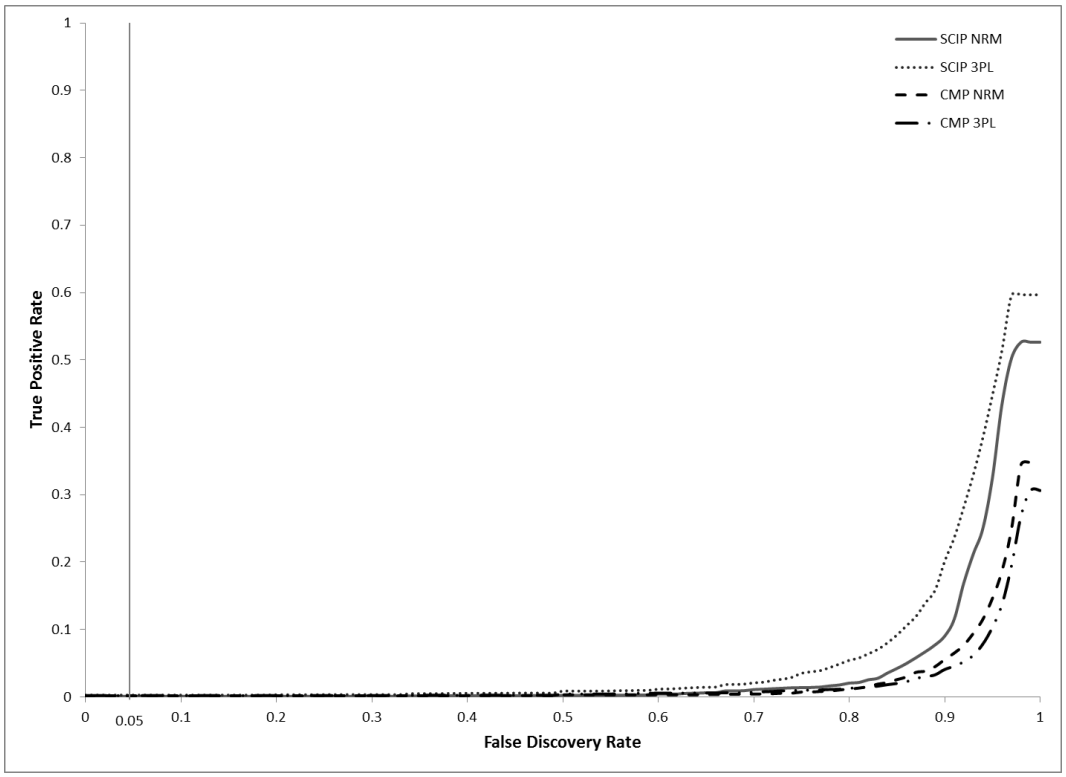

Figure 36: ROC Curves, all methods, true person parameters = -1, shift error length 3

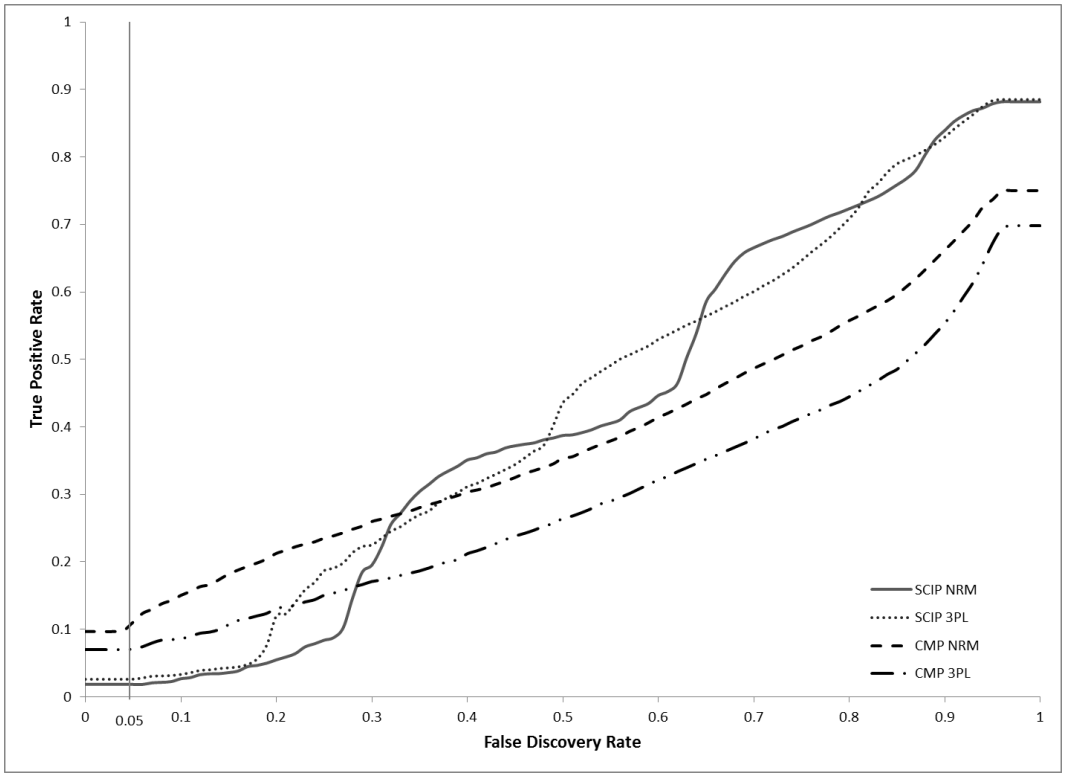

Figure 37: ROC Curves, all methods, true person parameters = 0, shift error length 3

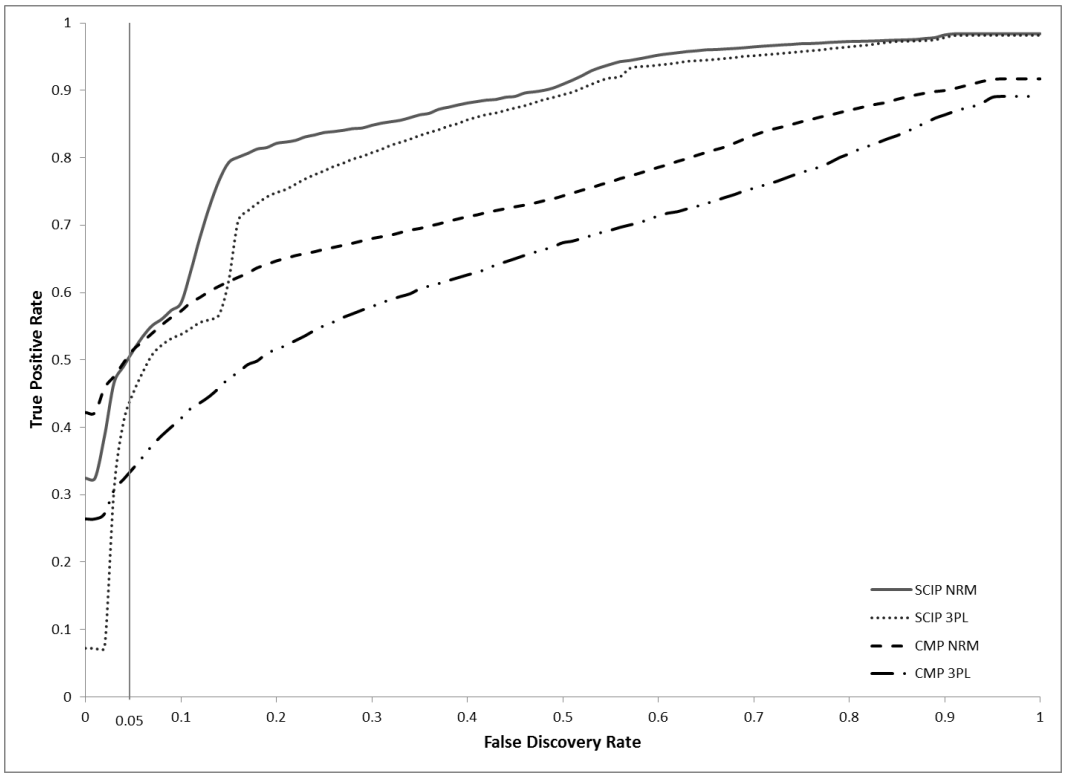

Figure 38: ROC Curves, all methods , true person parameters = 1, shift error length 3

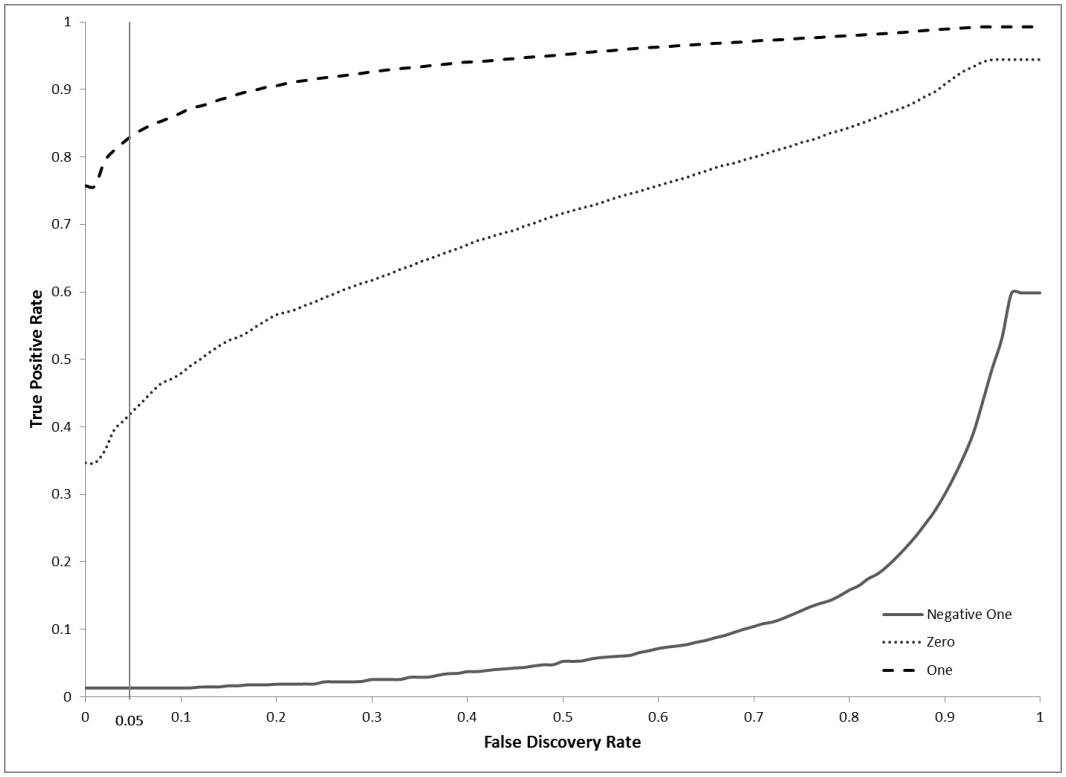

Figure 39: ROC Curves, CMP/3PL, true person parameters, shift error length 7

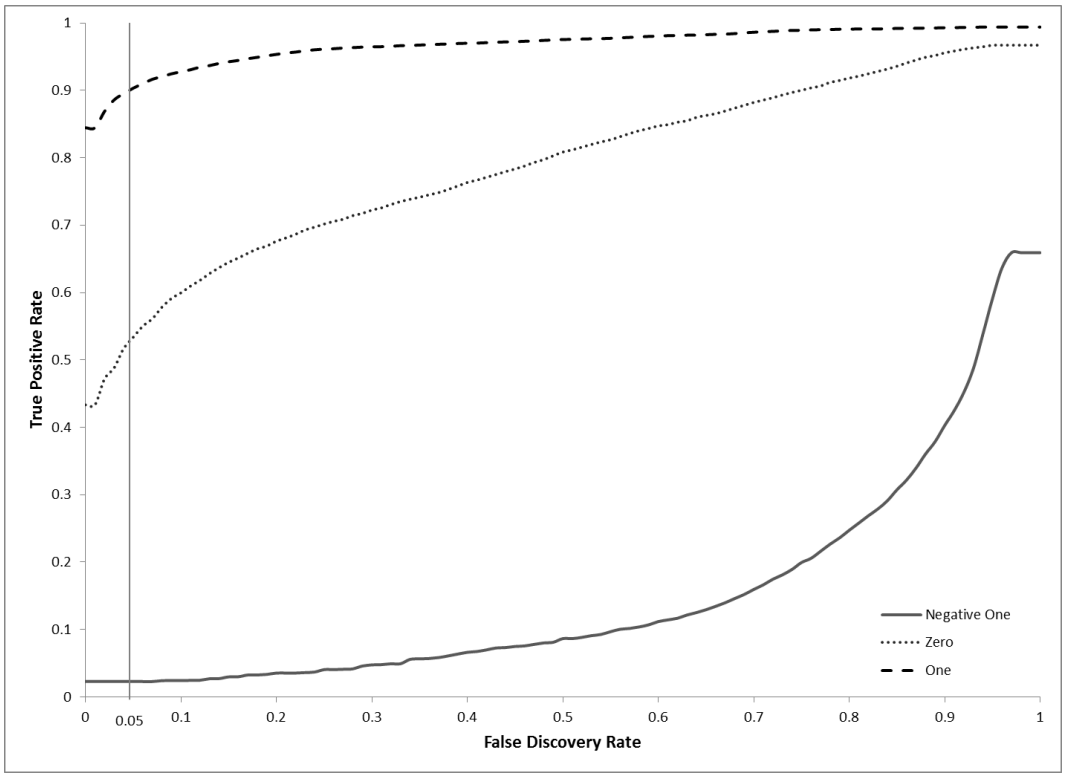

Figure 40: ROC Curves, CMP/NRM, true person parameters, shift error length 7

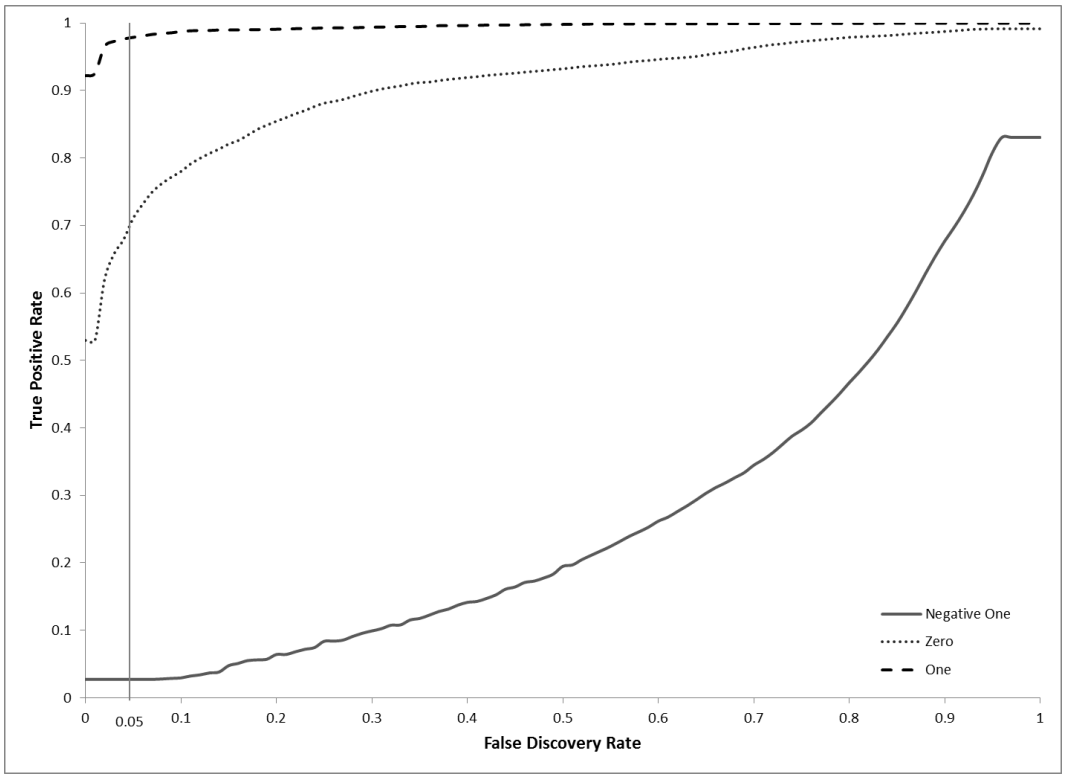

Figure 41: ROC Curves, SCIP/3PL, true person parameters, shift error length 7

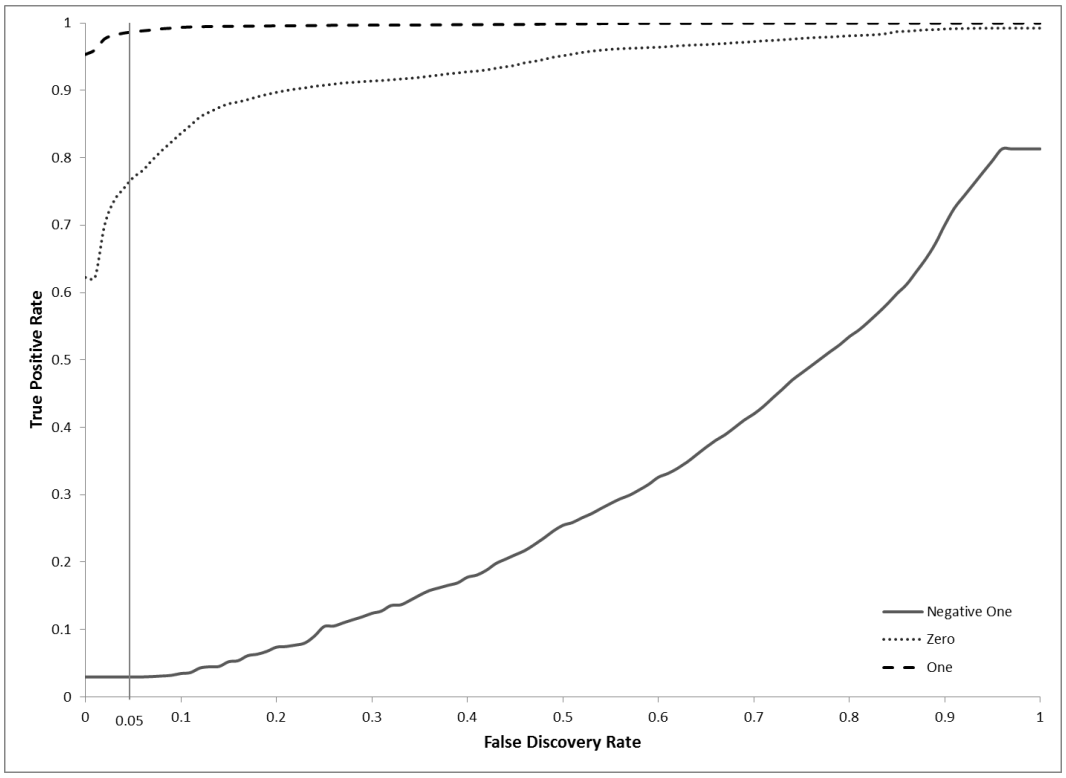

Figure 42: ROC Curves, SCIP/NRM, true person parameters, shift error length 7

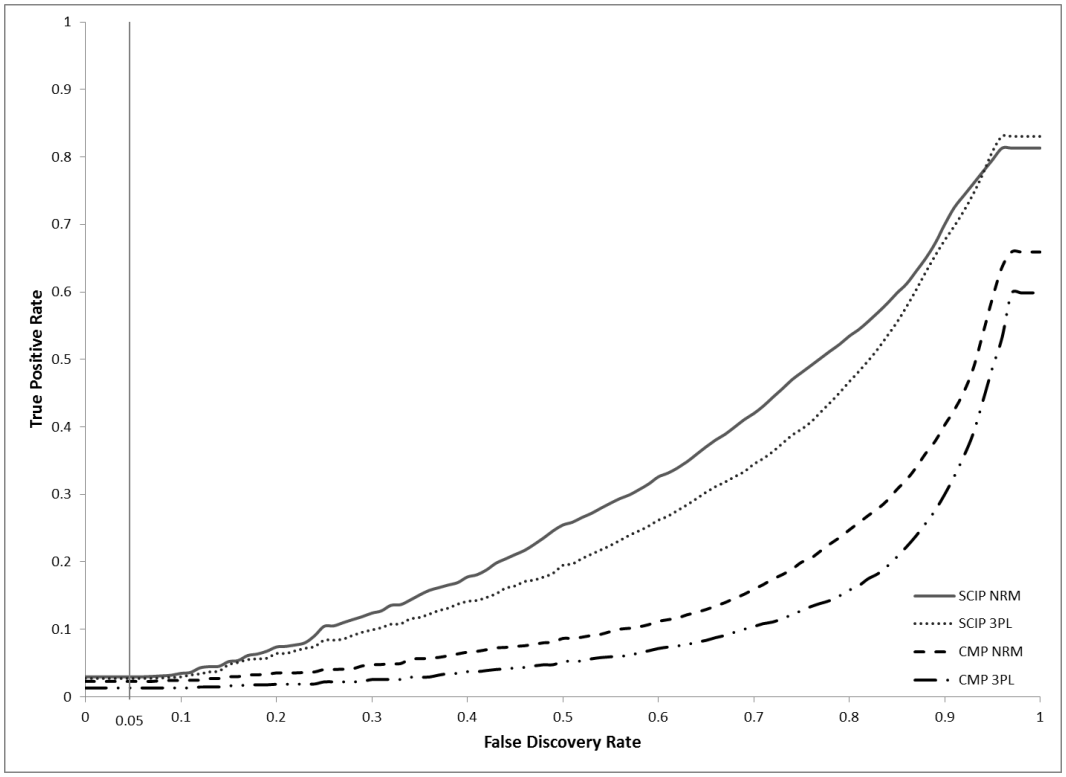

Figure 43: ROC Curves, all methods, true person parameters = -1, shift error length 7

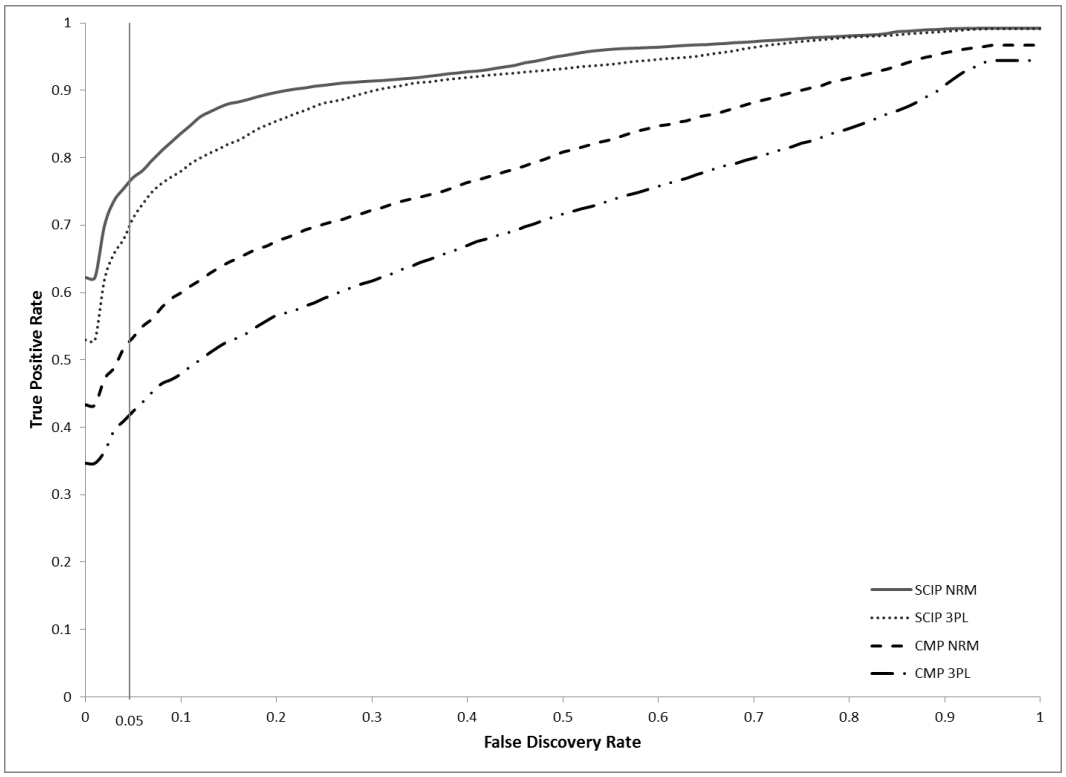

Figure 44: ROC Curves, all methods, true person parameters = 0, shift error length 7

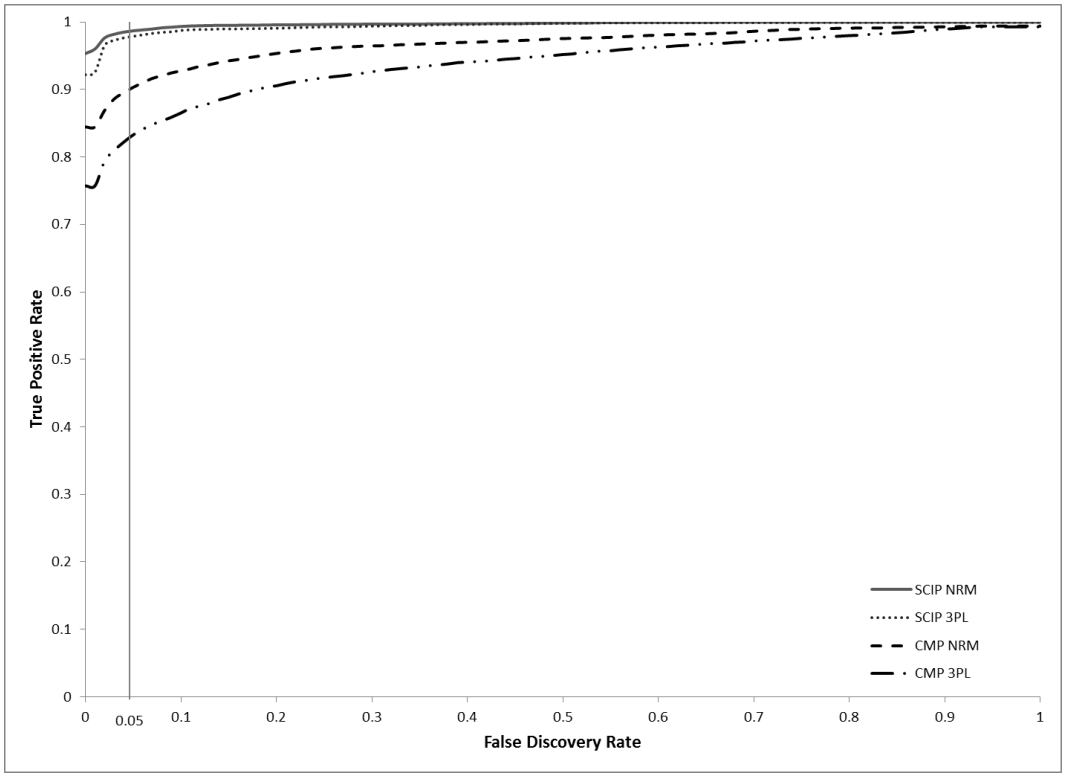

Figure 45: ROC Curves, all methods , true person parameters = 1, shift error length 7

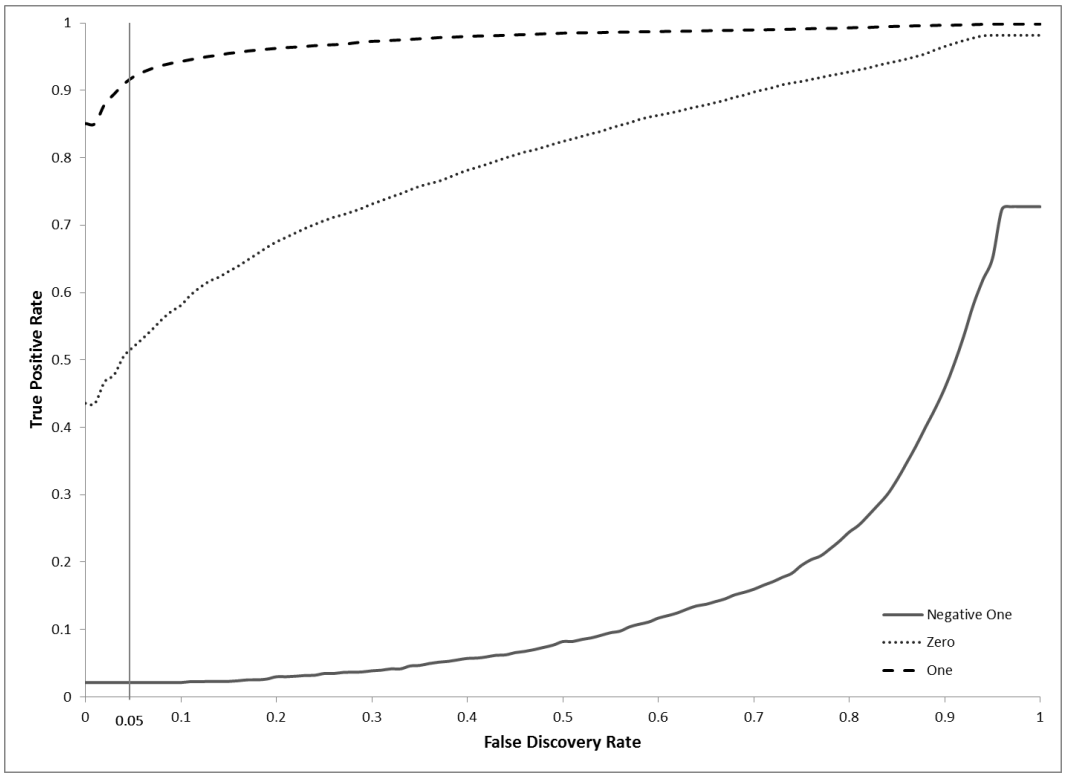

Figure 46: ROC Curves, CMP/3PL, true person parameters, shift error length 10

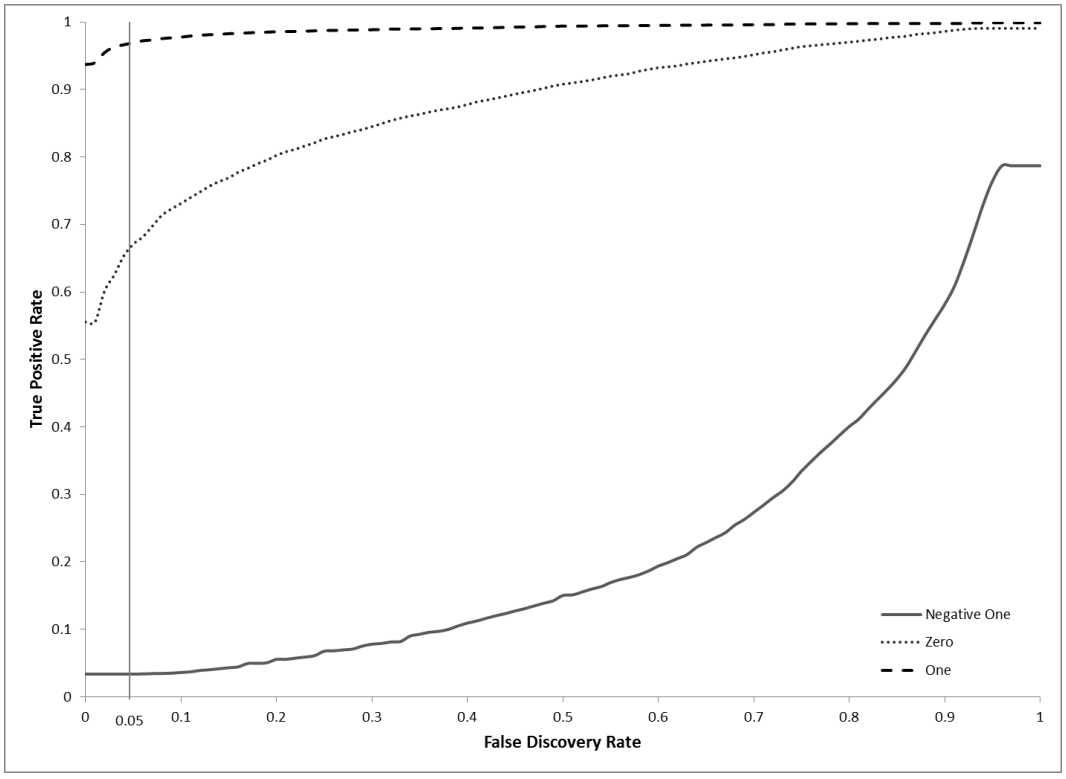

Figure 47: ROC Curves, CMP/NRM, true person parameters, shift error length 10

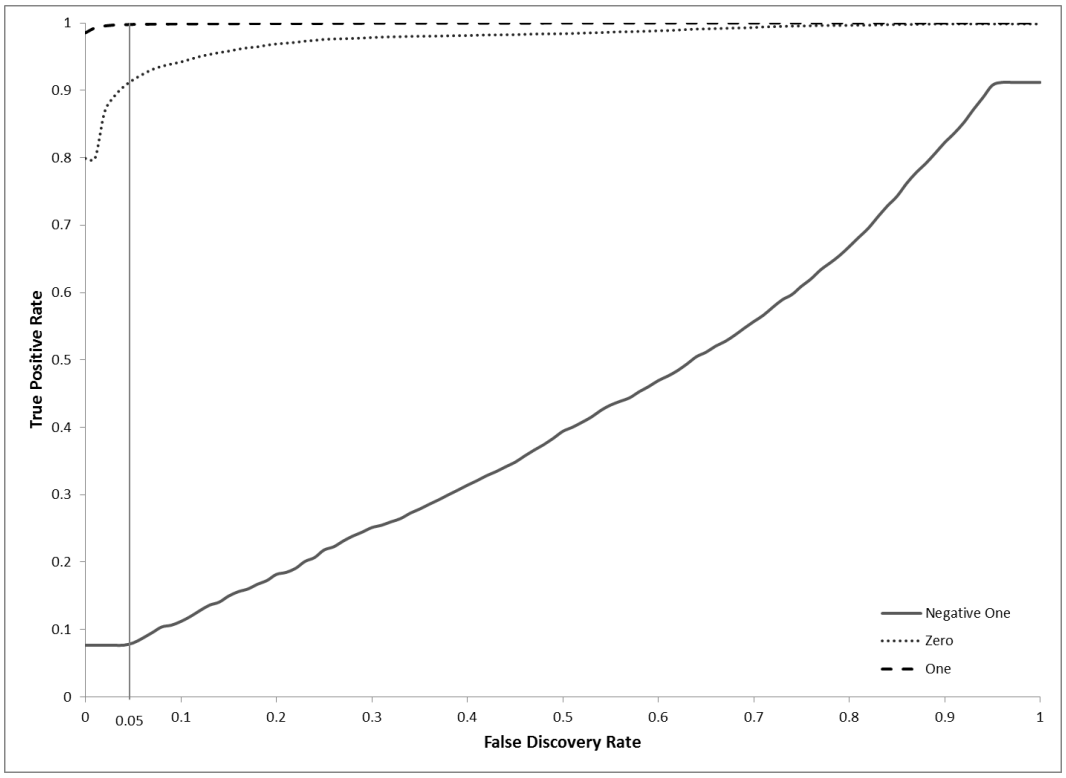

Figure 48: ROC Curves, SCIP/3PL, true person parameters, shift error length 10

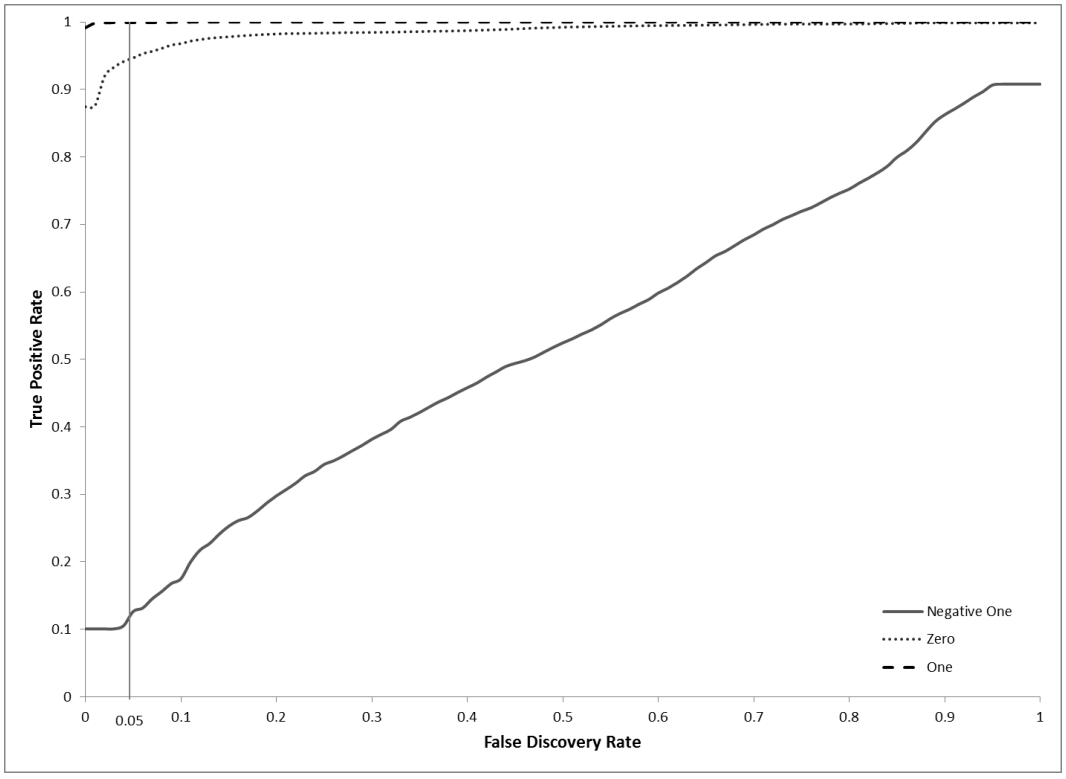

Figure 49: ROC Curves, SCIP/NRM, true person parameters, shift error length 10

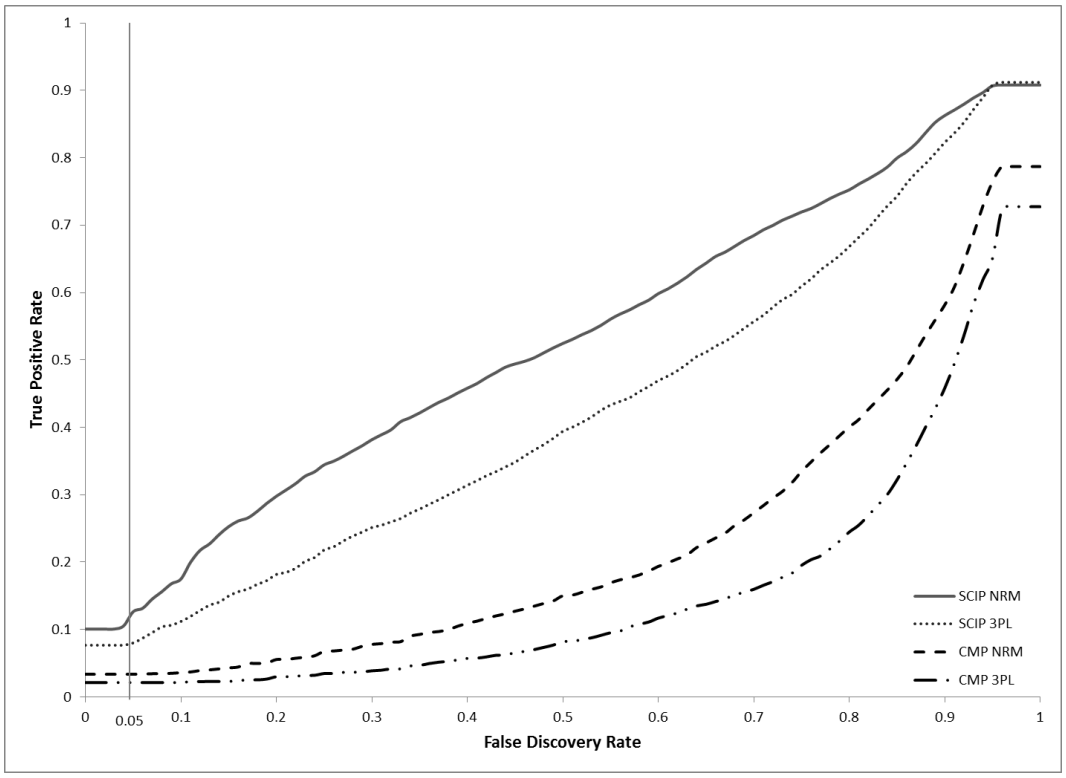

Figure 50: ROC Curves, all methods, true person parameters = -1, shift error length 10

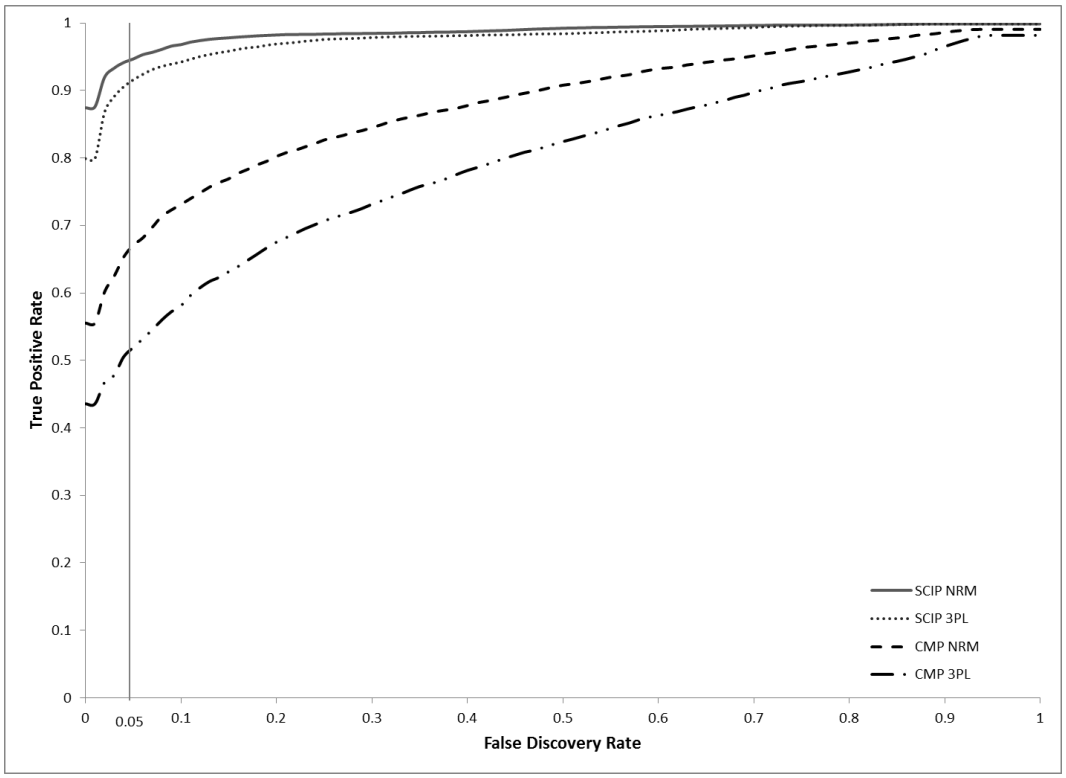

Figure 51: ROC Curves, all methods, true person parameters = 0, shift error length 10

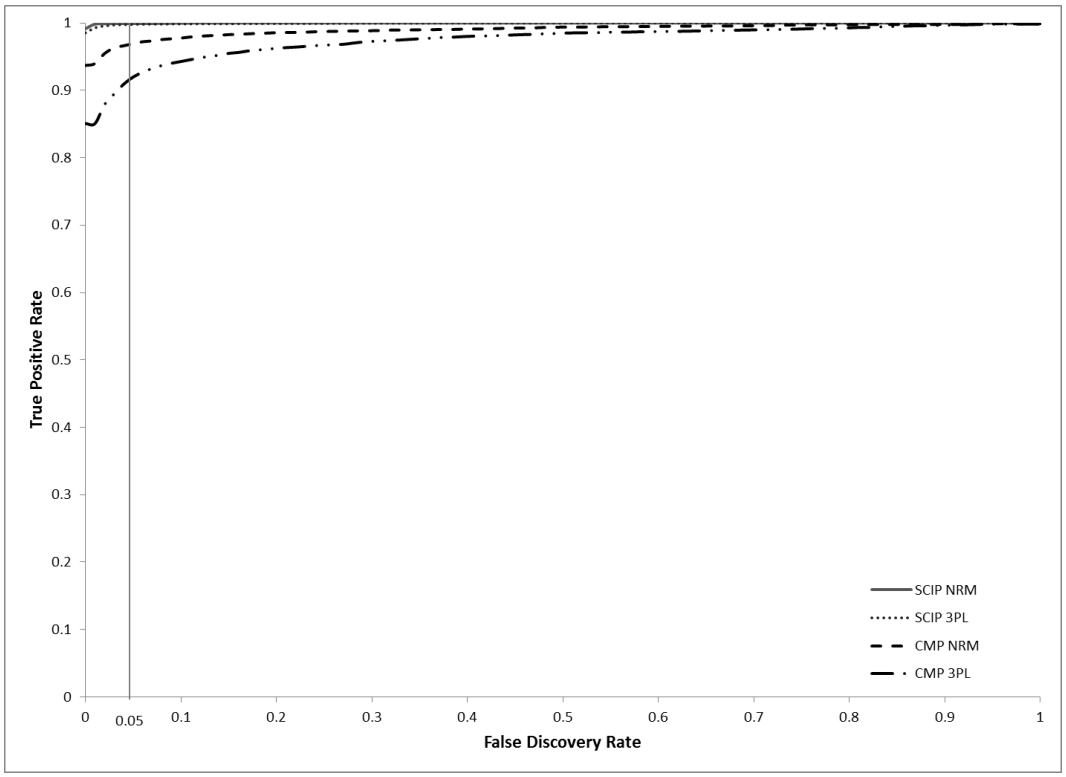

Figure 52: ROC Curves, all methods , true person parameters = 1, shift error length 10

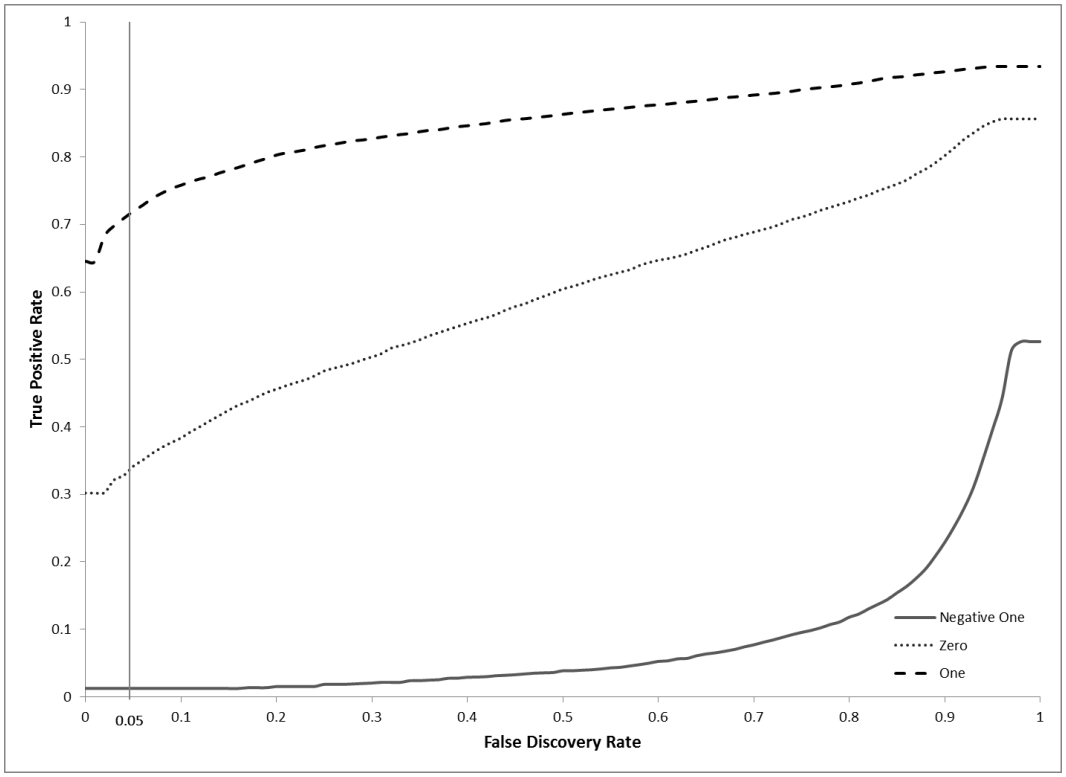

Figure 53: ROC Curves, CMP/3PL, true person parameters, mixed-length shifts

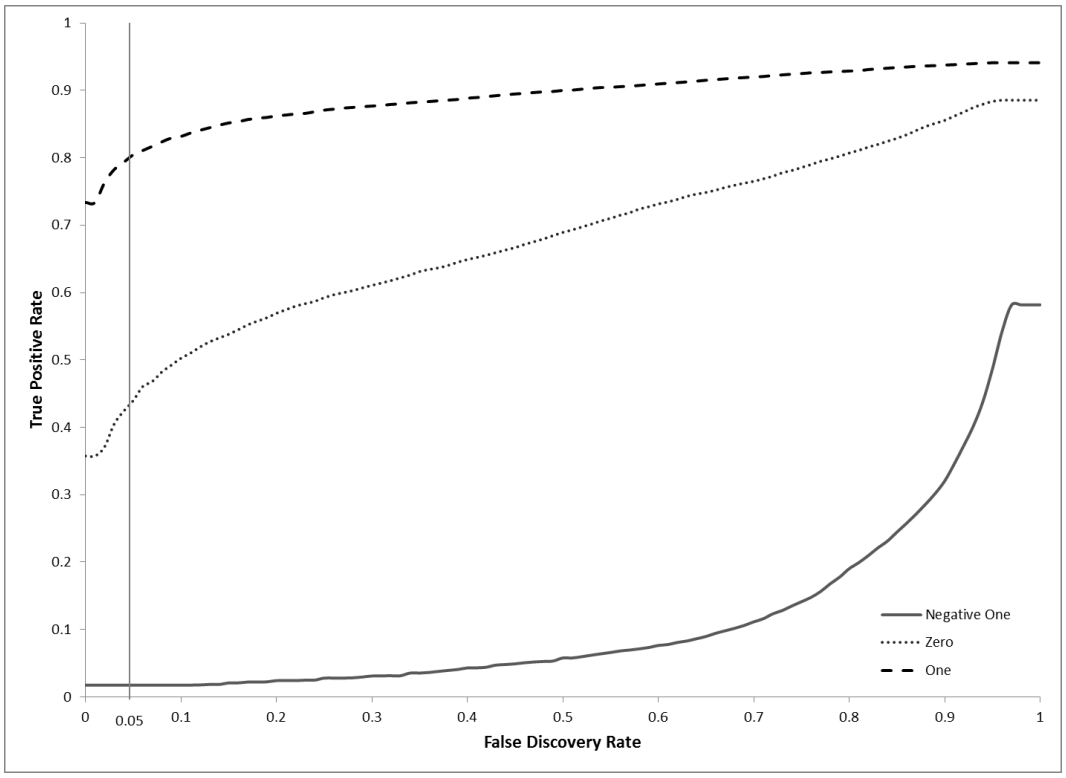

Figure 54: ROC Curves, CMP/NRM, true person parameters, mixed-length shifts

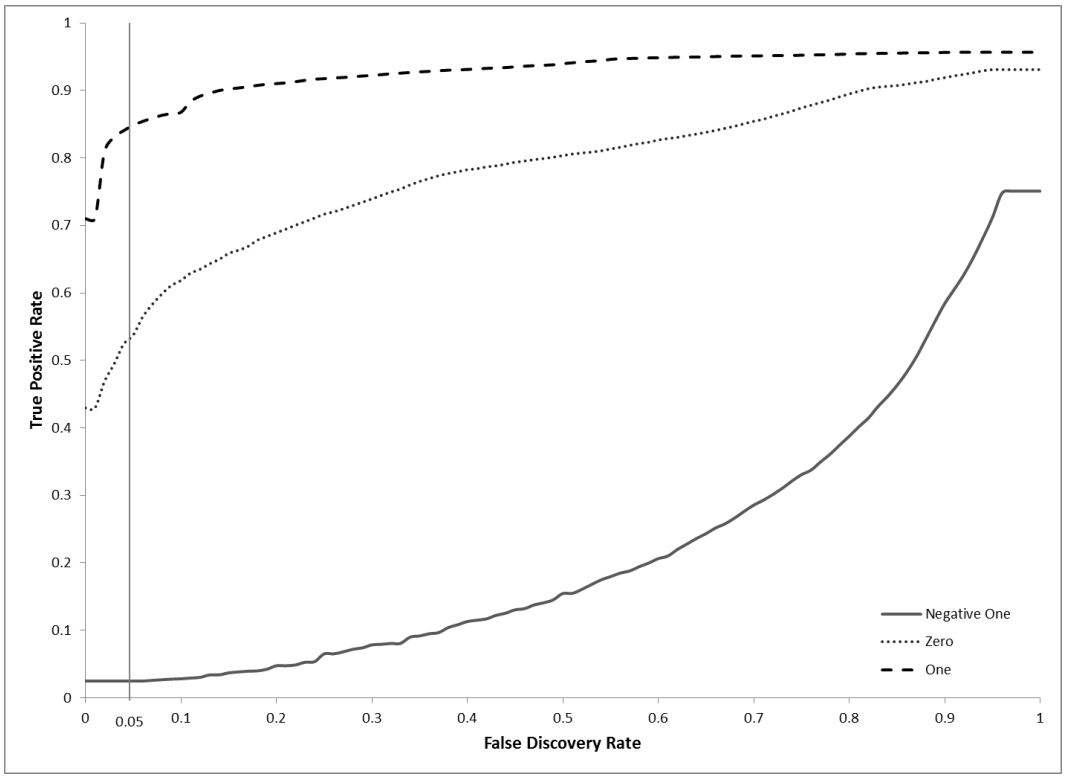

Figure 55: ROC Curves, SCIP/3PL, true person parameters, mixed-length shifts

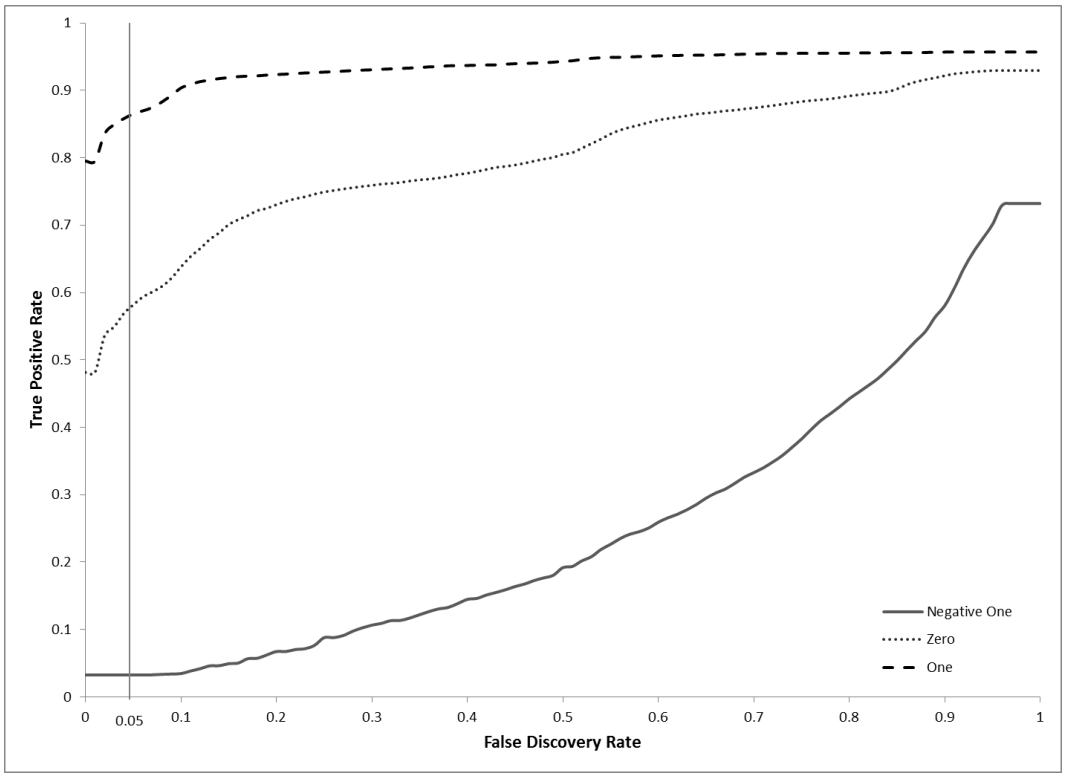

Figure 56: ROC Curves, SCIP/NRM, true person parameters, mixed-length shifts

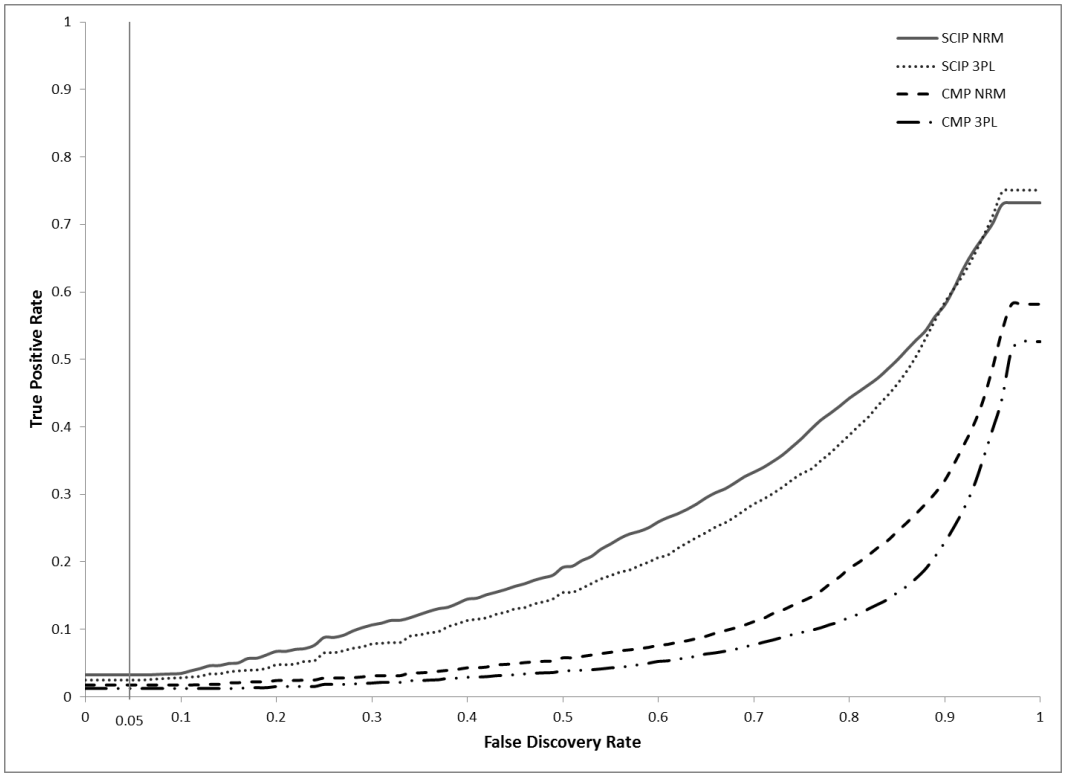

Figure 57: ROC Curves, all methods, true person parameters = -1, mixed-length shifts

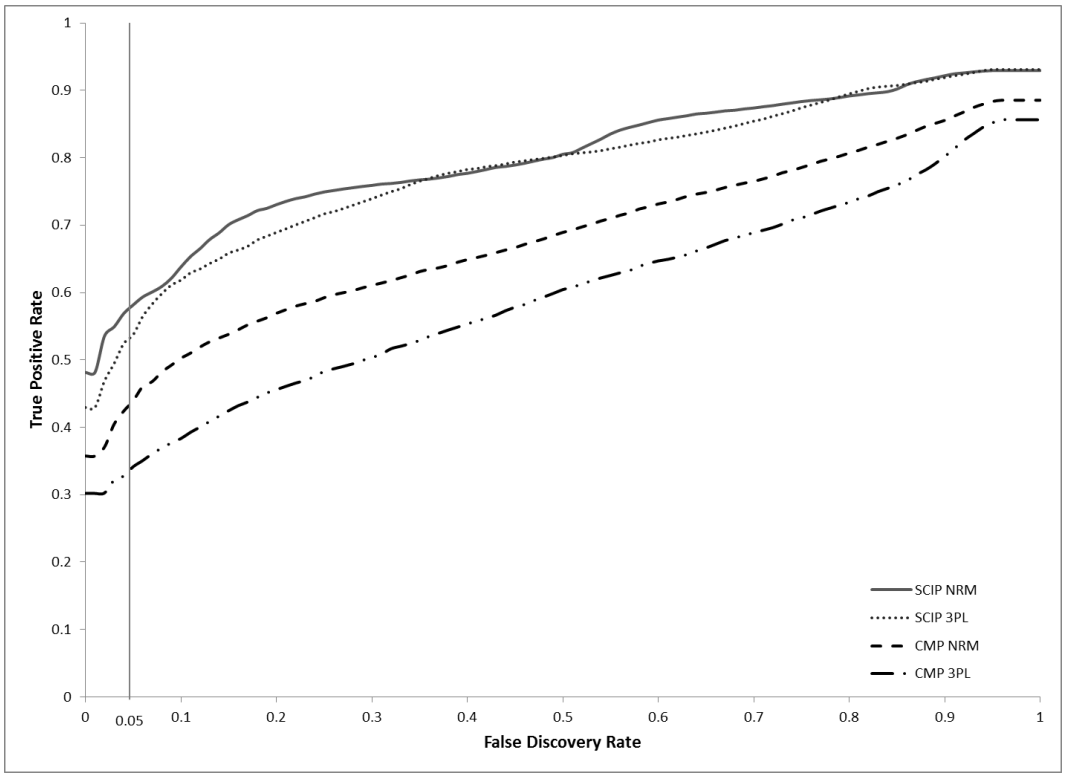

Figure 58: ROC Curves, all methods, true person parameters = 0, mixed-length shifts

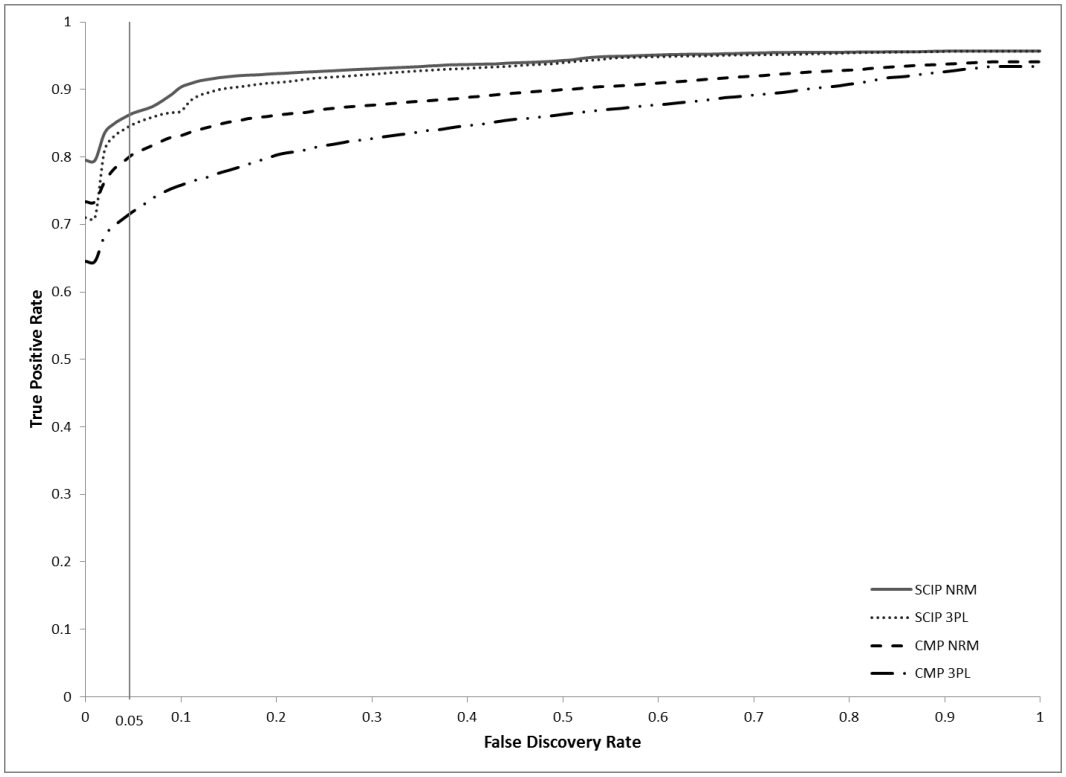

Figure 59: ROC Curves, all methods , true person parameters = 1, mixed-length shifts

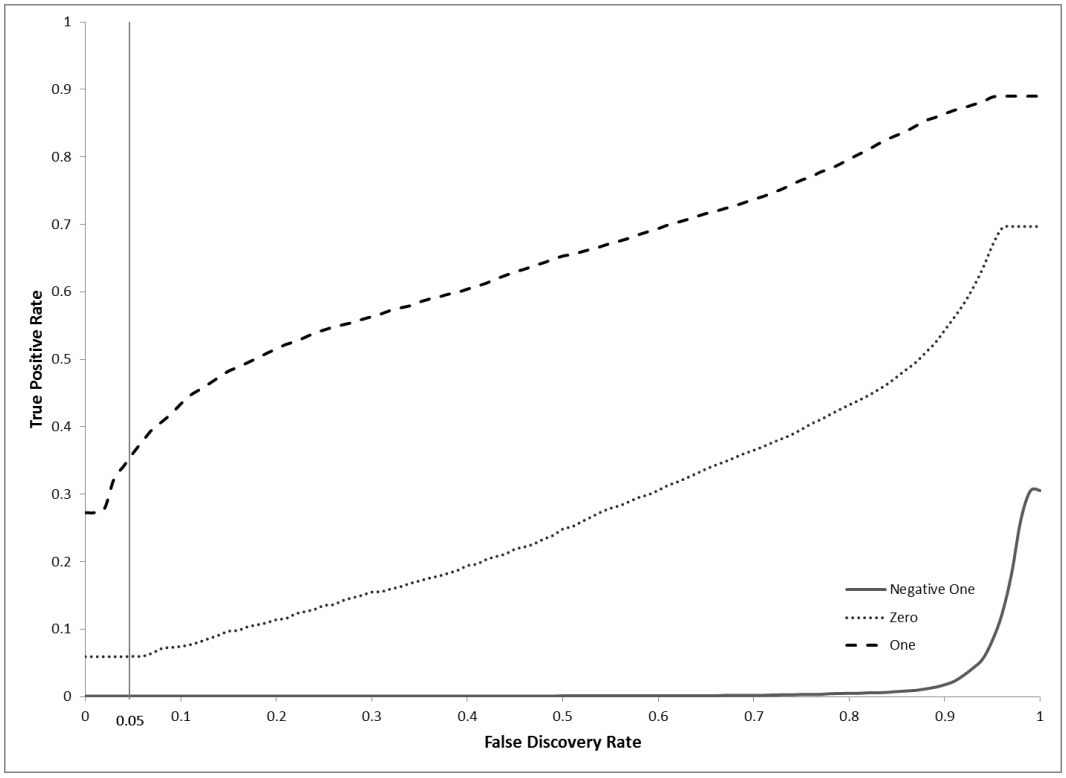

Figure 60: ROC Curves, CMP/3PL, estimated person parameters, shift length 3

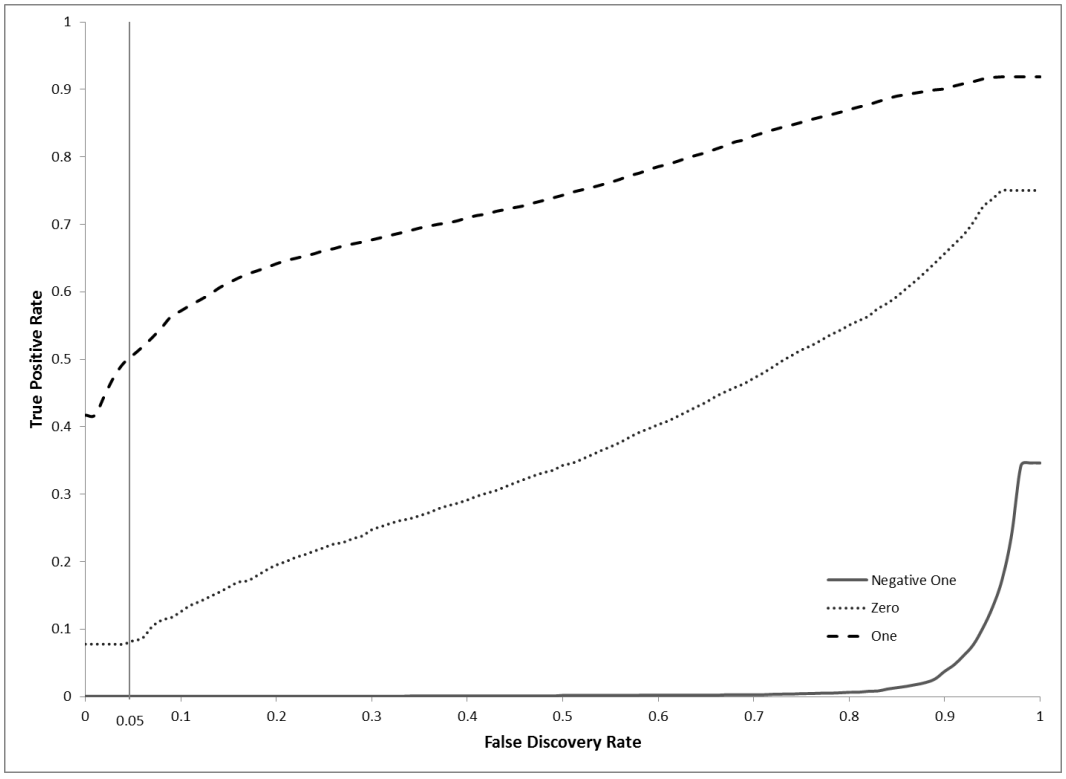

Figure 61: ROC Curves, CMP/NRM, estimated person parameters, shift length 3

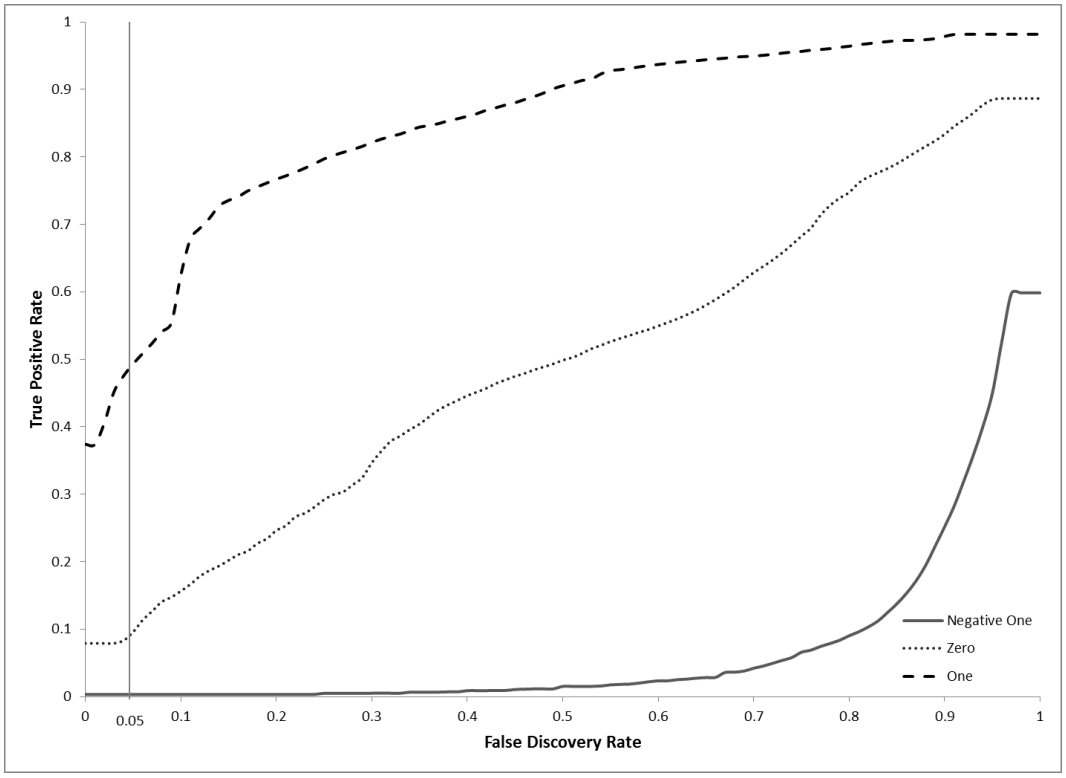

Figure 62: ROC Curves, SCIP/3PL, estimated person parameters, shift length 3

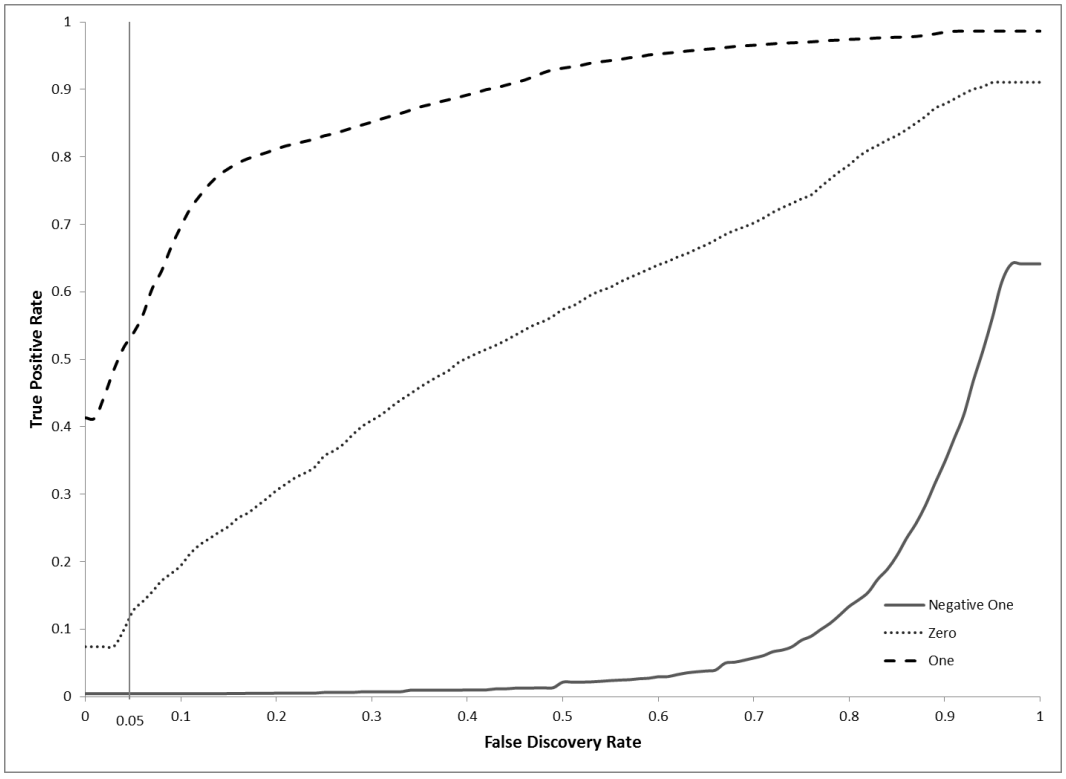

Figure 63: ROC Curves, SCIP/NRM, estimated person parameters, shift length 3

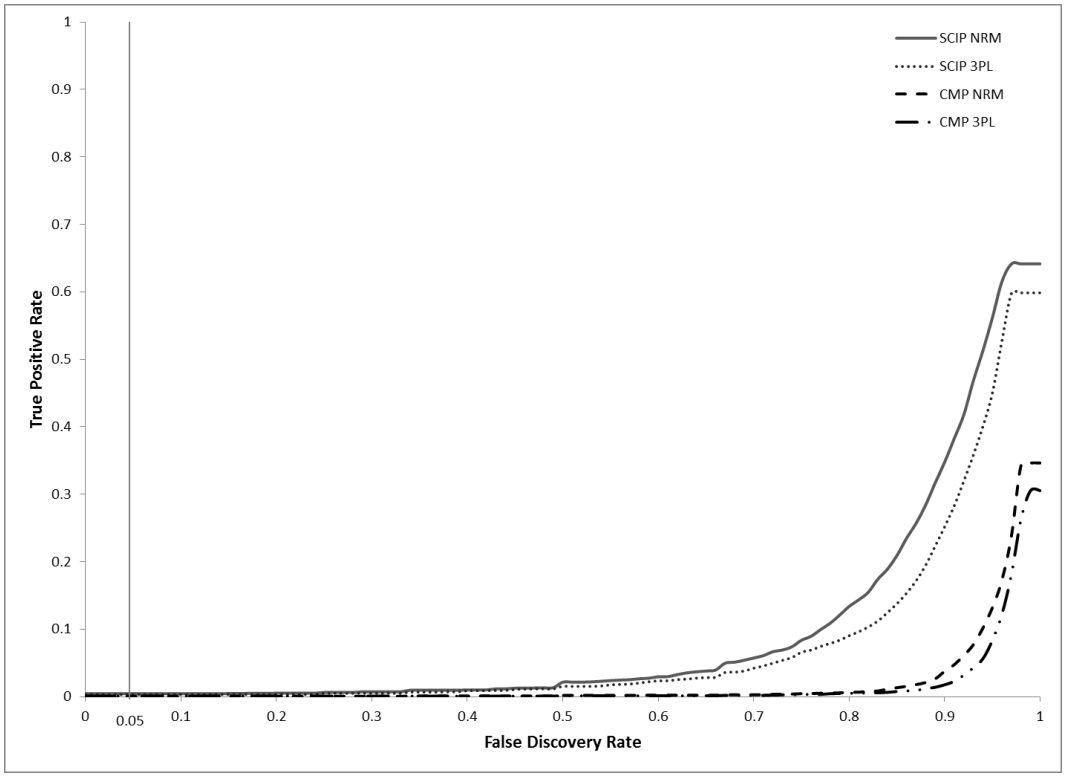

Figure 64: ROC Curves, all methods, estimated person parameters = -1, shift length 3

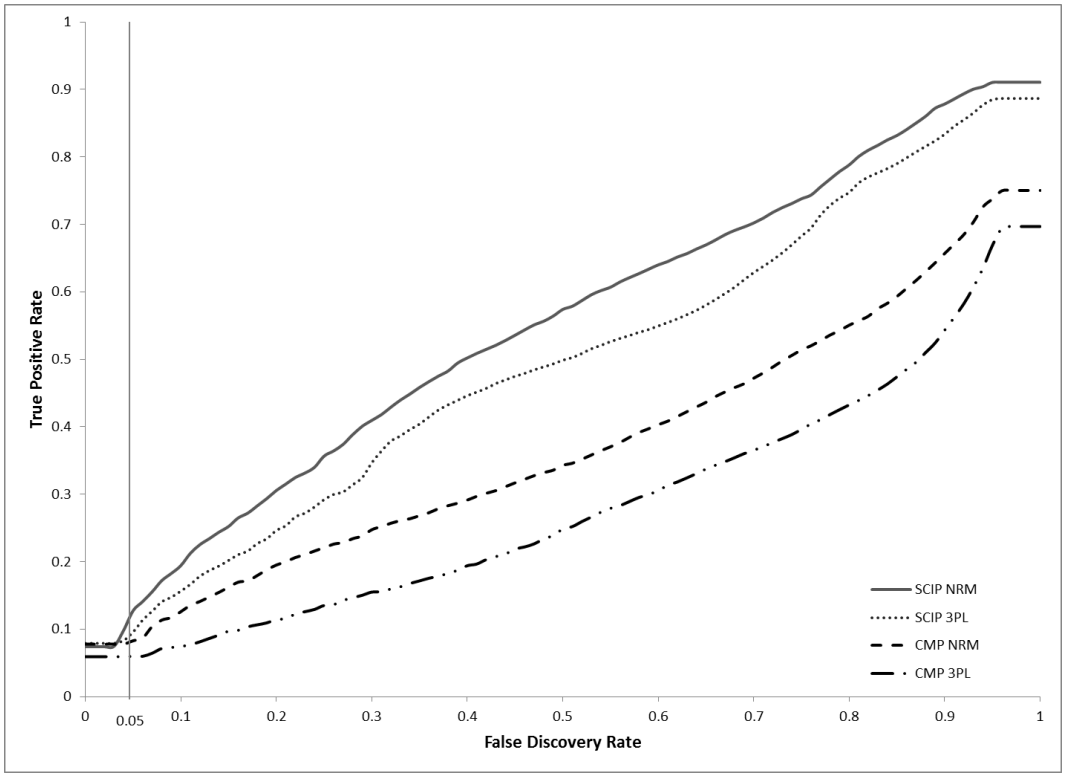

Figure 65: ROC Curves, all methods, estimated person parameters = 0, shift length 3

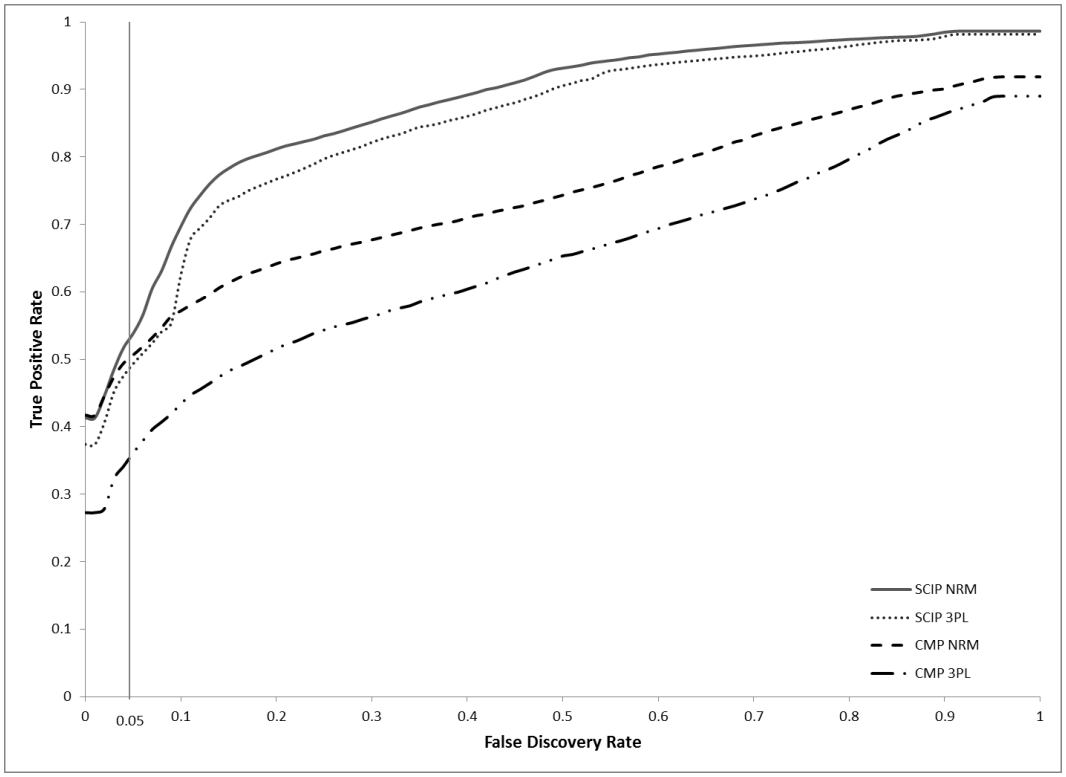

Figure 66: ROC Curves, all methods , estimated person parameters = 1, shift length 3

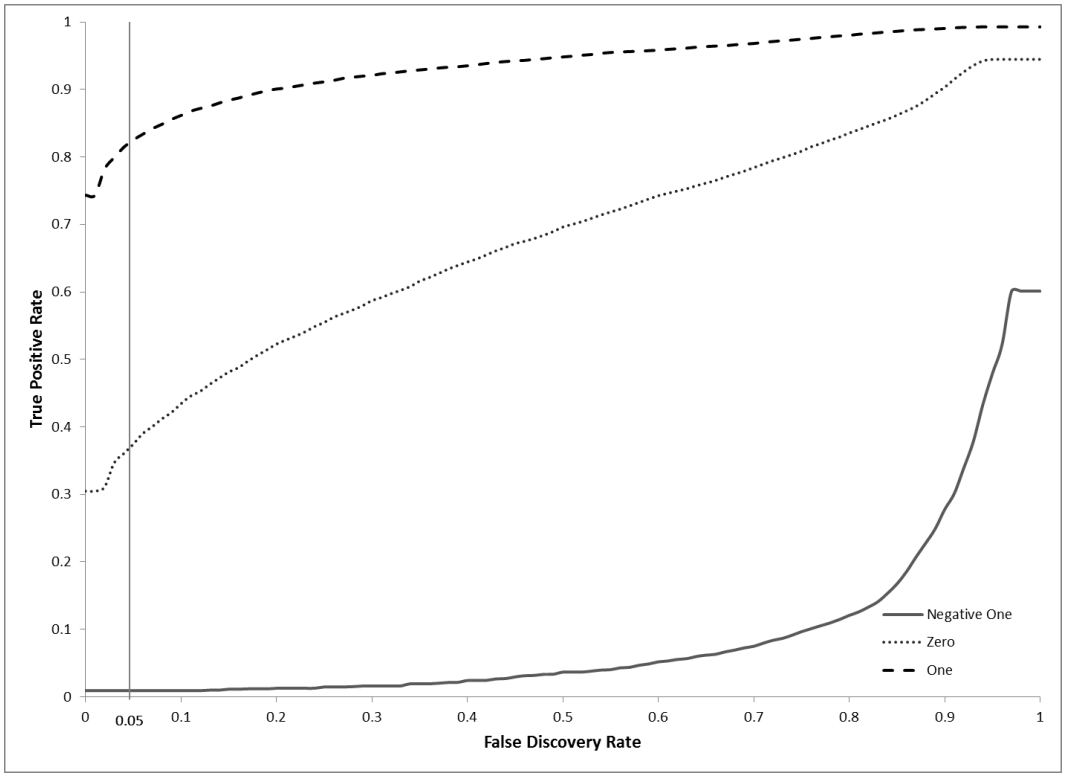

Figure 67: ROC Curves, CMP/3PL, estimated person parameters, shift length 7

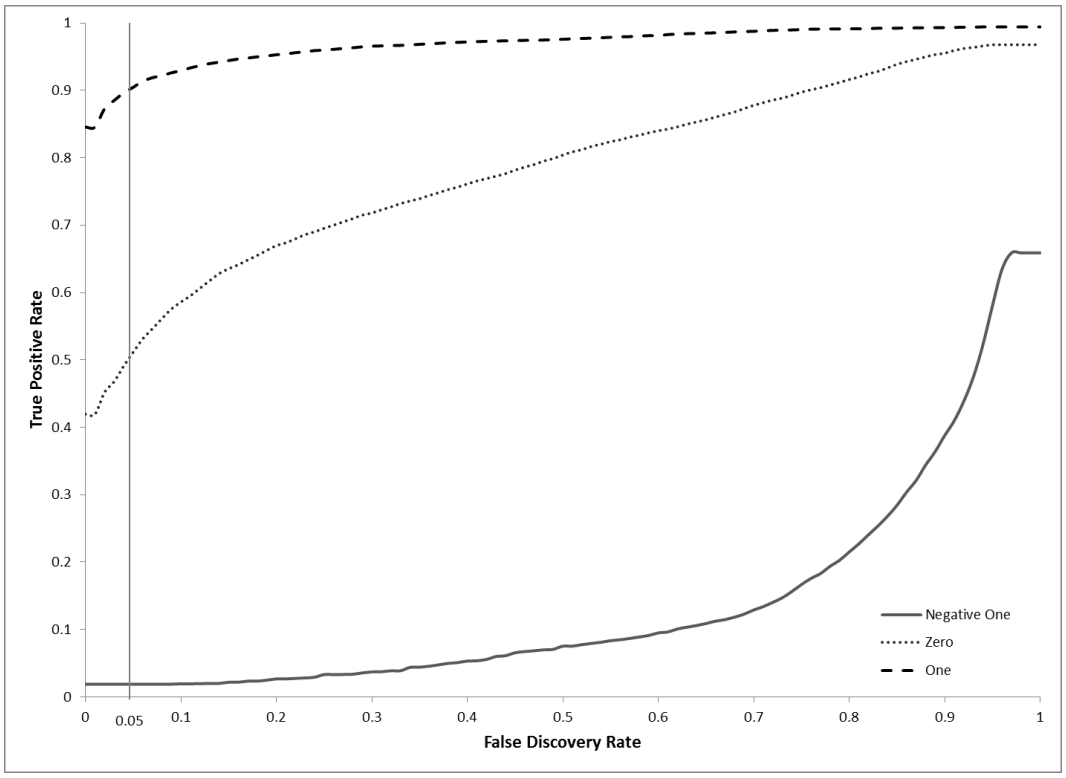

Figure 68: ROC Curves, CMP/NRM, estimated person parameters, shift length 7

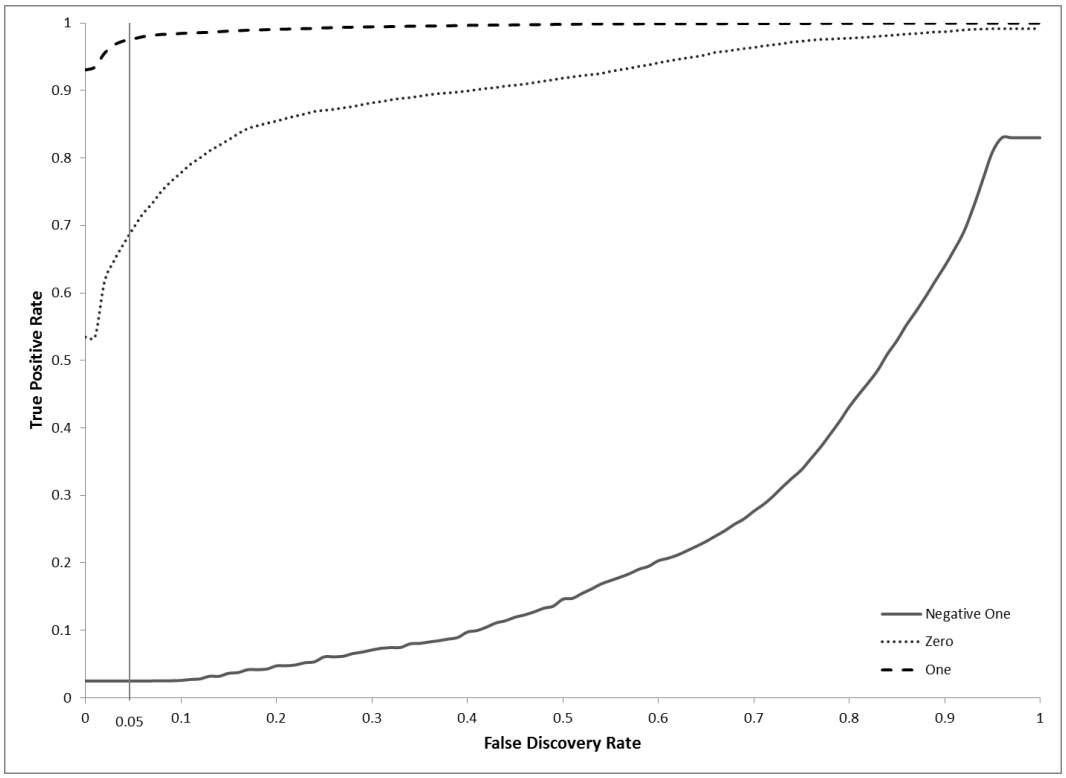

Figure 69: ROC Curves, SCIP/3PL, estimated person parameters, shift length 7

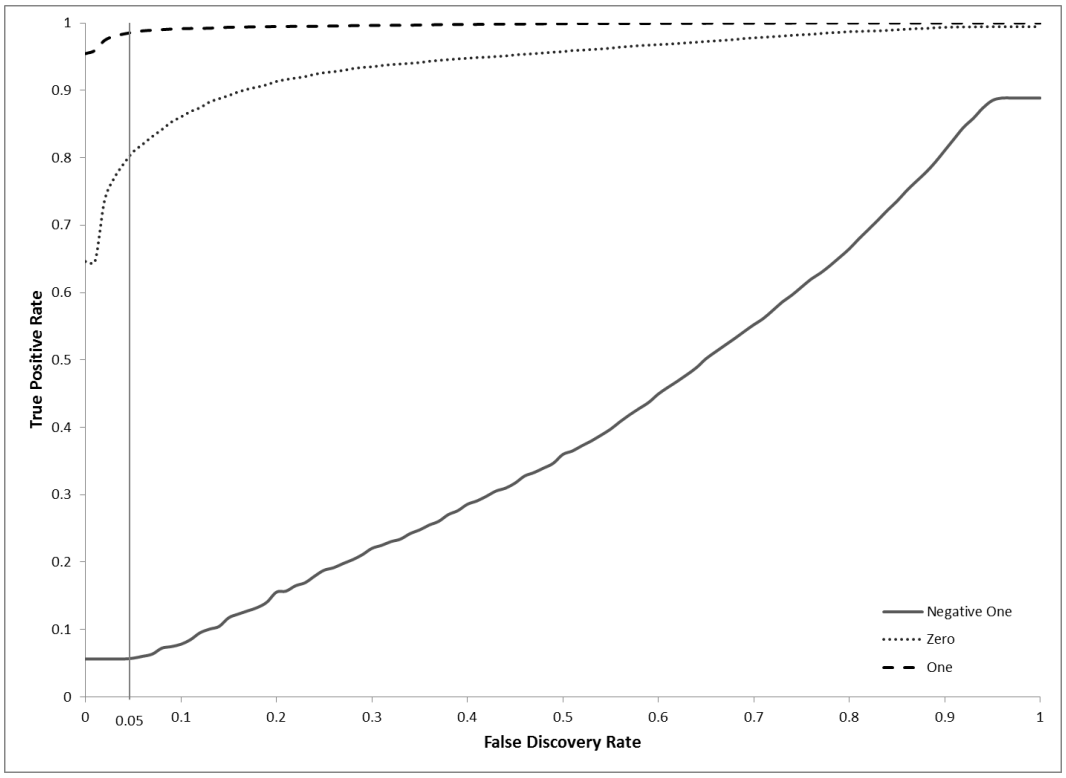

Figure 70: ROC Curves, SCIP/NRM, estimated person parameters, shift length 7

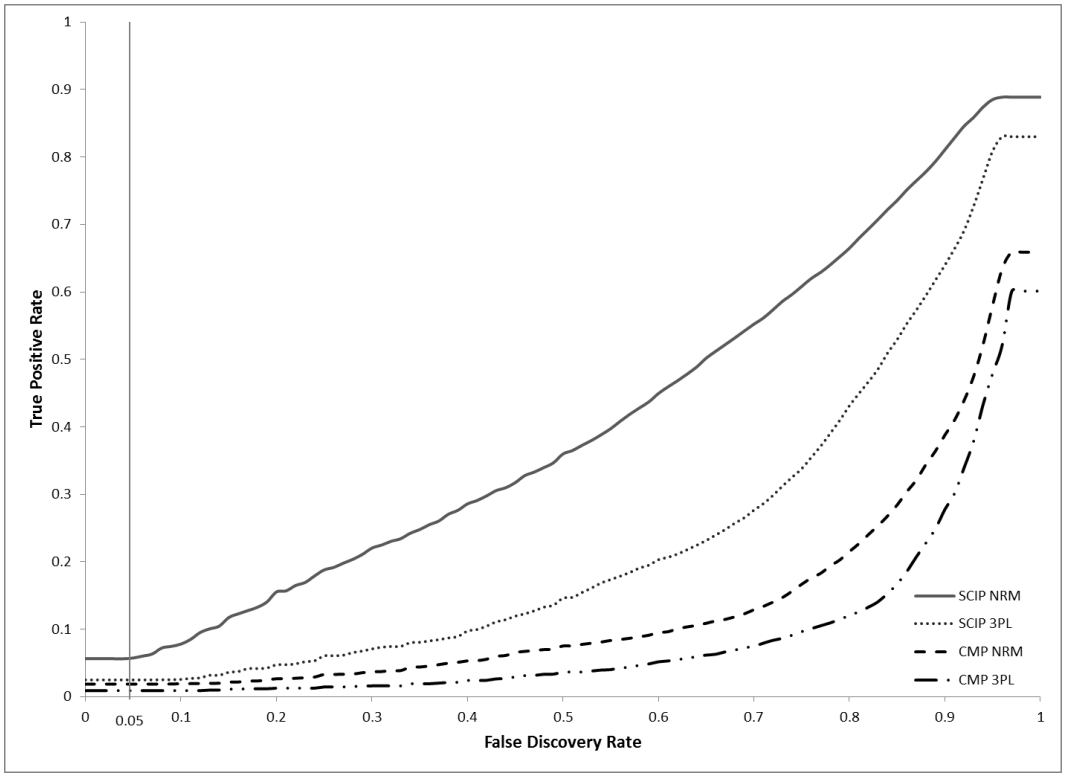

Figure 71: ROC Curves, all methods, estimated person parameters = -1, shift length 7

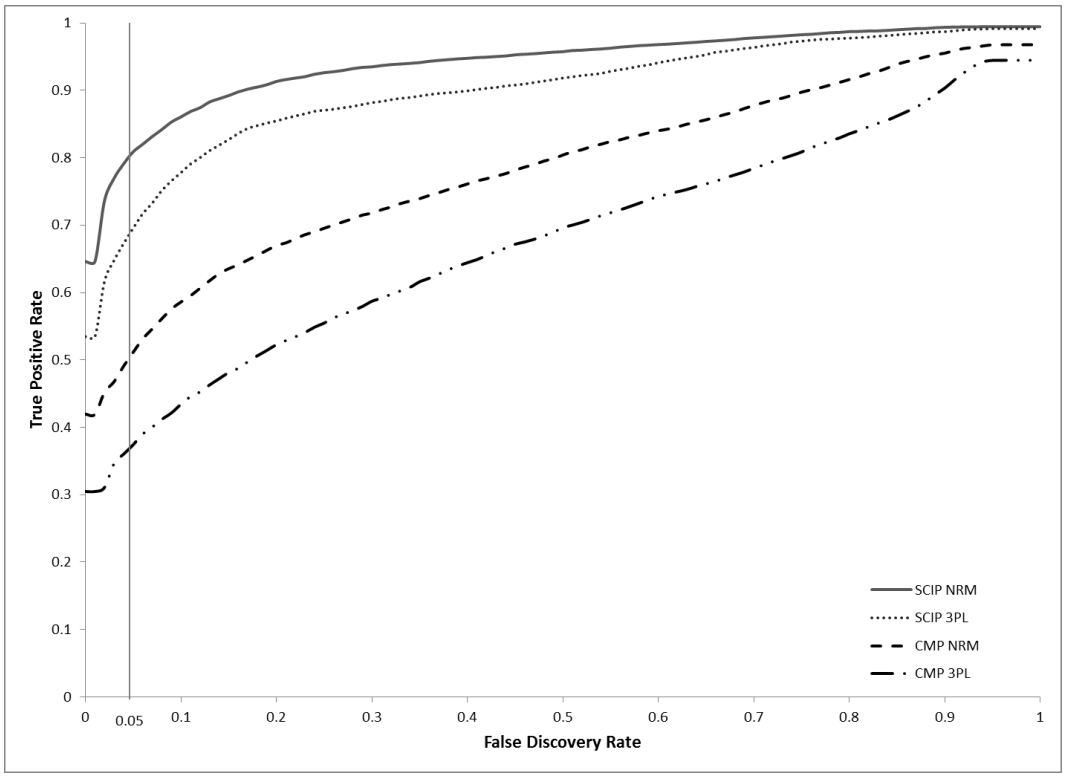

Figure 72: ROC Curves, all methods, estimated person parameters = 0, shift length 7

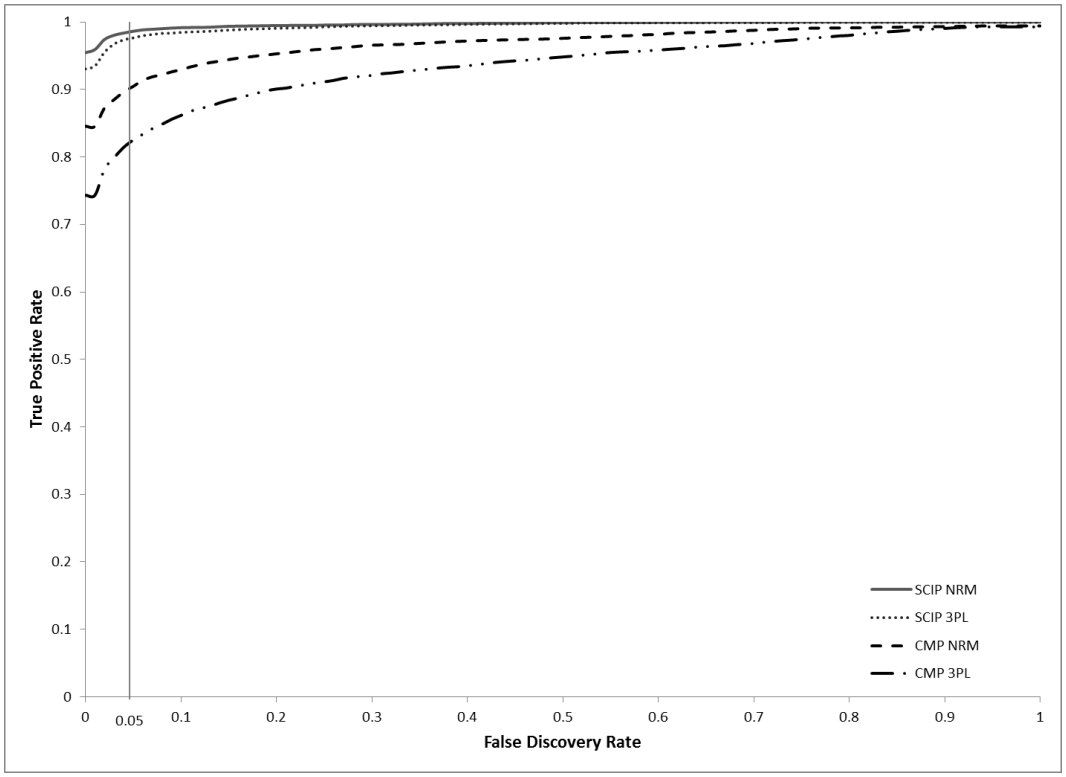

Figure 73: ROC Curves, all methods , estimated person parameters = 1, shift length 7

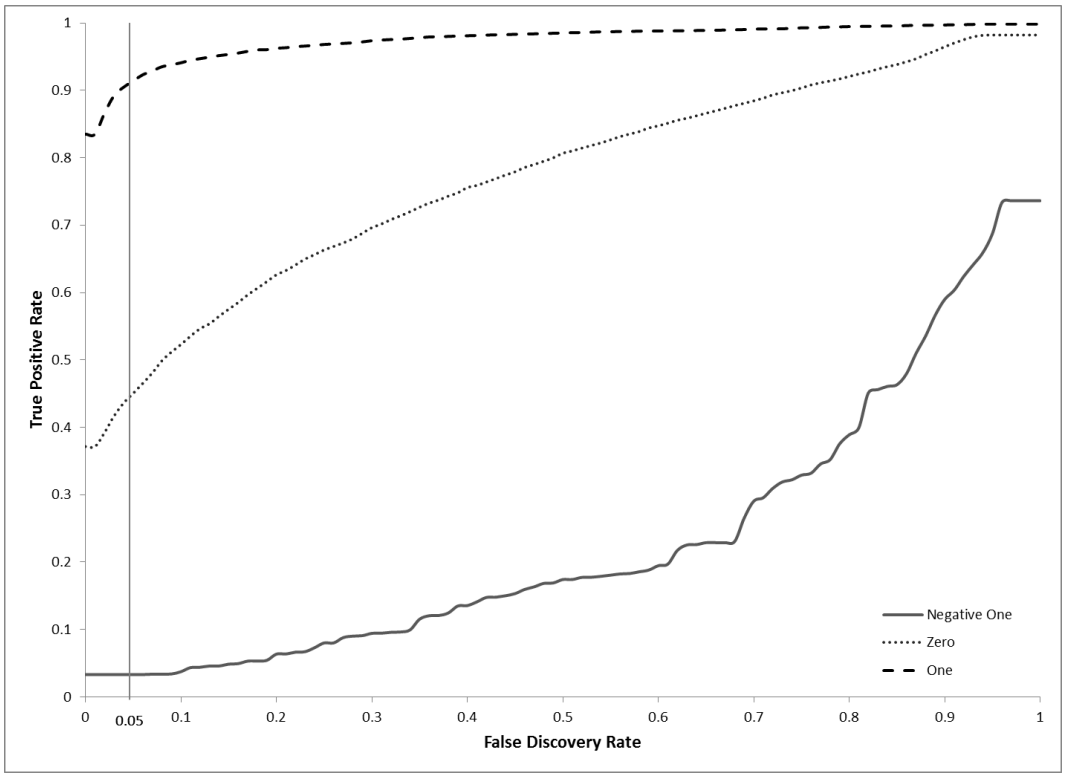

Figure 74: ROC Curves, CMP/3PL, estimated person parameters, shift length 10

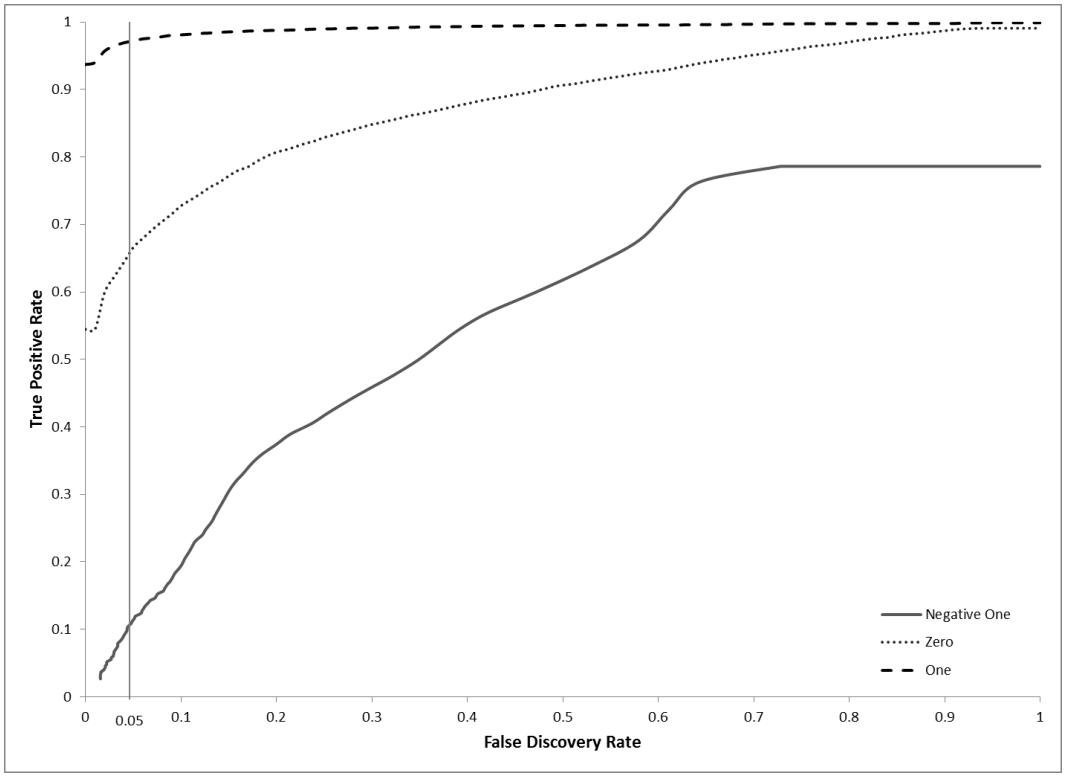

Figure 75: ROC Curves, CMP/NRM, estimated person parameters, shift length 10

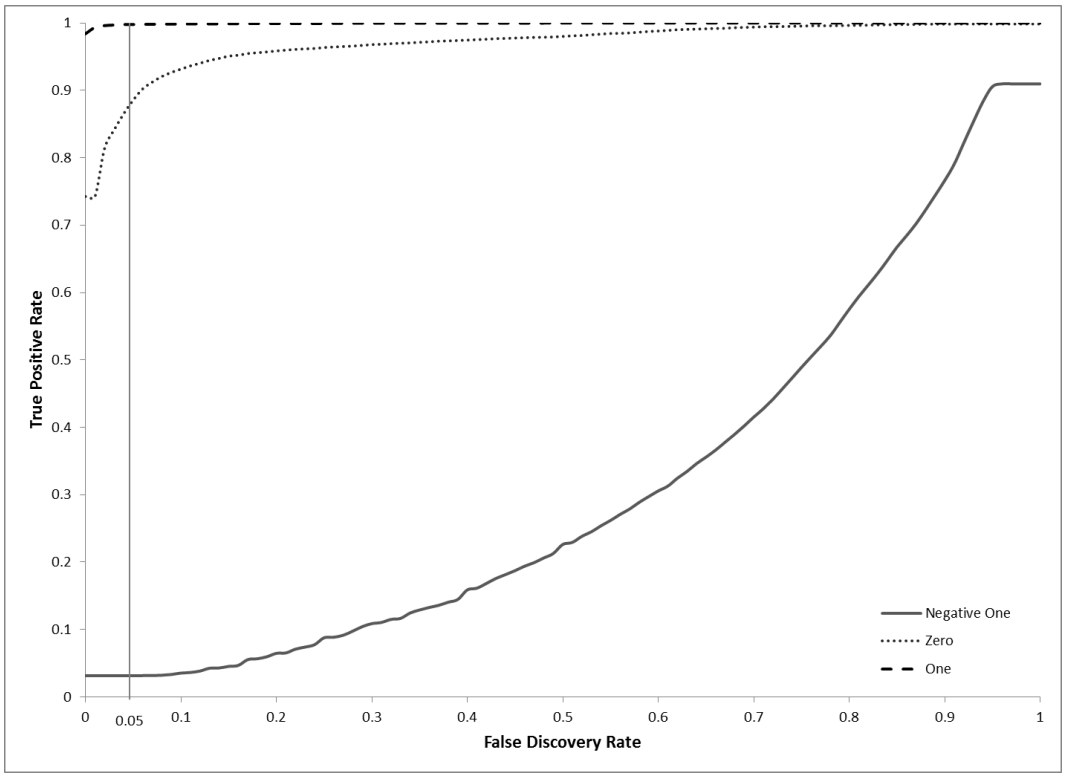

Figure 76: ROC Curves, SCIP/3PL, estimated person parameters, shift length 10

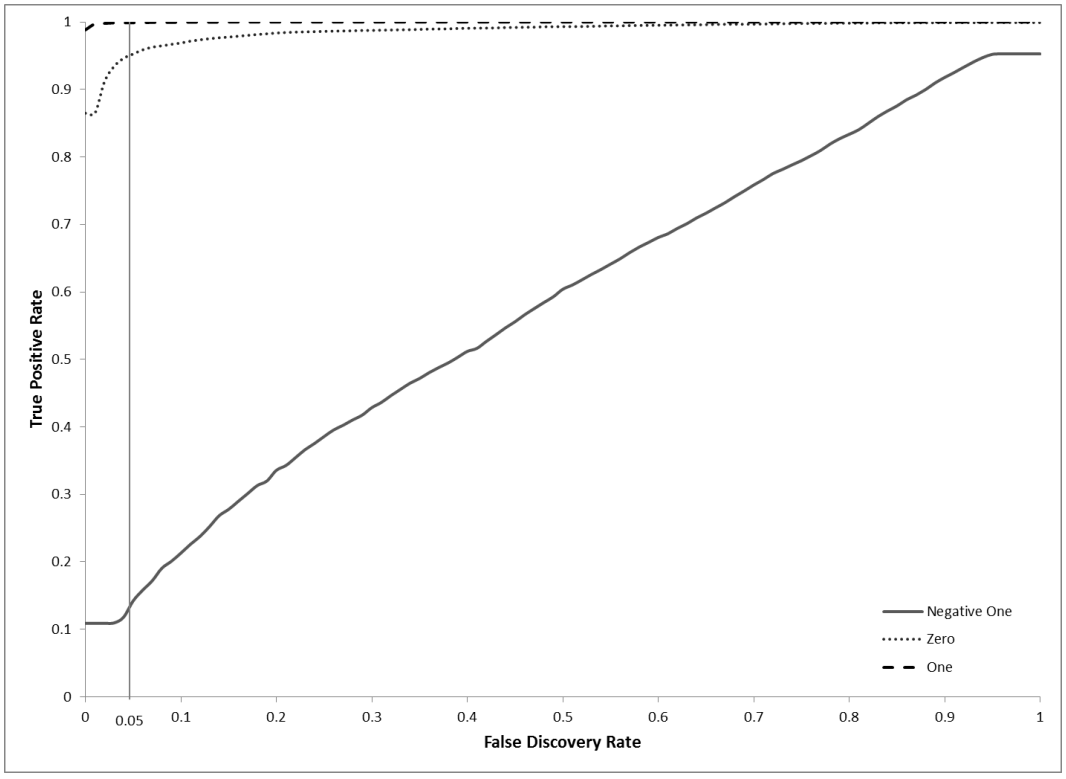

Figure 77: ROC Curves, SCIP/NRM, estimated person parameter levels, shift length 10

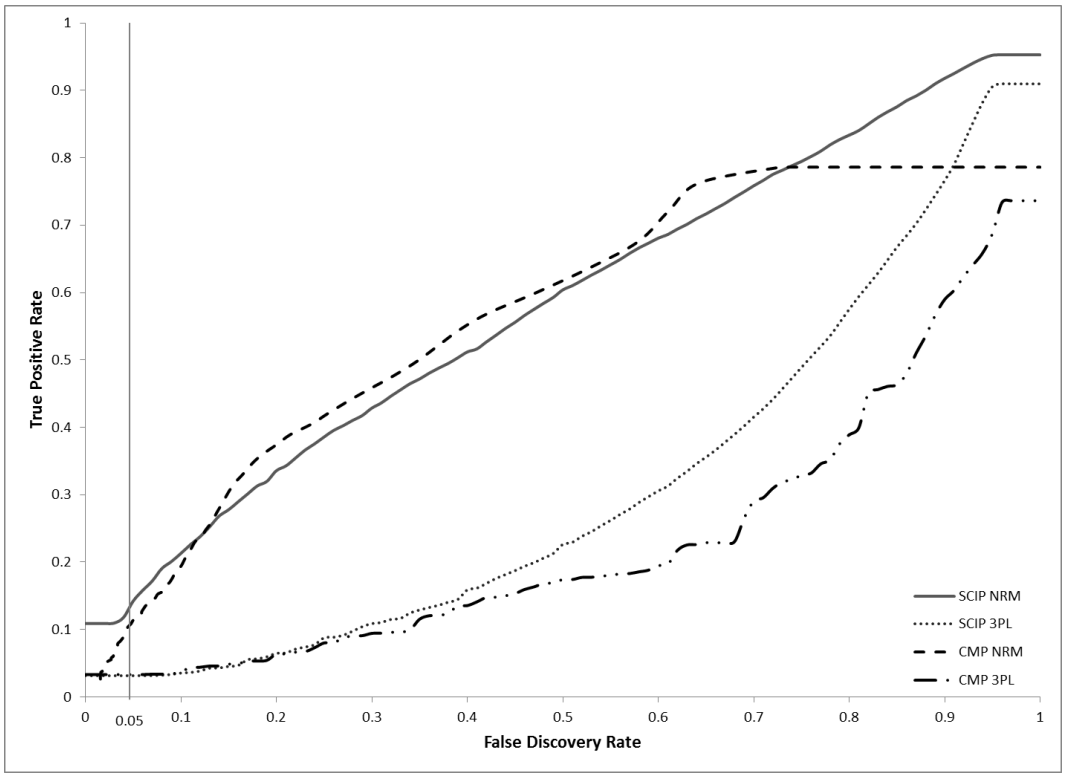

Figure 78: ROC Curves, all methods, estimated person parameters = -1, shift length 10

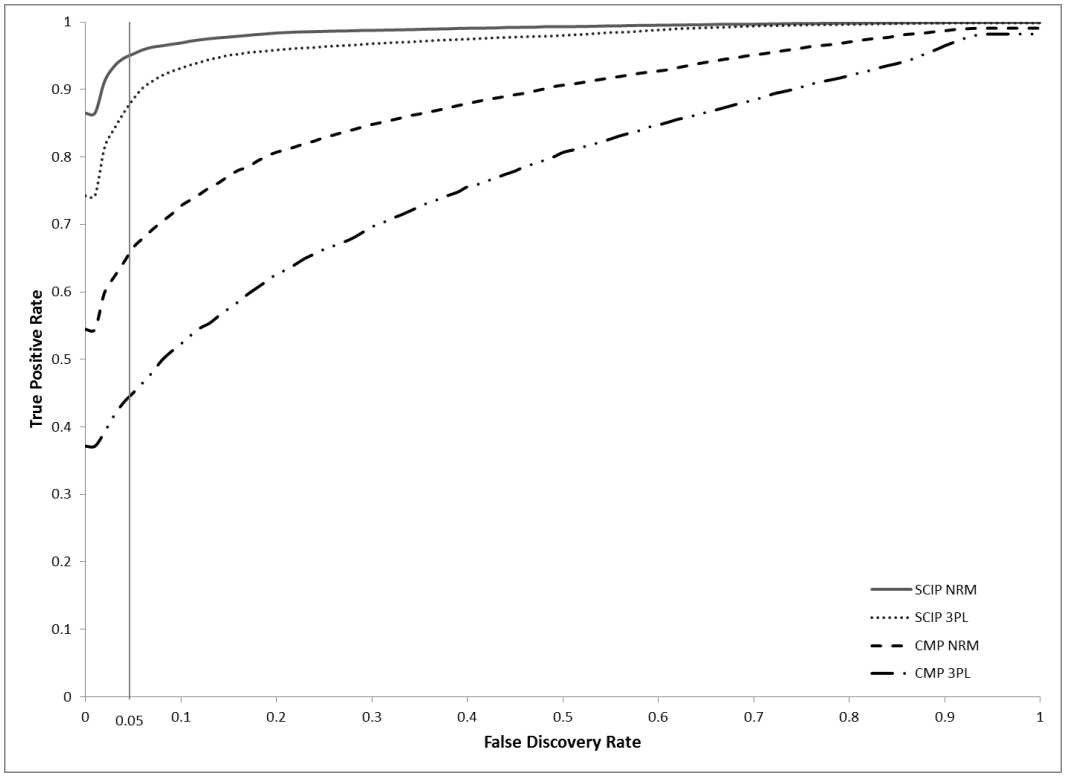

Figure 79: ROC Curves, all methods, estimated person parameters = 0, shift length 10

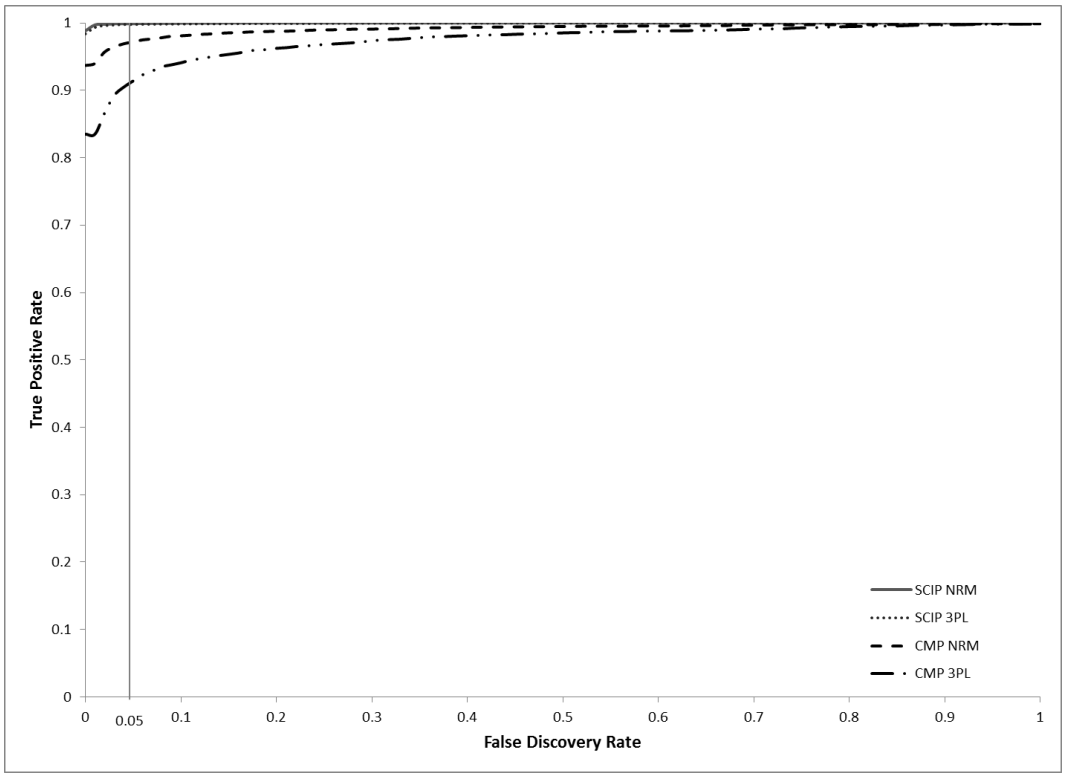

Figure 80: ROC Curves, all methods , estimated person parameters = 1, shift length 10

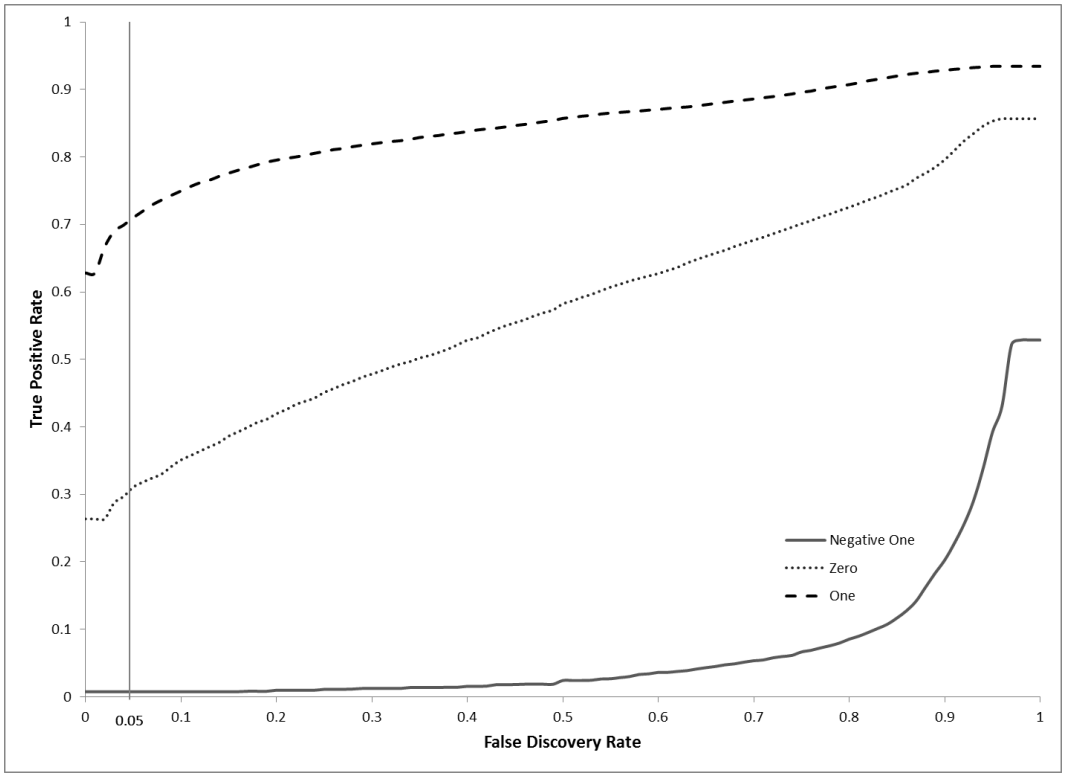

Figure 81: ROC Curves, CMP/3PL, estimated person parameters, mixed shifts

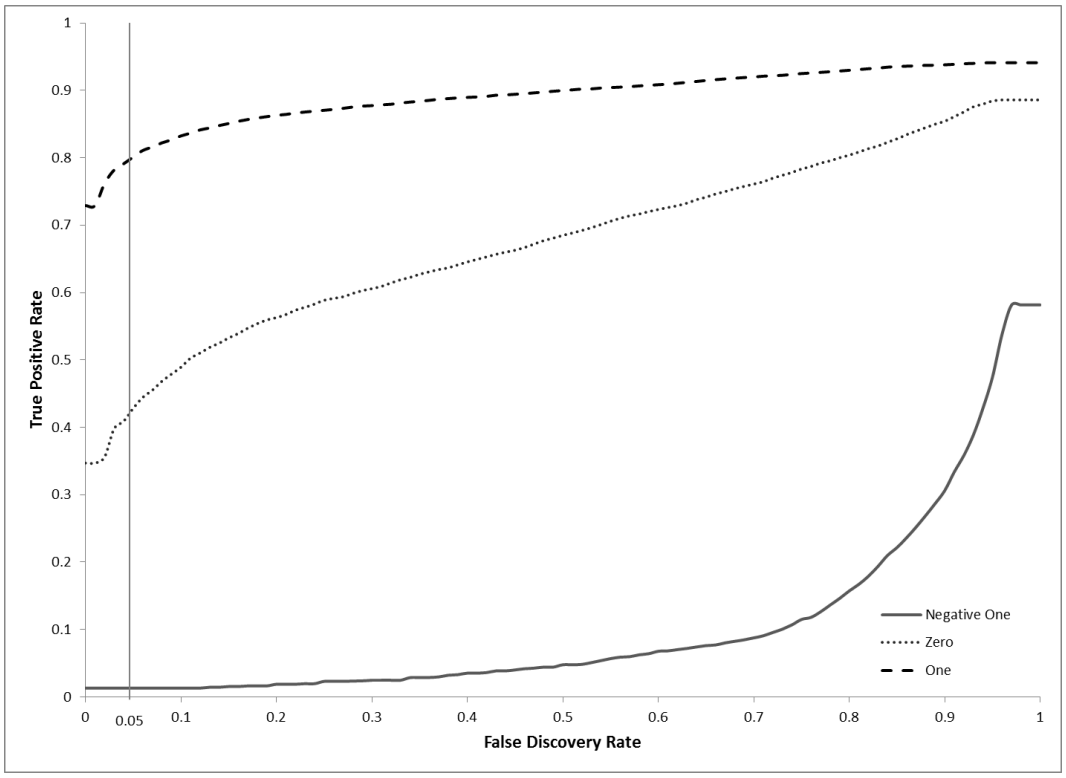

Figure 82: ROC Curves, CMP/NRM, estimated person parameters, mixed shifts

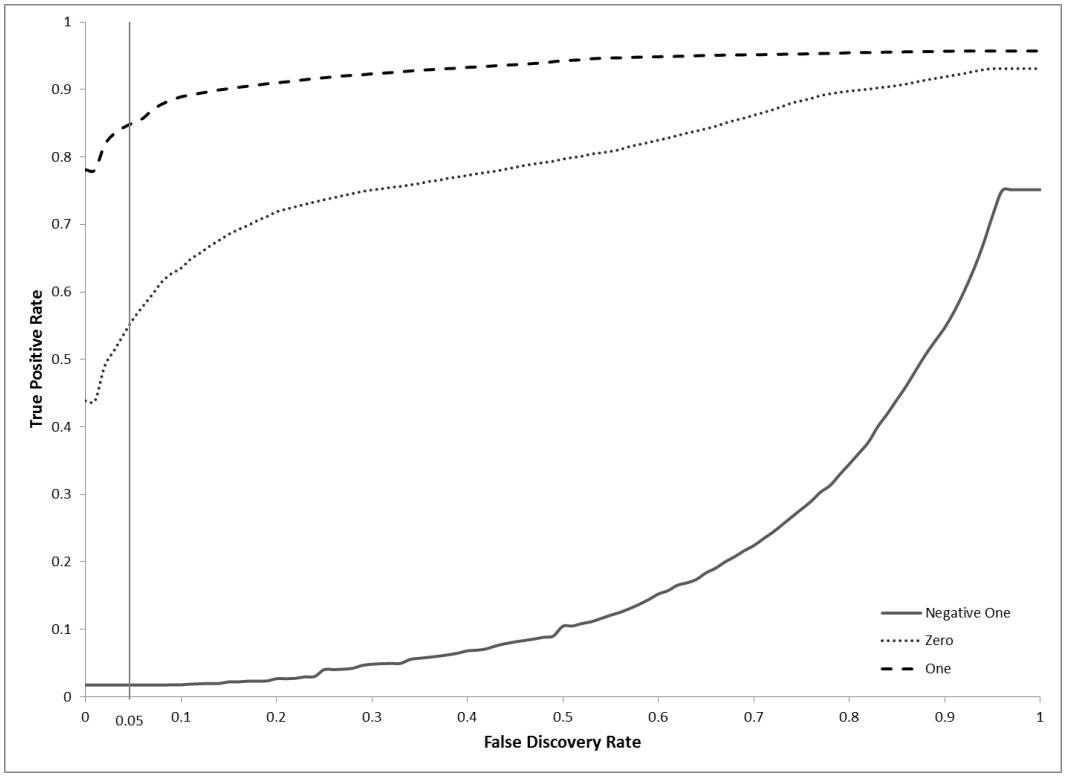

Figure 83: ROC Curves, SCIP/3PL, estimated person parameter levels, mixed shifts

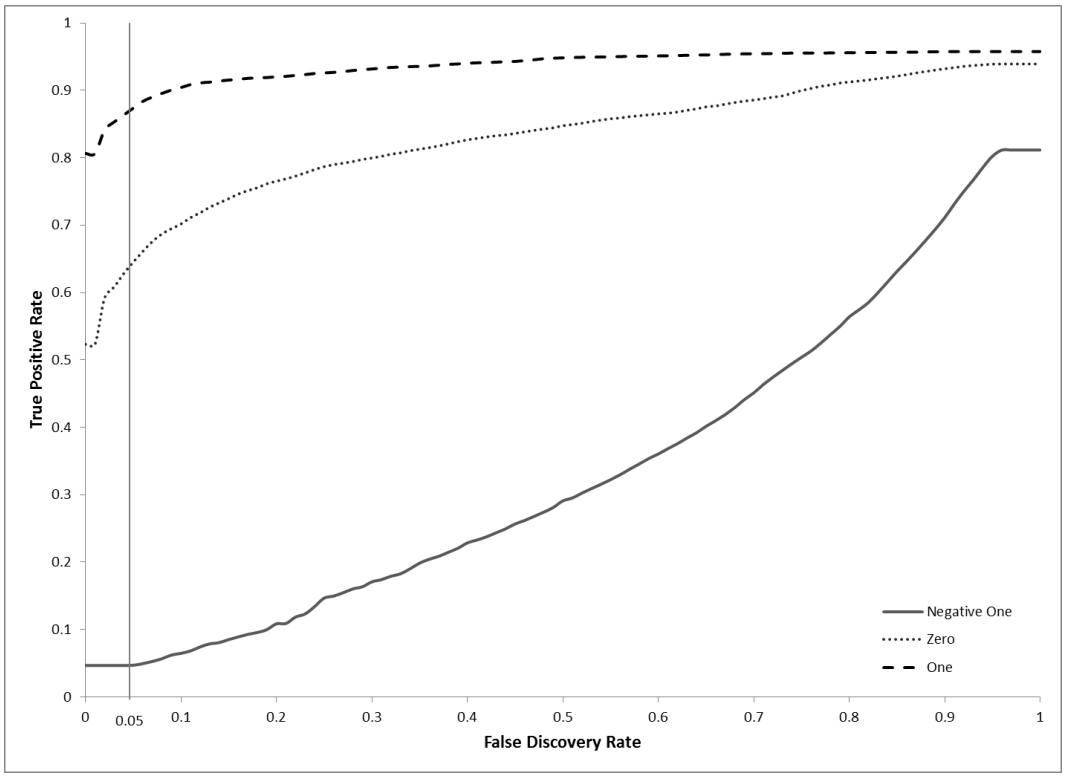

Figure 84: ROC Curves, SCIP/NRM, estimated person parameters, mixed shifts

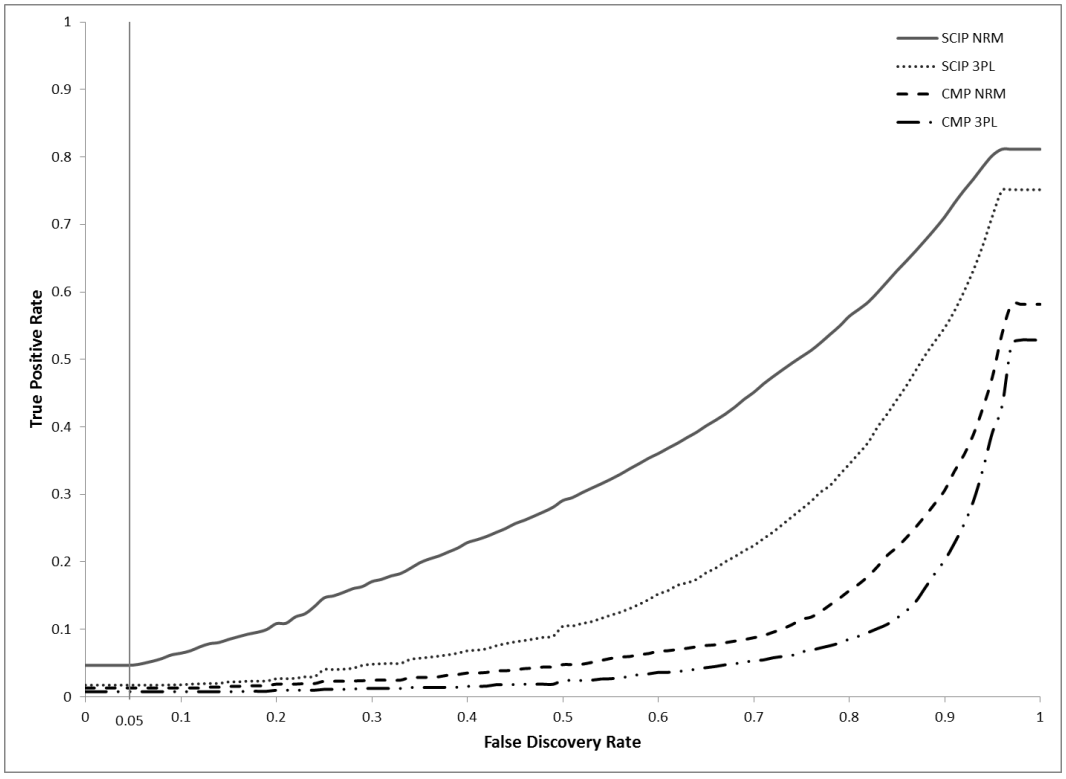

Figure 85: ROC Curves, all methods, estimated person parameters = -1, mixed shifts

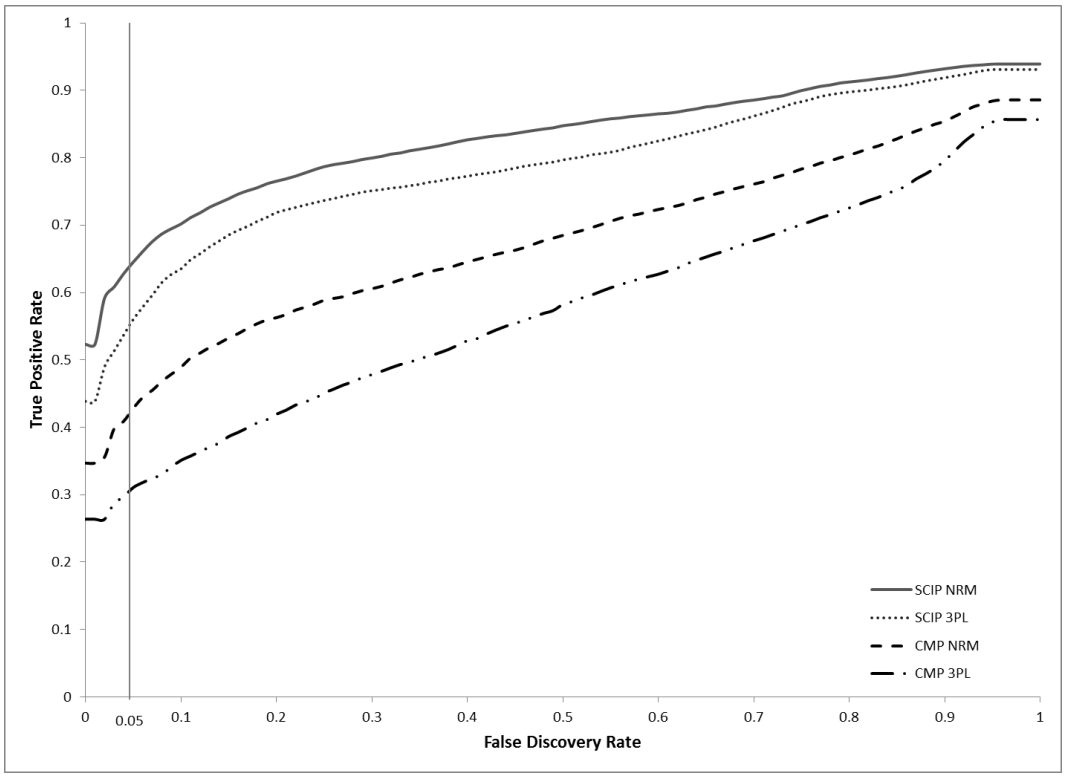

Figure 86: ROC Curves for all methods for estimated person parameters = 0, mixed shifts

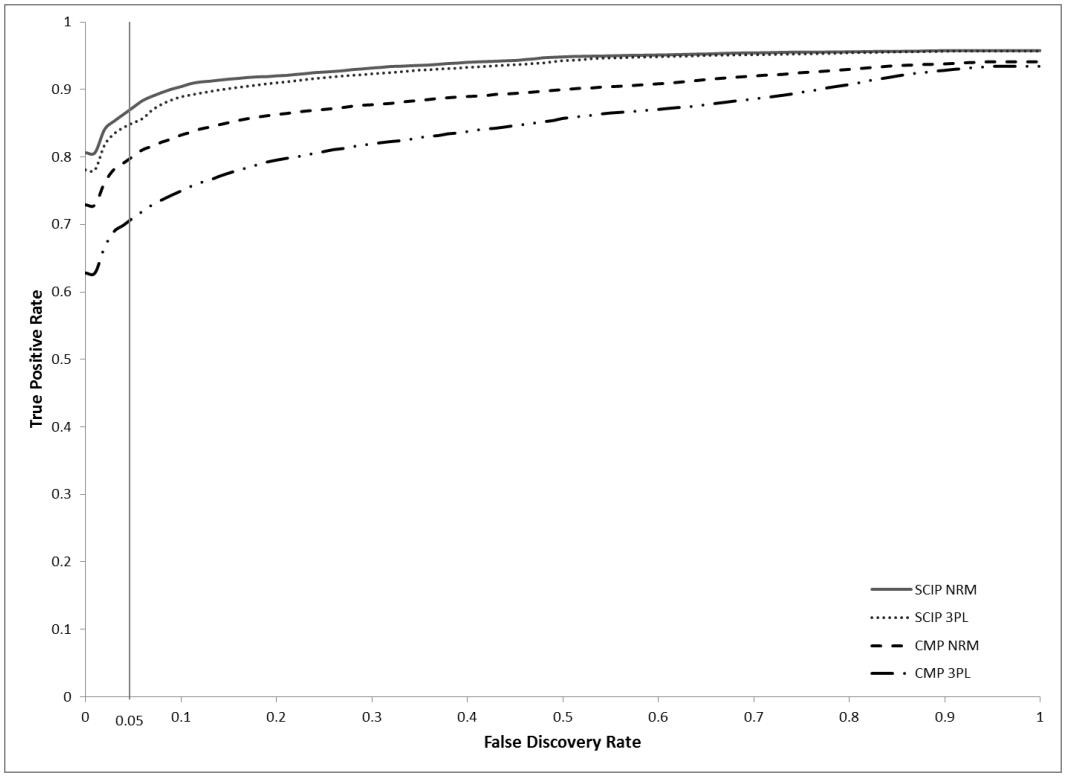

Figure 87: ROC Curves for all methods for estimated person parameters = 1, mixed shifts

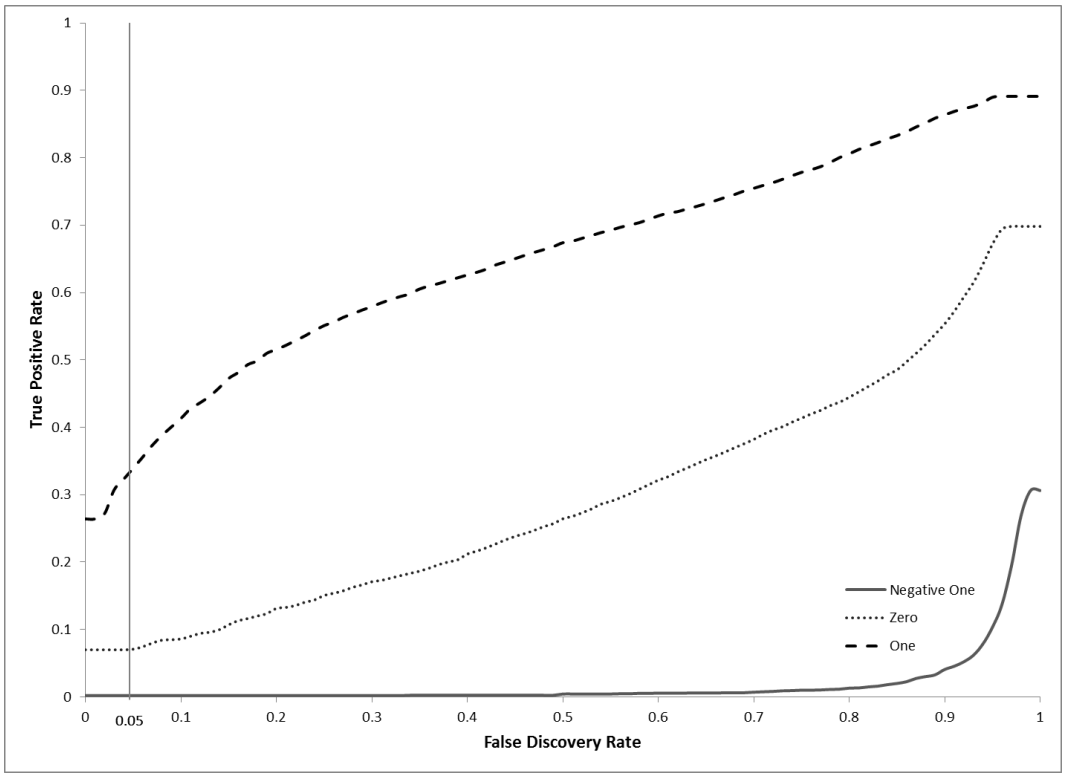

Figure 88: ROC Curves, CMP/3PL, bias-controlled person parameters, shift length 3

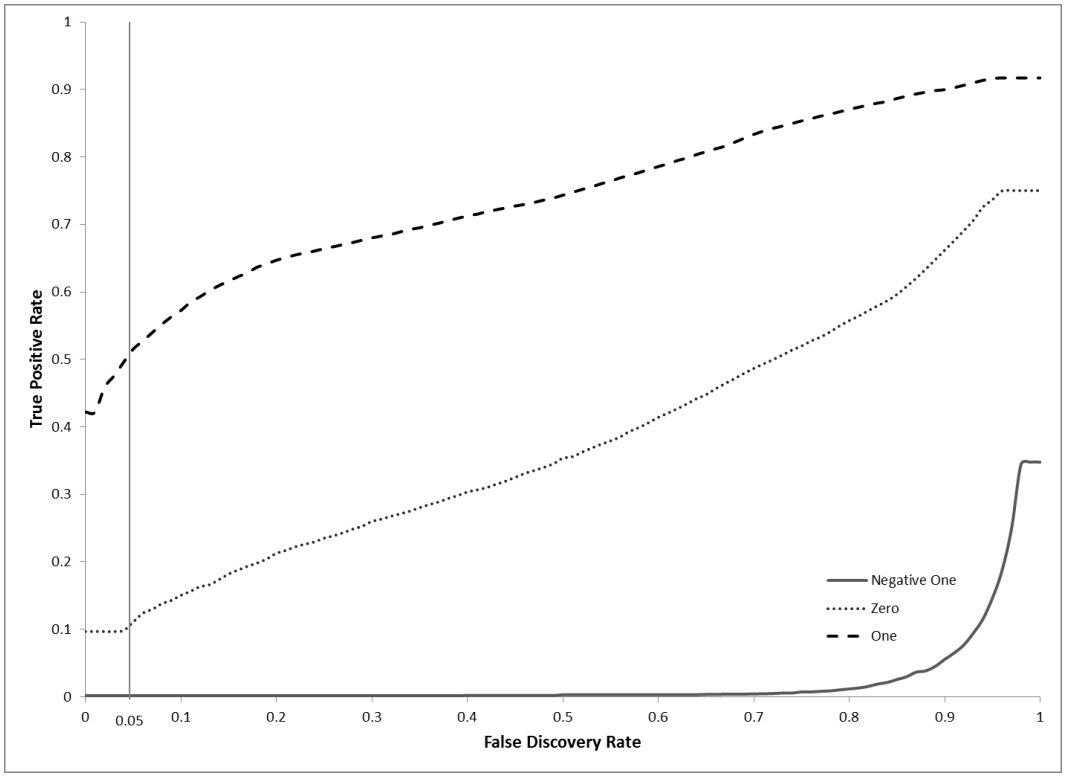

Figure 89: ROC Curves, CMP/NRM, bias-controlled person parameters, shift length 3

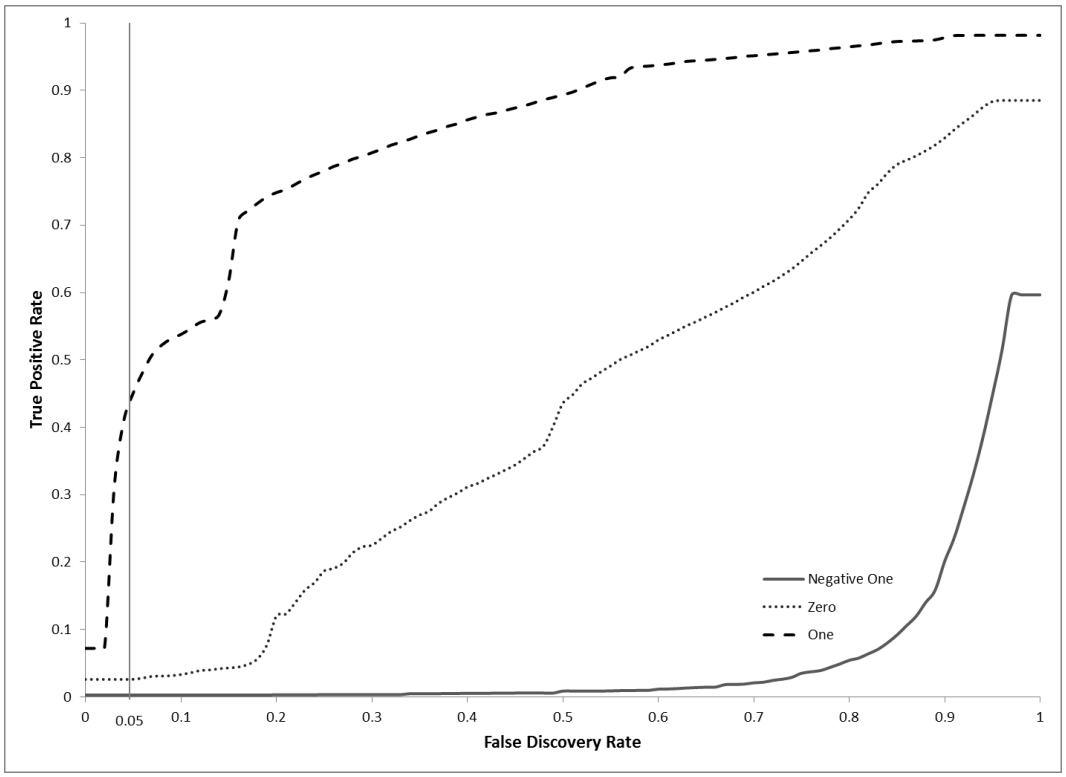

Figure 90: ROC Curves, SCIP/3PL, bias-controlled person parameters, shift length 3

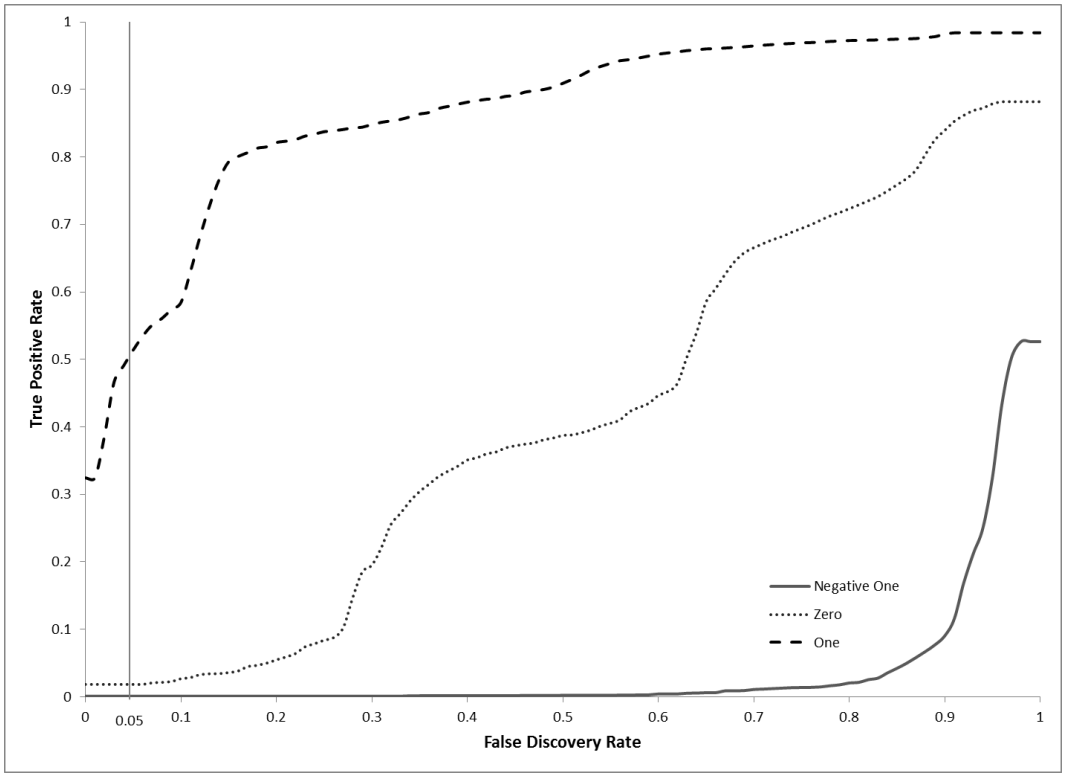

Figure 91: ROC Curves, SCIP/NRM, bias-controlled person parameters, shift length 3

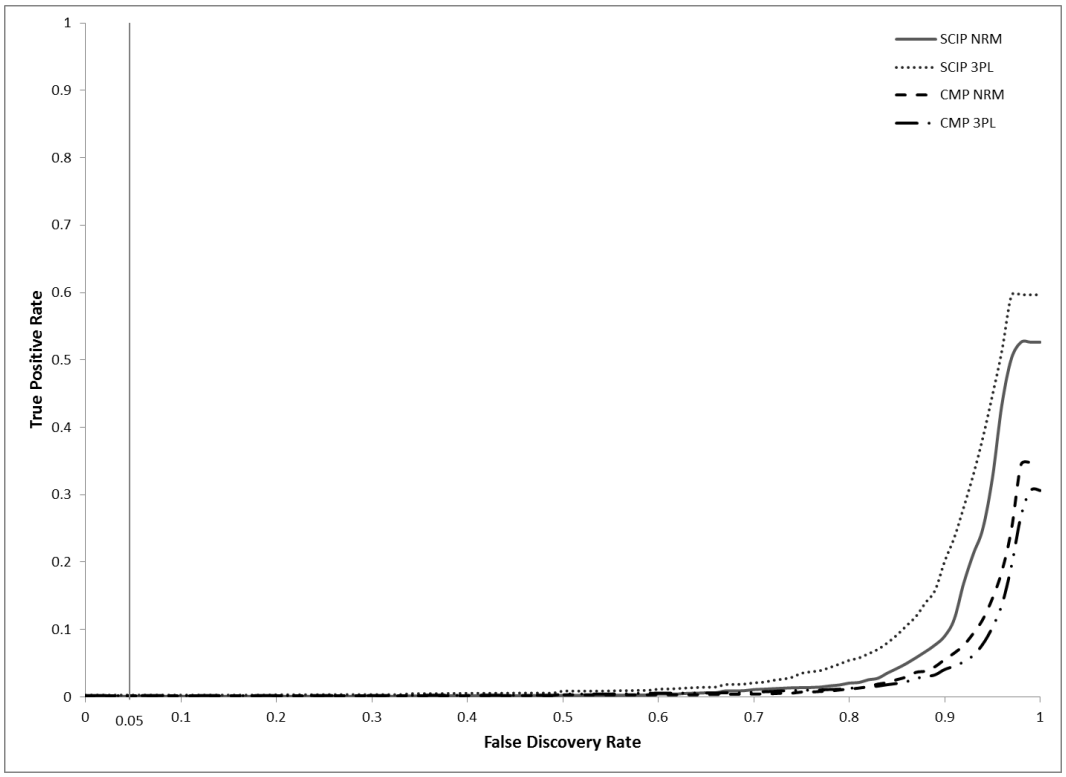

Figure 92: ROC Curves, all methods, bias-controlled parameters = -1, shift length 3

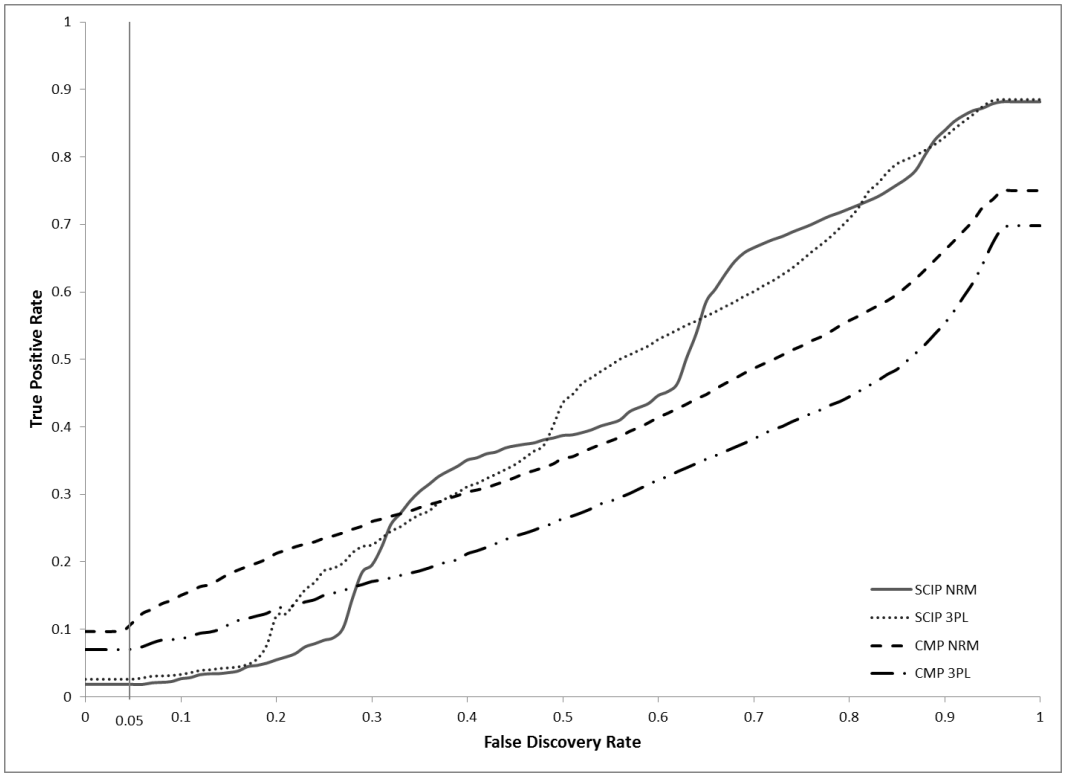

Figure 93: ROC Curves, all methods, bias-controlled person parameters = 0, shift length 3

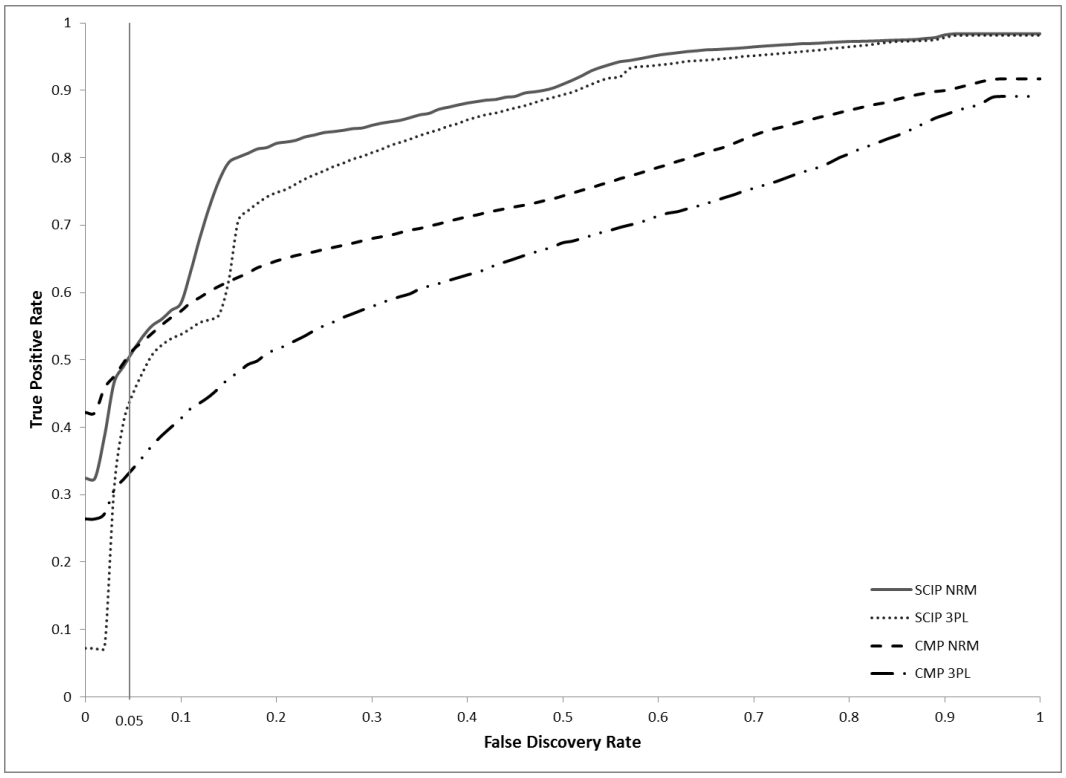

Figure 94: ROC Curves, all methods , bias-controlled person parameters = 1, shift length 3

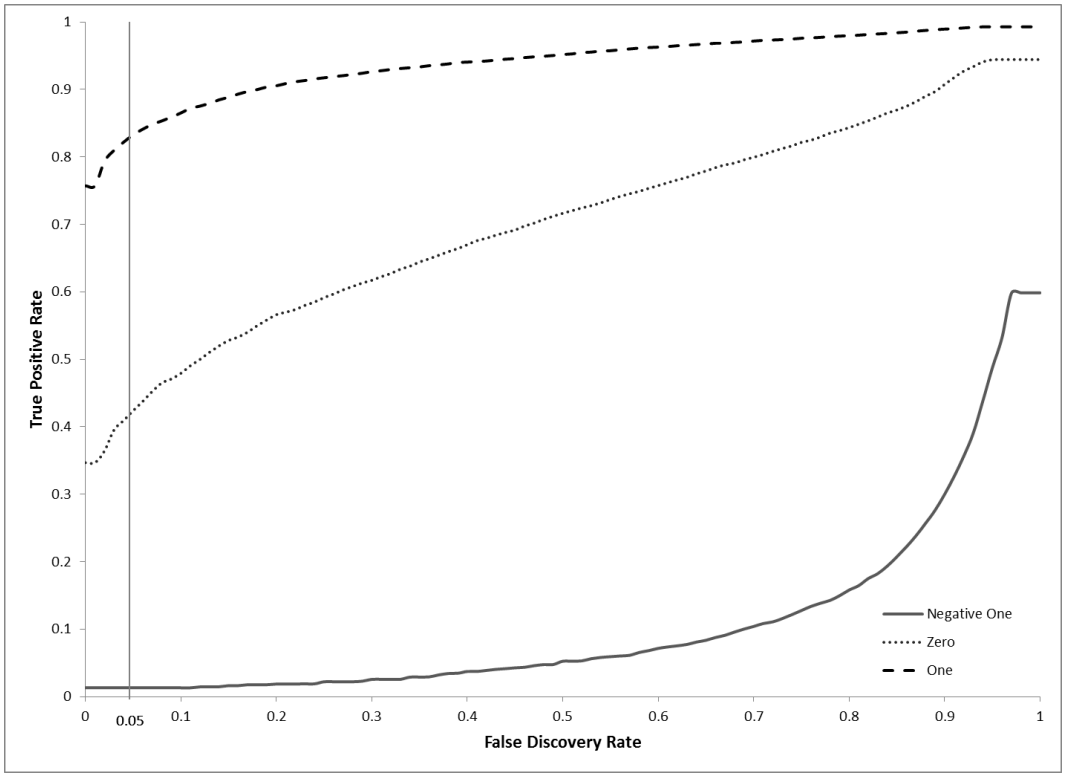

Figure 95: ROC Curves, CMP/3PL, bias-controlled person parameters, shift length 7

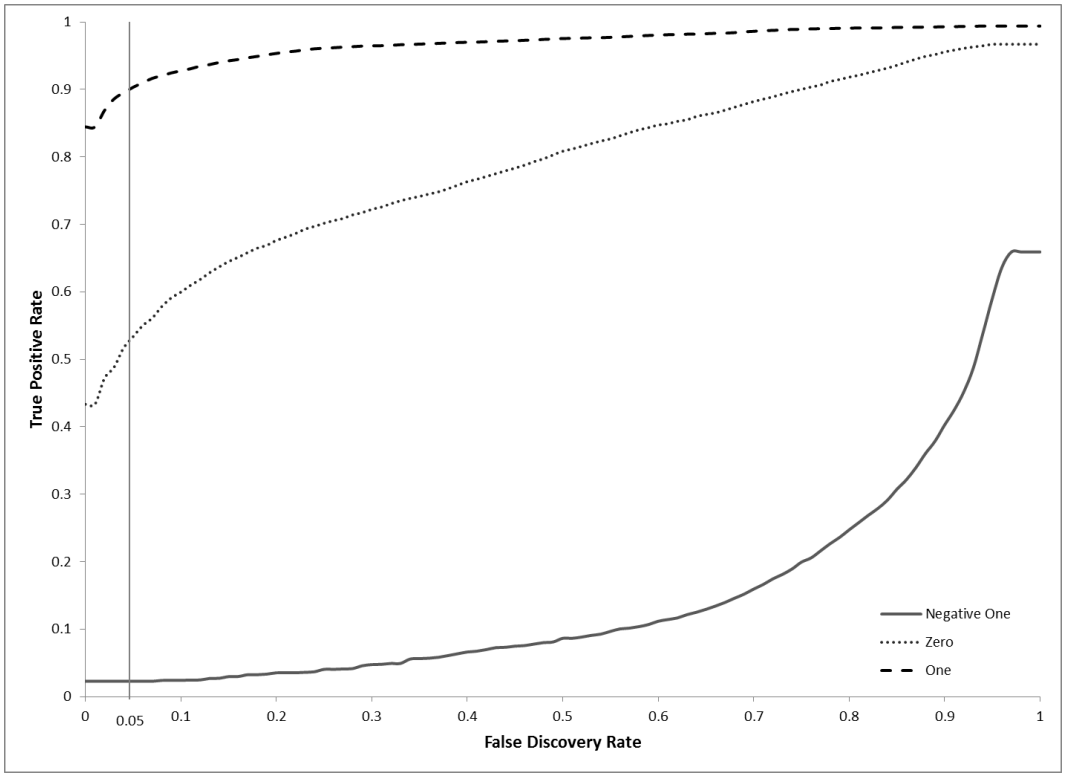

Figure 96: ROC Curves, CMP/NRM, bias-controlled person parameters, shift length 7

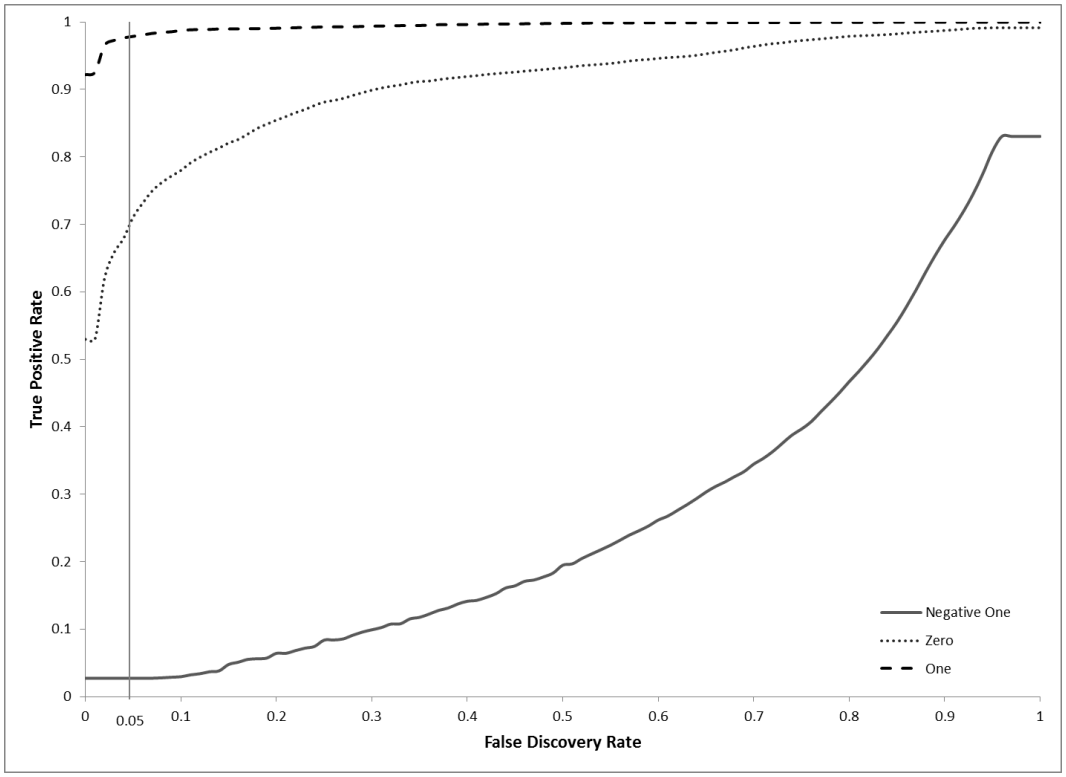

Figure 97: ROC Curves, SCIP/3PL, bias-controlled person parameters, shift length 7

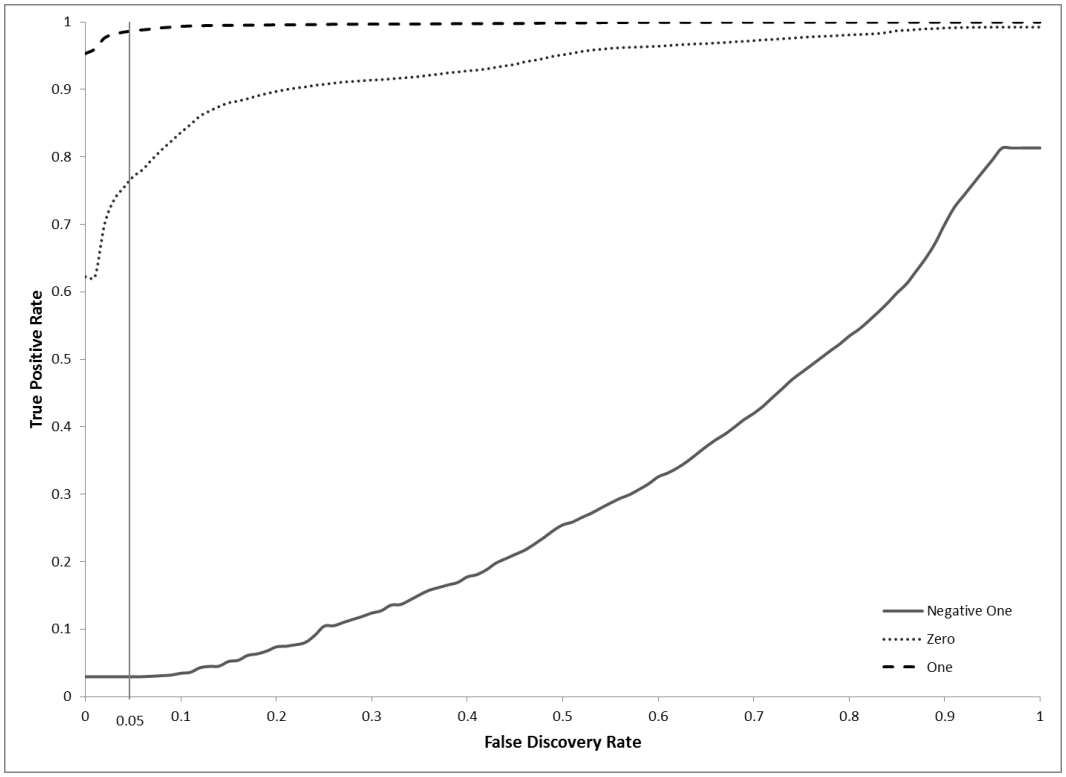

Figure 98: ROC Curves, SCIP/NRM, bias-controlled person parameters, shift length 7

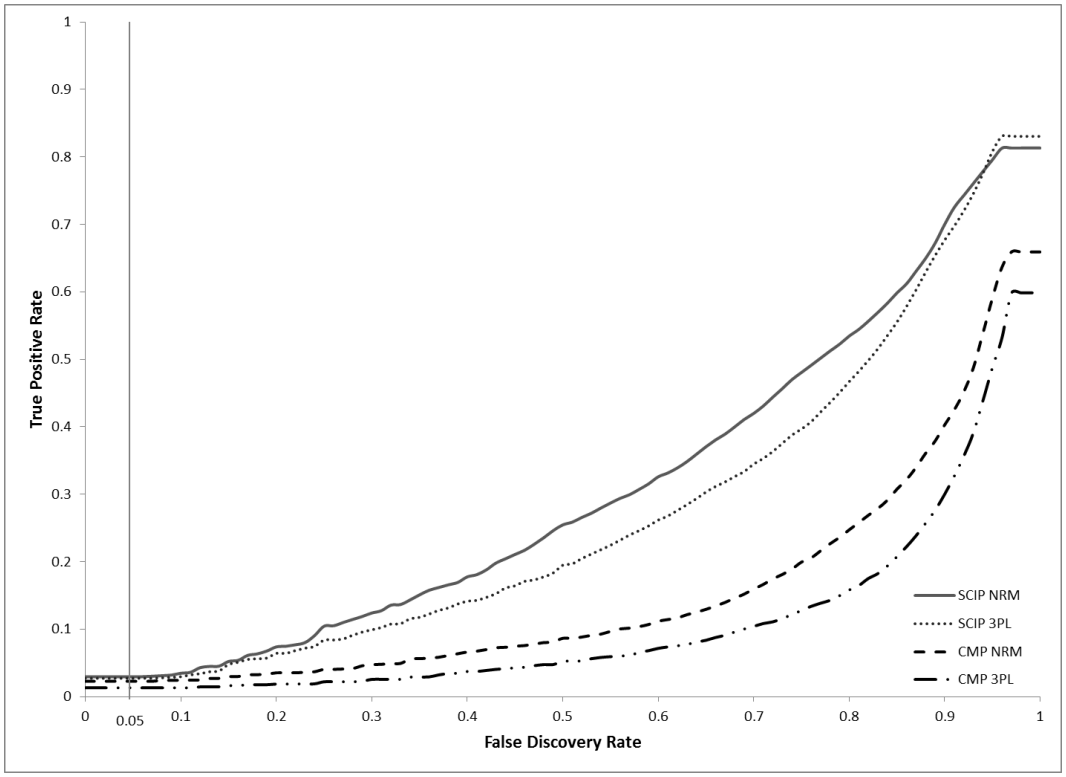

Figure 99: ROC Curves, all methods, bias-controlled parameters = -1, shift length 7

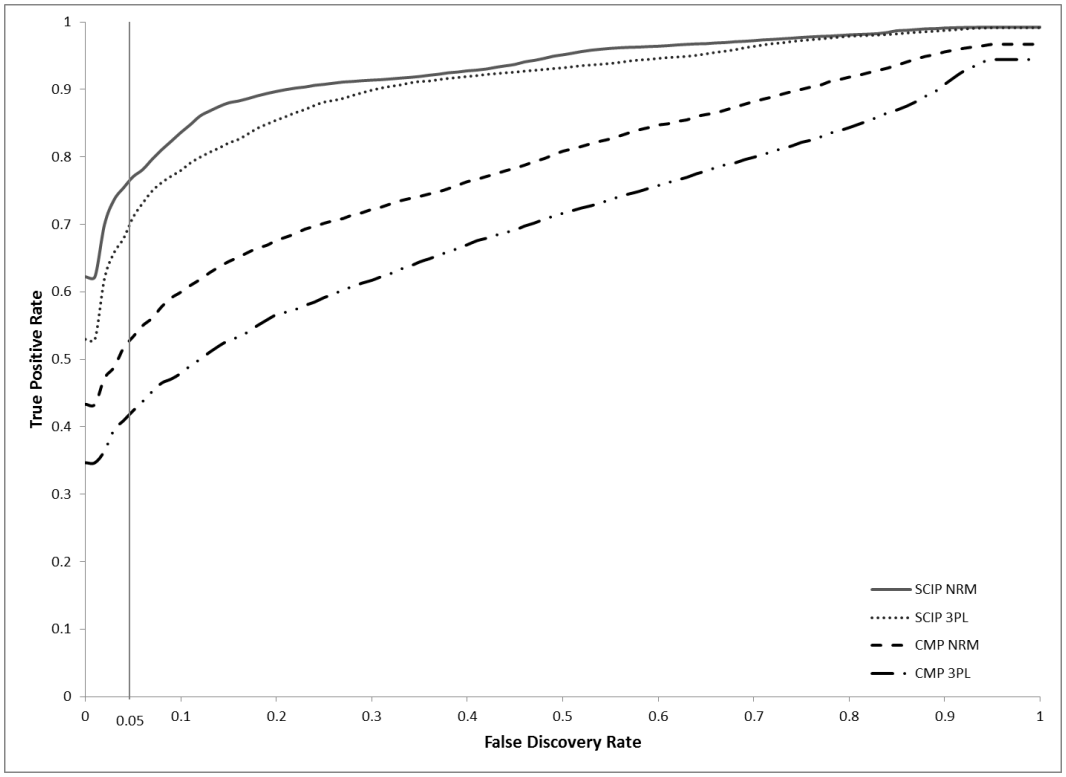

Figure 100: ROC Curves, all methods, bias-controlled parameters = 0, shift length 7

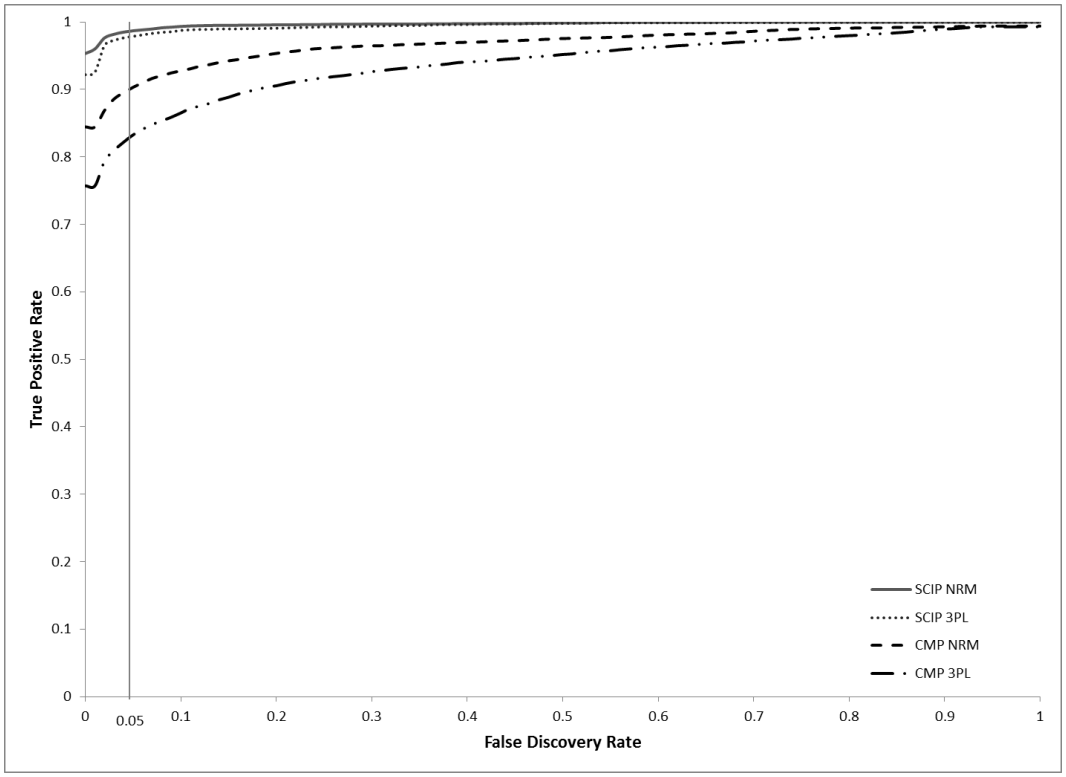

Figure 101: ROC Curves, all methods, bias-controlled parameters = 1, shift length 7

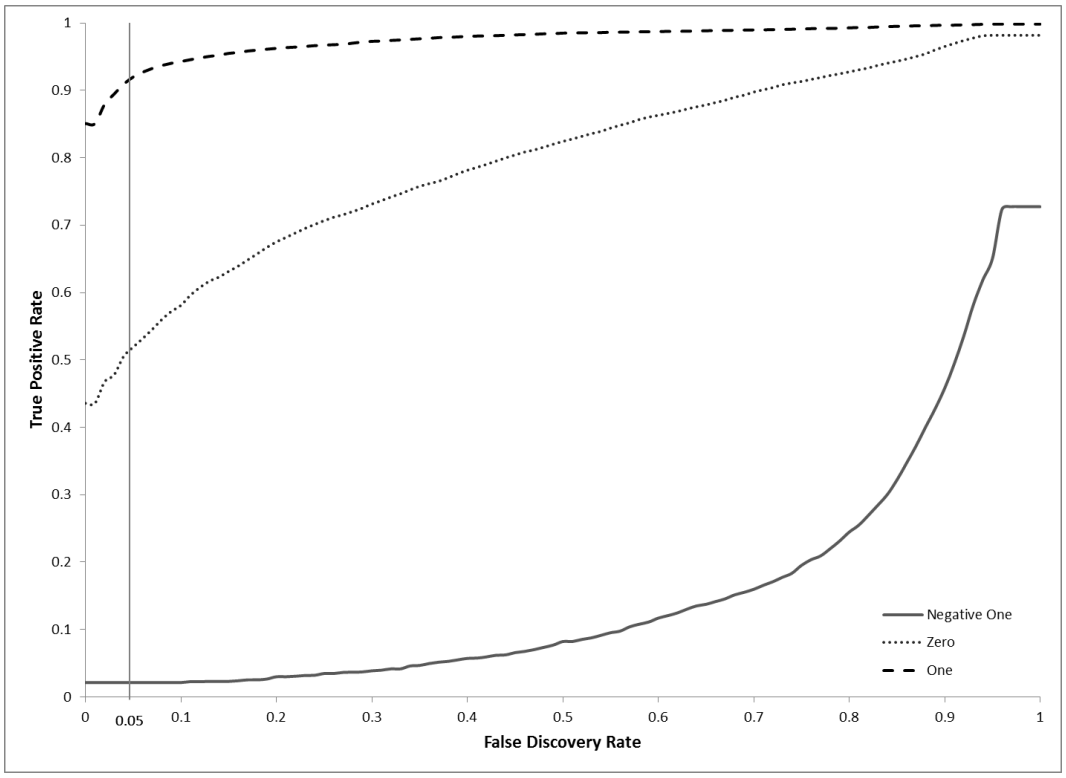

Figure 102: ROC Curves, CMP/3PL, bias-controlled parameters, shift length 10

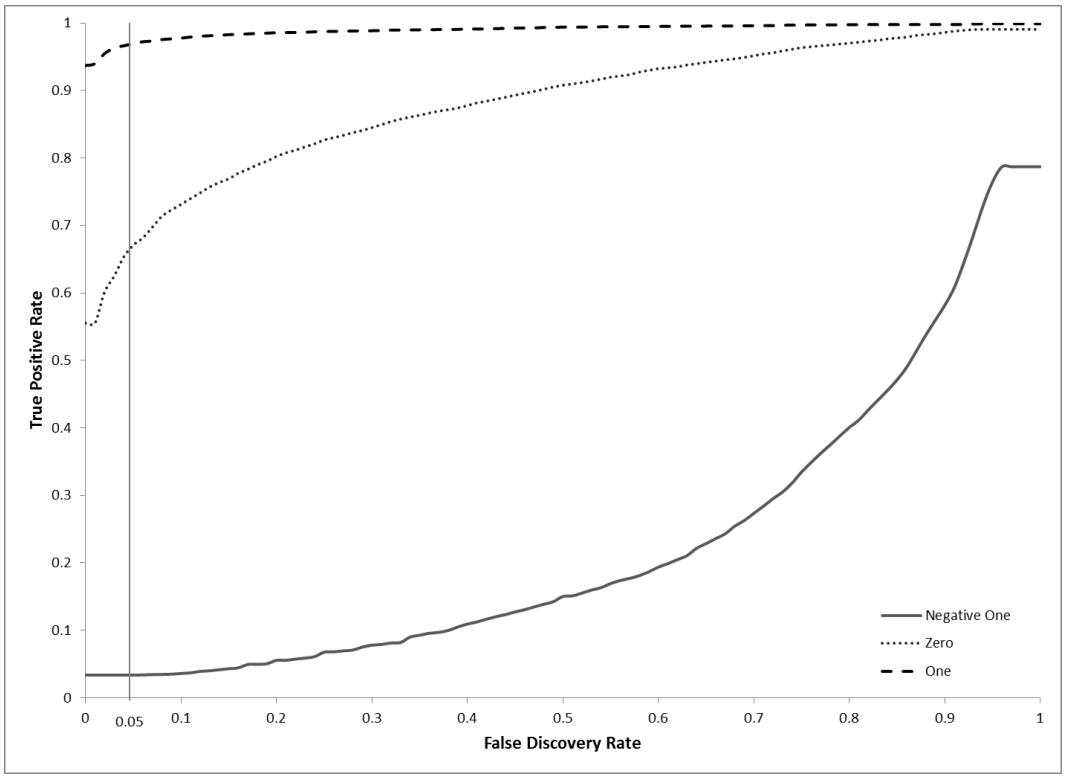

Figure 103: ROC Curves, CMP/NRM ,bias-controlled parameters, shift length 10

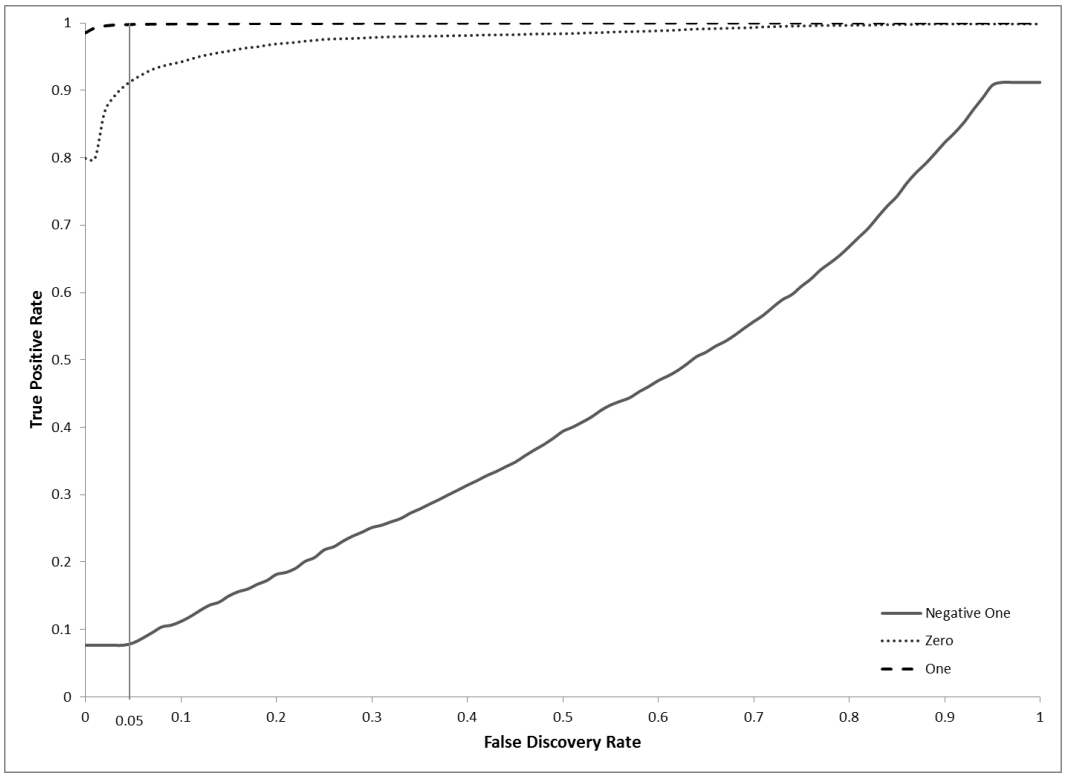

Figure 104: ROC Curves, SCIP/3PL, bias-controlled parameters, shift length 10

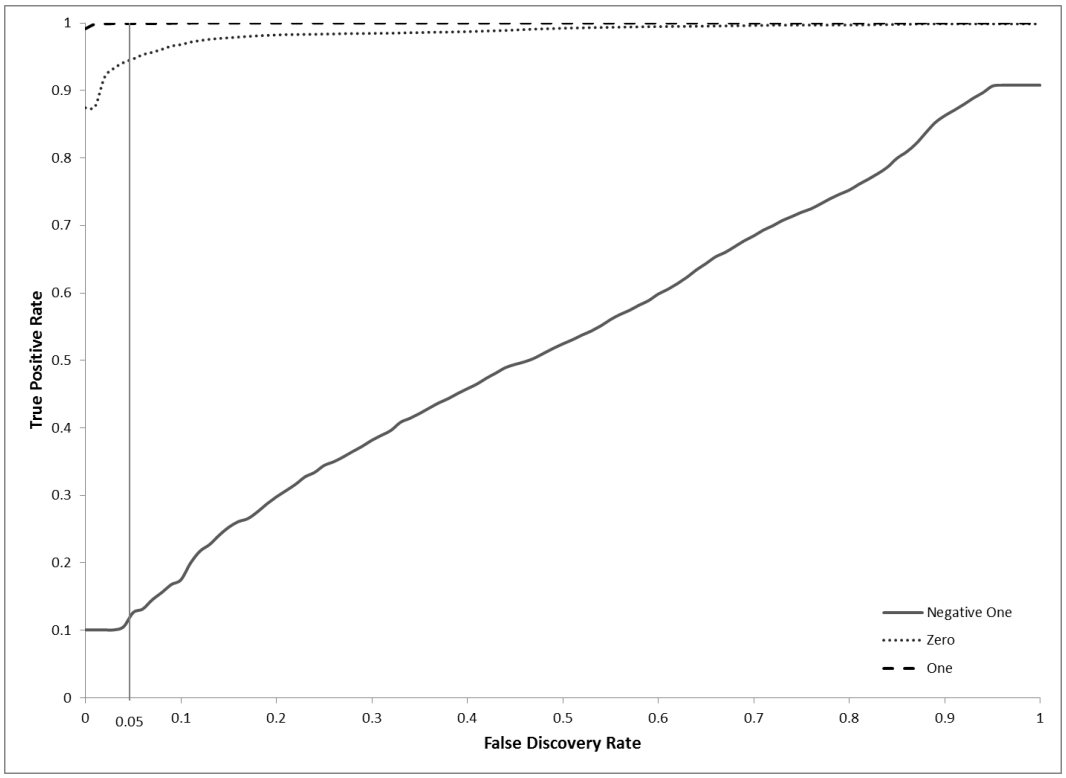

Figure 105: ROC Curves, SCIP/NRM, bias-controlled parameters, shift length 10

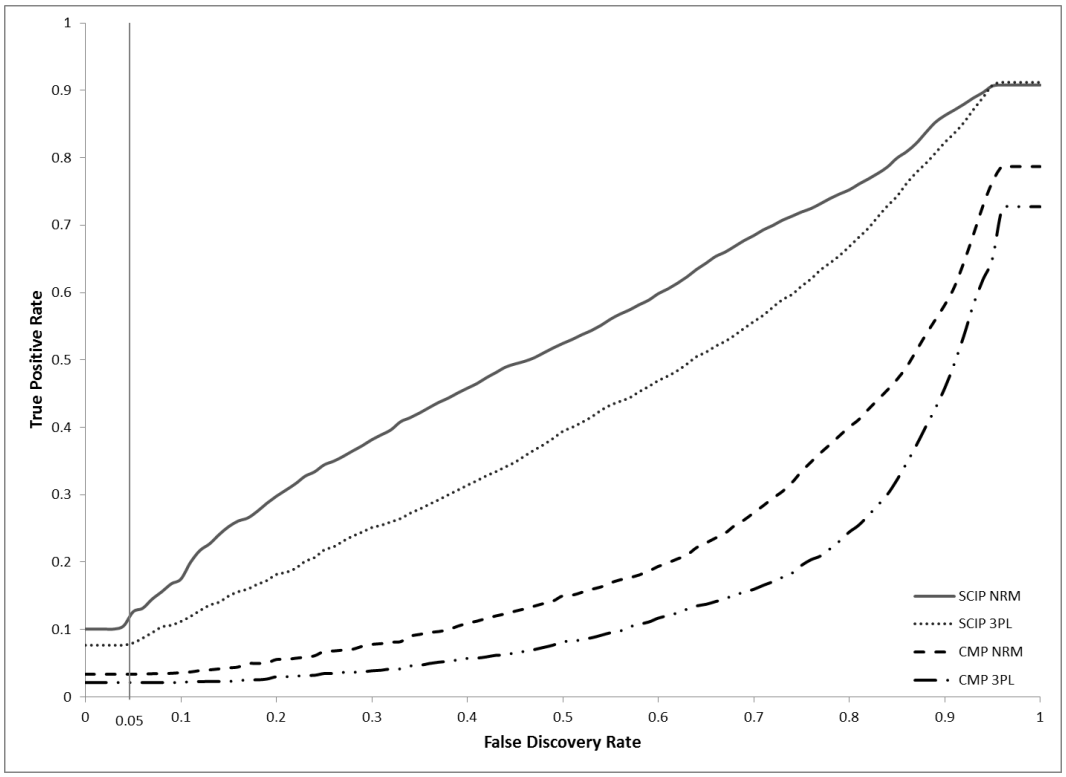

Figure 106: ROC Curves, all methods, bias-controlled parameters = -1, shift length 10

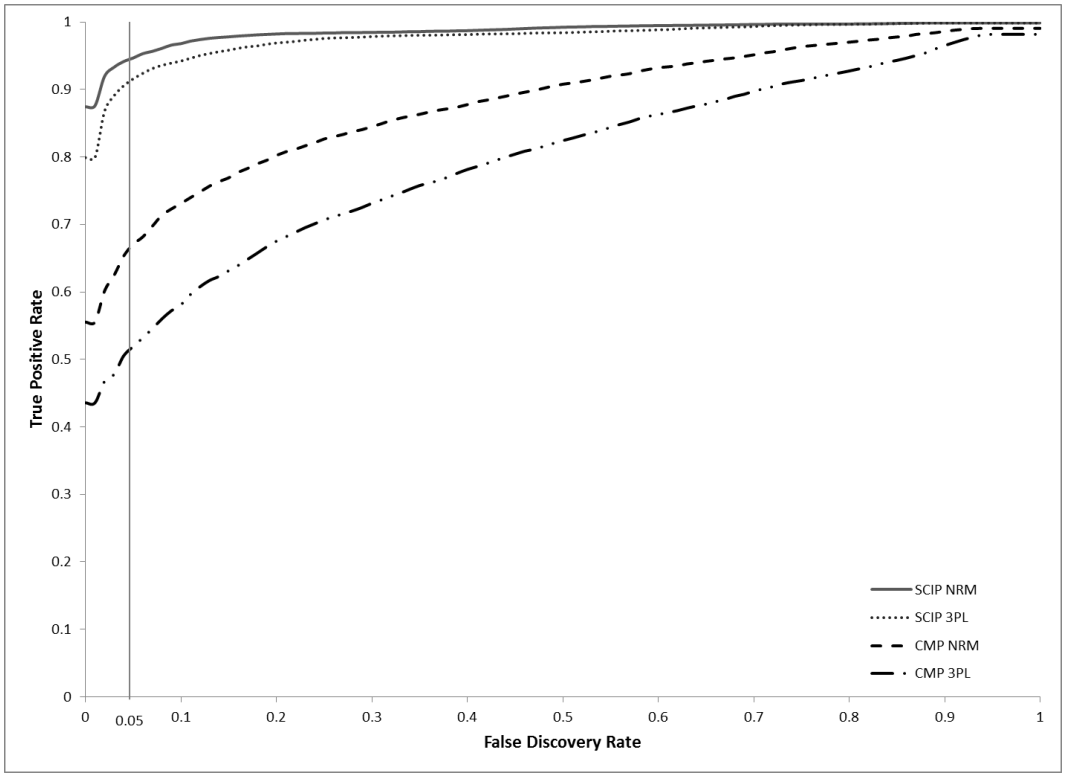

Figure 107: ROC Curves, all methods, bias-controlled parameters = 0, shift length 10

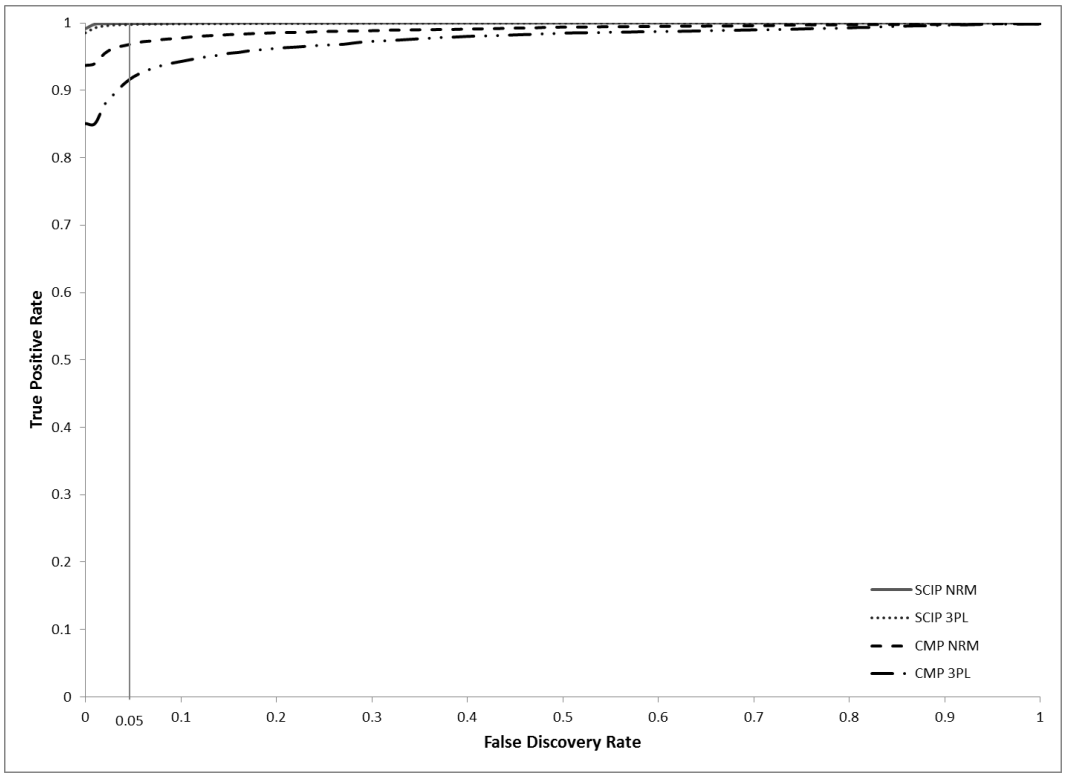

Figure 108: ROC Curves, all methods, bias-controlled parameters = 1, shift length 10

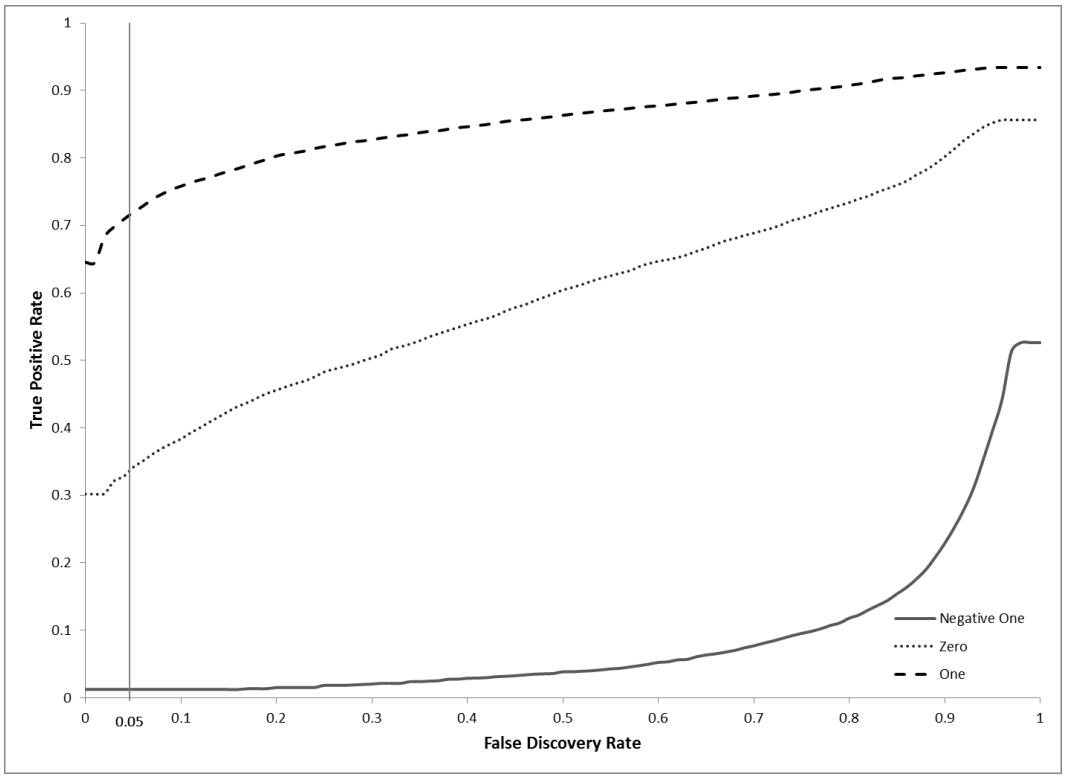

Figure 109: ROC Curves, CMP/3PL, bias-controlled person parameters, mixed shifts

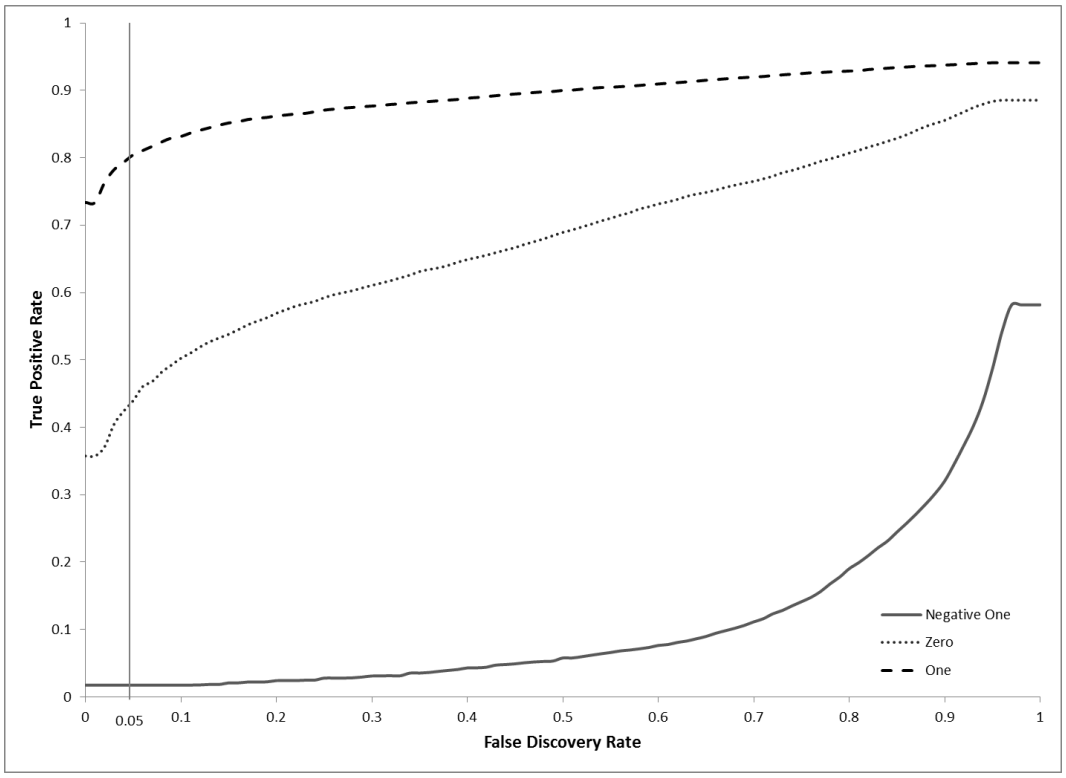

Figure 110: ROC Curves, CMP/NRM, bias-controlled person parameters, mixed shifts

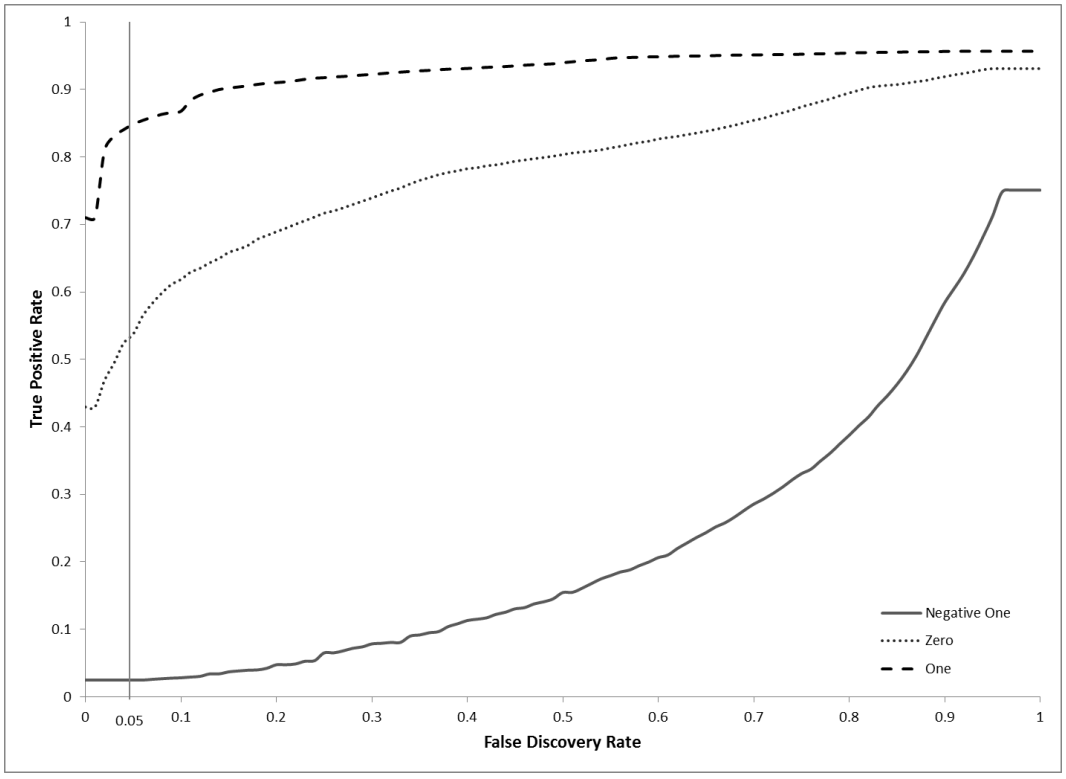

Figure 111: ROC Curves, SCIP/3PL, bias-controlled person parameters, mixed shifts

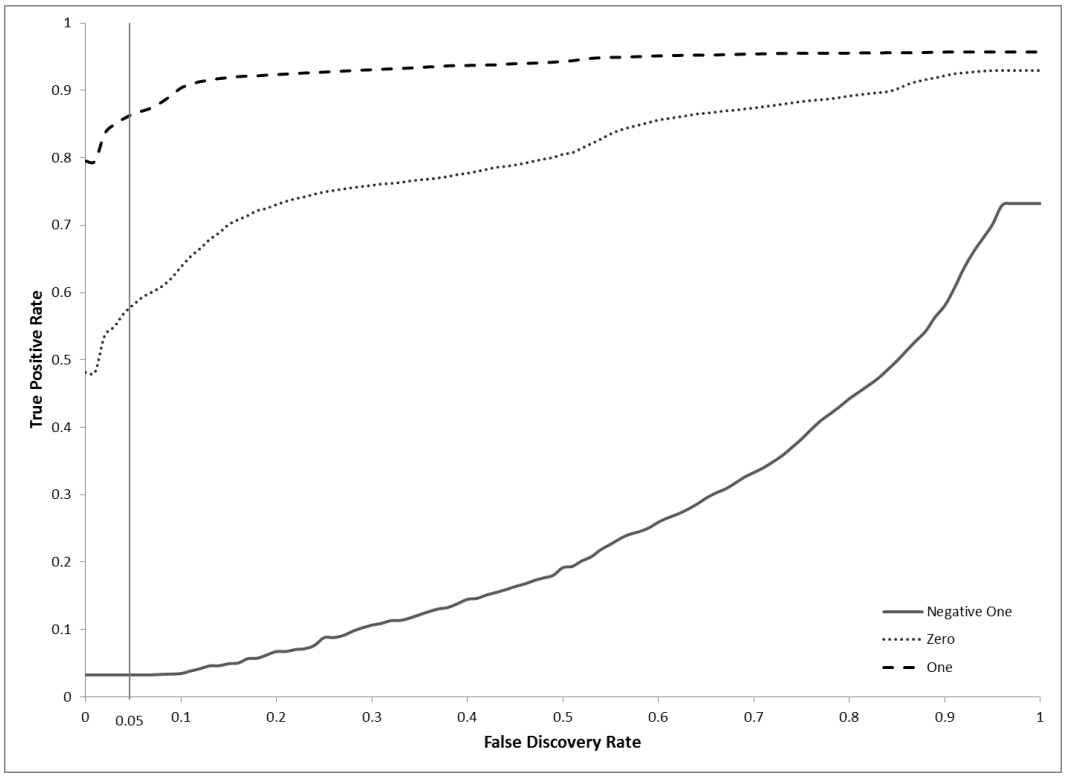

Figure 112: ROC Curves, SCIP/NRM, bias-controlled person parameters, mixed shifts

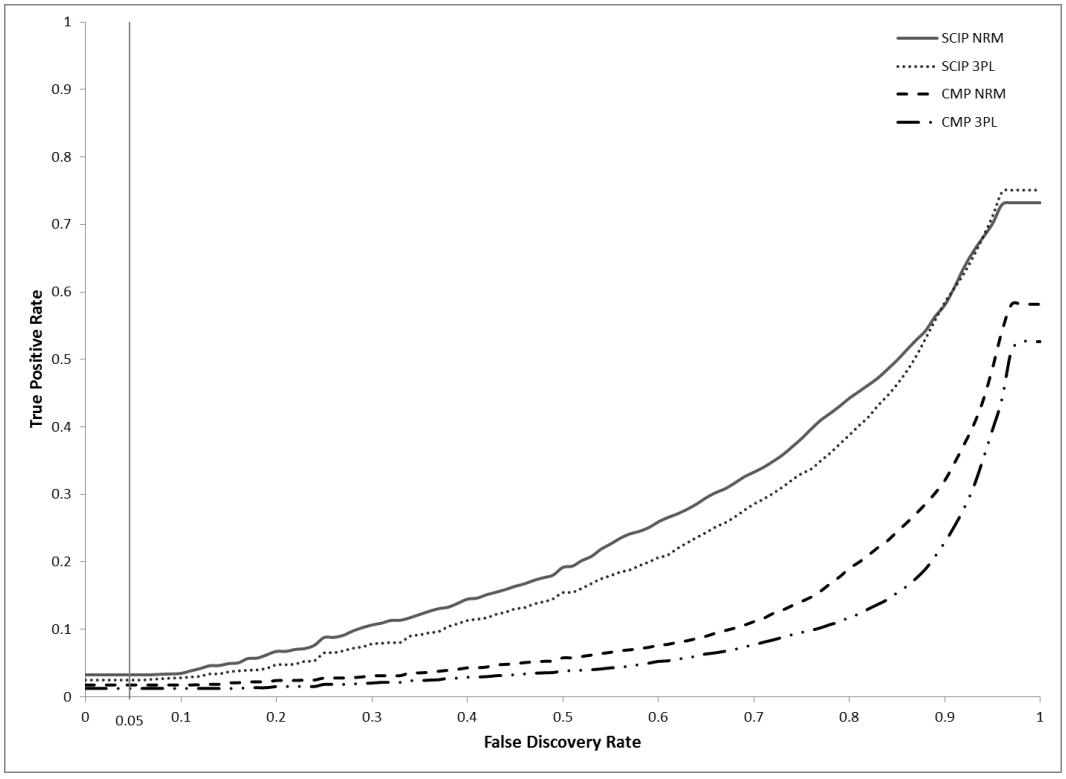

Figure 113: ROC Curves, all methods, bias-controlled parameters = -1, mixed shifts

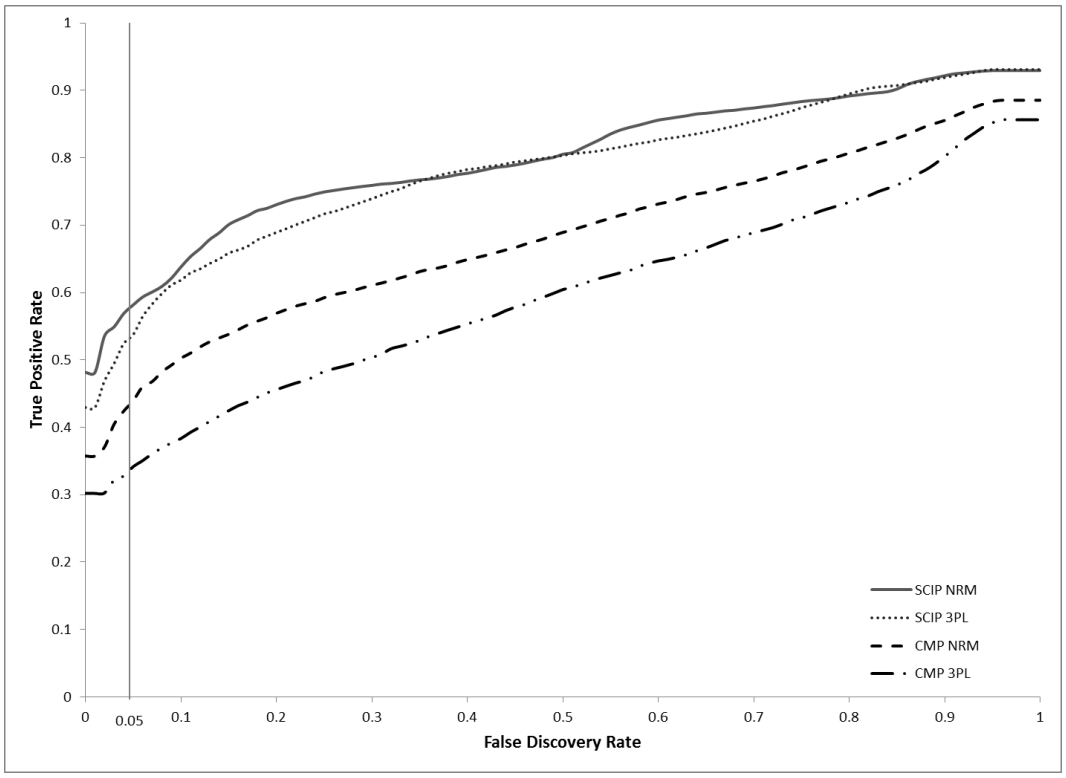

Figure 114: ROC Curves, all methods, bias-controlled parameters = 0, mixed shifts

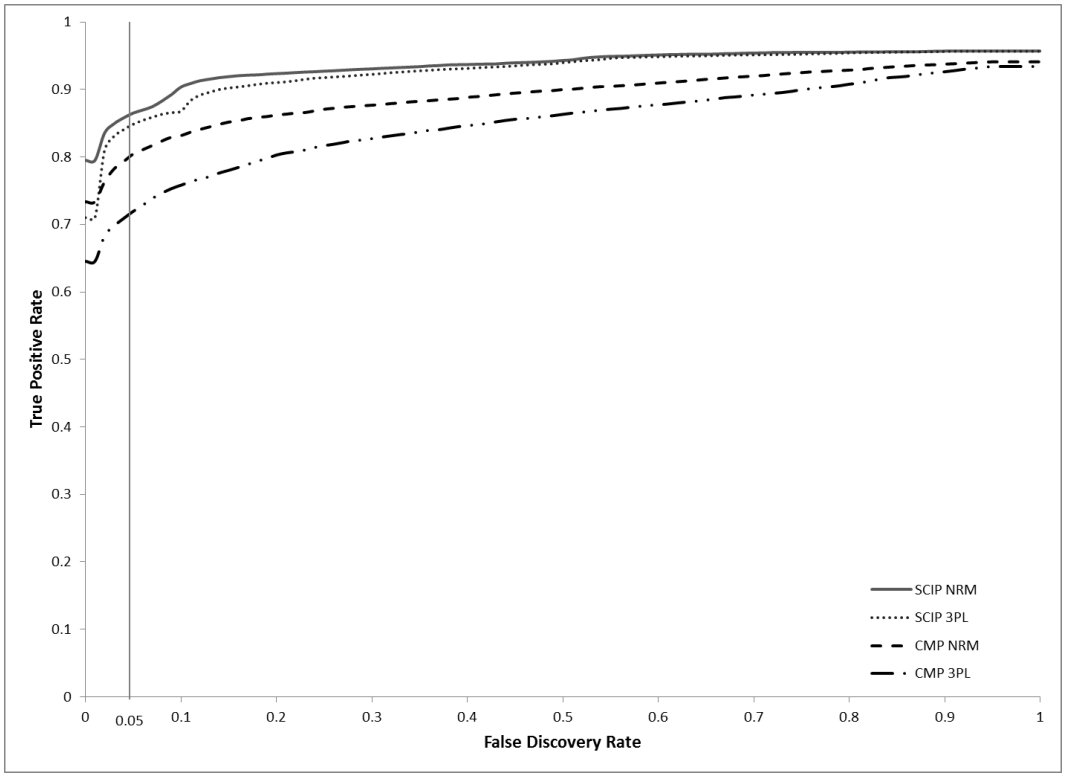

Figure 115: ROC Curves, all methods, bias-controlled parameters = 1, mixed shifts

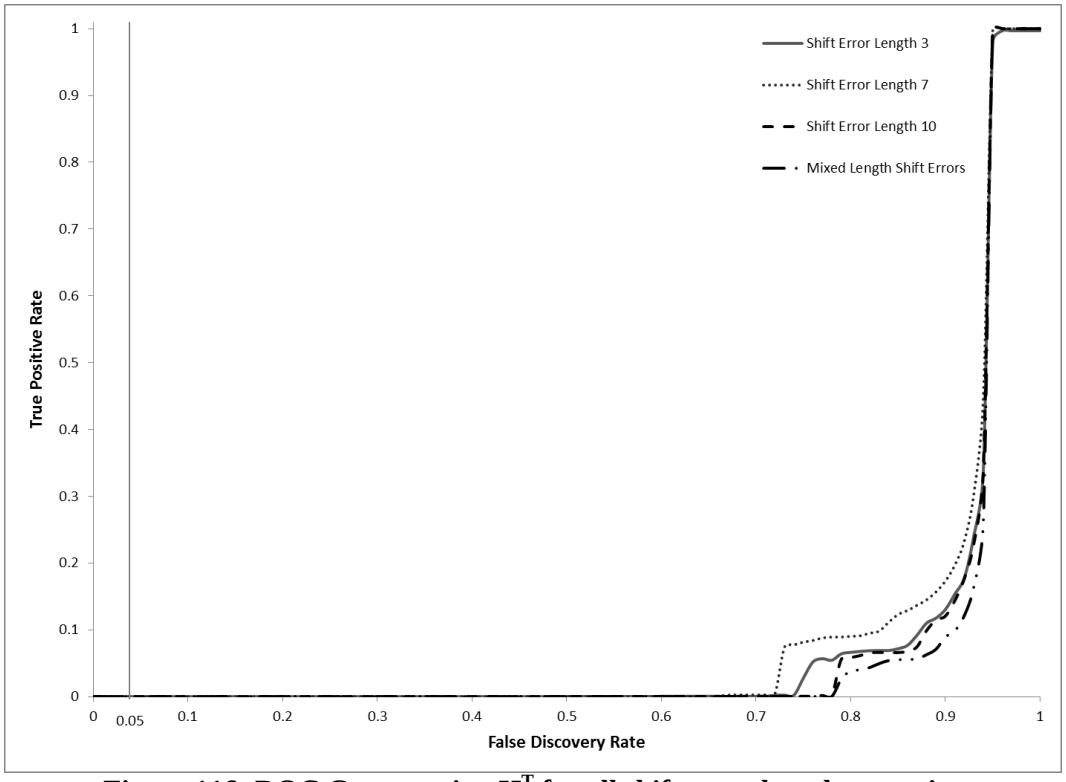

Figure 116: ROC Curves using H^T for all shift error length scenarios

REFERENCES

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC.

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69(345),* 44-49.

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement, 33(5),* 391-410.

Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement, 35*, 495-517.

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

Birnbaum, A. (1968). Some latent trait models. In Lord, F. M., & Novick, M. R. (Eds.) *Statistical theories of mental test scores.* Boston, MA: Addison-Wesley.

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

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive testing. *Journal of the American Statistical Association, 93*, 910–919.

Cai, L. (2012). flexMIRT™ version 1.88: A numerical engine for multilevel item factor analysis and test scoring. [Computer software]. Seattle, WA: Vector Psychometric Group.

Cashen, V. M., & Ramseyer, G. C. (1969). The use of separate answer sheets by primary age children. *Journal of Educational Measurement, 6(3),* 155-158

Cook, R. J., & Foster, C. C. (2012). *Application of the 3PL IRT model to the detection of shift errors.* Paper presented at the annual meeting of the Northeastern Educational Research Association. Rocky Hill, CT.

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

DeMars, C. E. (2008). *Scoring multiple choice items: A comparison of IRT and classical polytomous and dichotomous methods.* Paper presented at the National Council on Measurement in Education. New York, NY.

Dodeen, H., & Darabi, M. (2009). Person-fit: Relationship with four personality tests in mathematics. *Research Papers In Education, 24(1),* 115-126.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement, 28*, 105–113.

Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology, 72*, 10-18.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59–67.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59–79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171–191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*, 47–64.

Dunlap, J. W. (1940). Problems arising from the use of a separate answer sheet. *The Journal of Psychology, 10*, 3-48.

Frary, R. B. , Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics , 2*, 235-256.

Gaffney, R. F., & Maguire, T. O. (1971). Use of optically scored test answer sheets with young children. *Journal of Educational Measurement, 8(2)*, 103-106.

Green, M. G., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Gusfield, D. (1997) *Algorithms on strings, trees, and sequences.* Cambridge, UK: Cambridge University Press.

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review, 91*, 139-150.

Guttman, L. A. (1950). The basis for scalogram analysis. In Stouffer, S. A., Guttman, L.A., & Schuman, E. A. (Eds.) *Measurement and prediction. Volume 4 of Studies in social psychology in World War II.* Princeton: Princeton University Press.

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

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*, 133–146.

Hendrawan, I., Glas, C. W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement, 29(1),* 26-44.

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-Index: Statistical theory and empirical support.* Princeton, NJ: Educational Testing Services.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*, 16-25.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood IL: Dow Jones- Irwin.

International Test Commission (2005). *International guidelines on computer-based and Internet delivered testing, version 2005*. International Test Commission.

Ito, K., & Sykes, R. C. (2004). *Comparability of scores from norm-referenced paper-and-pencil and web-based linear tests for grades 4–12.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kane, M. (2006). Validity. In R. L. Linn (Ed.), *Educational Measurement (4th ed)*. New York: American Council on Education, Macmillan Publishing.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4*, 105–126.

Karabatsos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement In Education, 16(4)*, 277-298.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*, 22-37.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56*, 535–547.

Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications*, (pp. 97–110). New York: Springer-Verlag.

Klauer, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology, 43*, 193–206.

Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109–131). New York: Academic Press.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161–176.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–290.

Massachusetts Department of Education (2012). *Spring 2012 MCAS tests: Summary of state results.* Malden, MA.

Matter, M. K. (1985). *The relationship between achievement test response changes, ethnicity, and family income.* PhD Thesis. Austin, TX: University of Texas

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147–160.

McMorris, R. F., & Weideman, A. H. (1986). Answer changing after instruction on answer changing. *Measurement and Evaluation in Counseling and Development, 18*, 93-101.

McMorris, R. F., DeMers, L. P., & Schwarz, S. P. (1987). Attitudes, behaviors, and reasons for changing responses following answer-changing instruction. *Journal of Educational Measurement, 24*, 131–143.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311–314.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9(1)*, 3-8.

Meijer, R. R., & de Leeuw, E. D. (1993). Person fit in survey research: The detection of respondents with unexpected response patterns. In J. H. Oud & R. A. W. van Blokland-Vogelesang (Eds.), *Advances in longitudinal and multivariate analysis in the behavioral Sciences (pp. 235–245).* Nijmegen, The Netherlands: ITS.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8*, 261–272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25(2)*, 107-135.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111–120.

Molenaar, I. W.,& Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75–106.

Muller, D., Calhoun, E., & Orling., R. (1972). Test reliability as a function of answer sheet mode. *Journal of Educational Measurement, 9(4)*, 321-324.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement, 22*, 53–69.

Petridou, A., & Williams, J. (2010). The extent of mismeasurement for aberrant examinees. *Educational Assessment, 15(1)*, 42-68.

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 Reading tests. *Educational Computing Research, 32*, 153-166.

Pomplun, M., Frey, S., & Becker, D. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*, 337-354.

Ramseyer, G. C., & Cashen, V. M. (1971). The effect of practice sessions on the use of separate answer sheets by first and second graders. *Journal of Educational Measurement, 8(3)*, 177-181

Ramseyer, G. C., & Cashen, V. M. (1985). The relationship of level of eye–hand coordination and answer marking format to the test performance of first- and second-grade pupils: Implications for test validity. *Educational And Psychological Measurement, 45(2)*, 369-375.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*, 543–570.

Reise, S. P., & Flannery, W. P. (1996). Assessing person-fit measurement of typical performance applications. *Applied Measurement in Education, 9*, 9–26.

Rowan, B. E. (2010). *Comparability of paper-and-pencil and computer-based cognitive and non-cognitive measures in a low-stakes testing environment.* PhD Thesis. Harrisonburg, VA: James Madison University.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20*, 207–219.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education, 10*, 279-293.

Sato, T. (1975). *The construction and interpretation of S-P tables.* Tokyo: Meiji Tosho.

Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement, 28*, 163–171.

Shatz, M. A., & Best, J. B. (1987). Students' reasons for changing answers on objective tests. *Teaching of Psychology, 14(4)*, 241-242.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden, 7*, 131–145.

Sijtsma, K. (1998). Methodology review: non-parametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22,* 3-31.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16,* 149–157.

Skiena, S. & Sumazin, P. (2000a). Shift error detection in standardized exams (extended abstract). In Giancarlo, R. & Sankoff, D. (Eds.), *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching,* 264-276. London, UK: Springer-Verlag.

Skiena, S. & Sumazin, P. (2000b). Detecting and correcting shift errors in standardized exams. *Educational and Psychological Measurement, submitted.*

Skiena, S. & Sumazin, P. (2004). Shift error in standardized exams. *Journal of Discrete Algorithms, 2,* 313-331

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45,* 433–444.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46,* 359–372.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying in a multiple-choice test. *Journal of Educational Measurement, 39(2),* 115-132.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221–230.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13,* 201-214.

Thissen, D., (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26(2),* 161-176.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing (pp. 83–108).* New York: Academic Press.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties*. Lisse: Swets and Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13,* 267–298.

van der Linden, W. J., & Jeon, M. (2012). Modeling Answer Changes on Test Items. Journal Of Educational And Behavioral Statistics, 37(1), 180-199.

van der Linden, W. J., & Sotaridona, L. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement, 41*, 361–377.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201–219). Boston: Kluwer.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test*. Minneapolis MN: University of Minnesota, Department of Psychology.

Wise, S. L., Duncan, A. L., & Plake, B. S. (1985). The effect of introducing third graders to the use of separate answer sheets on the ITBS. *Journal Of Educational Research, 78(5)*, 306-09.

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch Models: Parts I & II. In Smith, E. V. & Smith, R. M. (Eds.), *Rasch measurement: Advanced specialized applications.* Maple Grove, MN: Journal of Applied Measurement Press.

Wollack, J. A. (1997). A nominal response model approach to detecting answer copying. *Applied Psychological Measurement, 21(4)*, 307-320.

Wright, B. D (1995). Diagnosing person misfit. *Rasch Measurement Transactions, 9*, 430–431.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement.* Chicago: Mesa Press.