University of Alberta

Comparison of Vertical Scaling Methods in the Context of NCLB

by

Andrea Julie Gotzmann

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Measurement Evaluation and Cognition

Educational Psychology

©Andrea Julie Gotzmann
Fall 2011
Edmonton, Alberta

ABSTRACT

Vertical scaling is the process of establishing a numerical test score scale across several age or grade levels. Given that the current literature does not indicate which of the different vertical scaling procedure works “best” for all situations. This study evaluated the performance of four vertical scaling procedures (concurrent calibration, fixed common item parameters, test characteristic curve, and hybrid characteristic curve), across two content areas (Reading and Mathematics), two score distribution types (normal and negatively skewed), and two sample sizes (1,500 and 3,000). Five outcome measures were used to evaluate the results: decision accuracy, decision consistency, conditional standard errors at each of two cut-scores, root-mean-squared-differences of the scale scores between scaling procedures, and correlations between scaling procedures’ final item parameters. The data used in this study was from a U.S. large scale testing program in Reading and Mathematics for grades 3 through 8. These data were used to simulate the type of score distribution and sample sizes considered with 100 replicates for these combinations.

The largest differences among the four vertical scaling procedures for Reading were found at the lower and upper grade levels, particularly for decision accuracy. Differences were found between the normal and skewed distributions, for decision accuracy where a different pattern of results were found. The accuracy results decreased markedly as grades increased for the skewed distribution. For Mathematics the largest differences across all outcome measures occurred across grade levels rather than across vertical scaling procedures. Sample size for both Reading and Mathematics did not seem to have an effect.

Practitioners should ensure high decision accuracy and consistency values across all grade levels, and that a particular scaling procedure does not result in undesirable results. If a state program allows different procedures for different content areas, then the hybrid characteristic curve procedure would be most appropriate for Reading and the test characteristic procedure most appropriate for Mathematics. However, if the procedure must be the same, then the hybrid characteristic curve procedure could be used for both Reading and Mathematics. Measurement specialists can use these results to guide their implementation of vertical scaling for their state assessment programs.

ACKNOWLEDGEMENTS

I would like to acknowledge several people that have assisted with this journey. First, I would like to acknowledge the unending support and assistance from my supervisors Dr. Mark J. Gierl and Dr. W. Todd Rogers. Dr. Gierl encouraged me to come back to the University of Alberta to complete my doctoral studies and has supported me throughout my program, particularly through writing my dissertation and job search. If not for your constant encouragement I might not have finished and started on my new journey with the Medical Council of Canada. Dr. Todd Rogers. I would like to thank you for your constant encouragement, even when I had thought I could not finish, and support throughout my program, especially reading endless drafts of my dissertation. I appreciate your thoughtful feedback even when it drove me a bit crazy.

I would like to acknowledge my committee members on their feedback and thoughtful insight on my dissertation topic. Dr. Ying Cui, Dr. Marilyn Abbott, Dr. Les Hayduk and Dr. Ruth Childs has given me much thoughtful conversation and encouraged me to expand my thinking and approaches to research.

I would like to acknowledge Dr. Jacqueline P. Leighton and Dr. Cheryl Poth, professors which I had the pleasure of working for and with in CRAME. Both of you have been encouraging and helpful in my job search and skill development. In addition, I would like to acknowledge all of my CRAME friends and colleagues that have been helpful throughout my program. I have appreciated

all of the support and discussions, especially the long talks with Mary, Louise, and Cecilia.

Support for my program has been financially provided by the Social Sciences and Humanities Research Council with a full scholarship for three full years of my four year program. This support was especially important to assist me and my family during my studies.

This research would not have been achieved had the U.S. state Department of Education not provided the data for this study. I thank my contacts that made this possible and hope this research assists with future scaling applications.

Finally, I would like to acknowledge my family, my brothers (Wilfred, Conrad, Roland, Rudy, and Gerald) and sisters (Mariane, Rosanne, and Judy) as well as my nephews (William and Jonathan) has been encouraging and helpful throughout my program even when you didn’t understand everything I was doing. To my parents (Reinhold and Alice) I would like to thank you for your support and belief I could accomplish anything I set my mind to. Most importantly I would like to acknowledge the great sacrifice and support that my daughter Adina Gotzmann has contributed the most to my current success. I know that living with a PhD student is very tough especially when your parent is at wits end you were always encouraging, supportive and funny when I needed a laugh. Adina you were also very helpful in checking all of my data and numbers in the endless tables in this document. Thank you for knowing when I got too frustrated and insisted I take a break when I got too frustrated and I hope a similar future in academia awaits you.

TABLE OF CONTENTS

ABSTRACT ii

LIST OF TABLES xii

LIST OF FIGURES xvi

CHAPTER 1 INTRODUCTION 1

Purpose of the Study 5

Evaluation criteria 7

Research Questions 7

Organization of Dissertation 8

CHAPTER 2 REVIEW OF THE LITERATURE 9

Description of Vertical Scaling 9

Vertical Scaling Designs 17

Vertical Scaling Data Collection Design. 17

Traditional Vertical Scaling Methods. 19

IRT Logistic Models. 20

IRT Estimation Methods. 23

IRT Scaling Methods. 24

Concurrent Calibration. 24

Fixed Common Item Parameters. 27

IRT Linking Methods. 29

Mean-Mean. 30

Mean-Sigma. 31

Stocking and Lord Test Characteristic Curve Method. 32

Hybrid Test Characteristic Curve Method. 35

Literature Review 36

Real Data Studies. 37

Traditional and IRT Scaling Comparisons. 37

IRT Scaling Comparisons. 39

Simulation Studies. 48

Traditional and IRT Scaling Comparisons. 48

IRT Scaling Comparisons. 53

Shortcomings of Reported Research 62

CHAPTER 3 METHODS 65

Research Design 65

Sample 65

Procedure. 67

Calibration and Data Simulation Procedures. 67

Scaling Procedures. 72

Dependent Variables 78

Scoring Procedure. 78

Cut-scores rescaling. 79

Evaluation Measures. 81

Evaluation of Assignments to the three proficiency levels. 81

Evaluation of IRT outcomes. 84

CHAPTER 4 RESULTS: READING 85

Decision Consistency and Accuracy 86

Normal distribution 1,500 sample size 86

Normal distribution 3,000 sample size 91

Skewed distribution 1,500 sample size 94

Skewed distribution 3,000 sample size 97

Summary for Decision Accuracy and Decision Consistency 100

Conditional Standard Error 101

Normal distribution 1,500 sample size 101

Normal distribution 3,000 sample size 104

Skewed distribution 1,500 sample size 106

Skewed distribution 3,000 sample size 108

Summary for Conditional Standard Error 109

Root-Mean-Squared-Difference 110

Normal distribution 1,500 sample size 110

Normal distribution 3,000 sample size 114

Skewed distribution 1,500 sample size 115

Skewed distribution 3,000 sample size 116

Summary for Root-Mean-Squared-Difference 118

Correlations between Item Parameters 118

Normal distribution 1,500 sample size 118

Normal distribution 3,000 sample size 122

Skewed distribution 1,500 sample size 125

Skewed distribution 3,000 sample size 128

Summary for Correlations.. 131

CHAPTER 5 RESULTS: MATHEMATICS..132

Decision Consistency and Accuracy.. 132

Normal distribution 1,500 sample size 132

Normal distribution 3,000 sample size 138

Skewed distribution 1,500 sample size 141

Skewed distribution 3,000 sample size 145

Summary for Decision Accuracy and Decision Consistency 148

Conditional Standard Error 149

Normal distribution 1,500 sample size 149

Normal distribution 3,000 sample size 152

Skewed distribution 1,500 sample size 153

Skewed distribution 3,000 sample size 155

Summary of Conditional Standard Error 156

Root-Mean-Squared-Difference 157

Normal distribution 1,500 sample size 157

Normal distribution 3,000 sample size 161

Skewed distribution 1,500 sample size 162

Skewed distribution 3,000 sample size 163

Summary of Root-Mean-Squared-Difference................................ 164

Correlations between Item Parameters 164

Normal distribution 1,500 sample size 164

Normal distribution 3,000 sample size 167

Skewed distribution 1,500 sample size 170

Skewed distribution 3,000 sample size 173

Summary of Correlations 175

CHAPTER 6 DISCUSSION AND CONCLUSIONS 177

Summary of Study and Research Questions 177

Summary of Findings 178

Vertical Scaling Procedures 179

Concurrent Calibration 179

Fixed Common Item Parameters. 180

Test Characteristic Curve... 181

Hybrid Characteristic Curve. 182

Distribution Type 183

Content Area 184

Sample Size... 186

Limitations of the Study.. 186

Discussion 187

Distribution type 187

Content area 188

Standard setting outcome measures 190

Conclusions... 192

Recommendations for Practice 194

Recommendations for Future Research 196

REFERENCES 197

Appendix A: BILOG-MG and ICL command files...209

Appendix B: Results Tables Scale Score Mean and Standard Deviations, Decision Accuracy and Consistency, Conditional Standard Error, and Root-Mean-Squared-Difference 223

LIST OF TABLES

Table 1. *Approximate Student Population per Grade* 65

Table 2. *Number of items per form for Reading and Mathematics* 67

Table 3. *Reading and Mathematics Mean item parameters for the on-level and off-level test forms by grade*.. 69

Table 4. *Mean thetas for the normal and negatively skewed distributions* 70

Table 5. *Final Cut-scores for Reading and Mathematics*................................... 81

Table 6. *Decision Accuracy and Consistency for Vertical Scaling for Reading, normal distribution, 1,500 sample size*... 87

Table 7. *Conditional Standard Error for Reading, normal distribution, 1,500 sample size condition* 102

Table 8. *Root-Mean-Squared-Difference for Vertical Scaling procedures for Reading, normal distribution, 1,500 sample size*.. 111

Table 9. *Correlations of item parameters for Vertical Scaling for Reading, normal distribution, 1,500 sample size*... 120

Table 10. *Correlations of item parameters for Vertical Scaling for Reading, normal distribution, 3,000 sample size*... 123

Table 11. *Correlations of item parameters for Vertical Scaling for Reading, skewed distribution, 1,500 sample size*... 126

Table 12. *Correlations of item parameters for Vertical Scaling for Reading, skewed distribution, 3,000 sample size*... 129

Table 13. *Decision Accuracy and Consistency for Vertical Scaling for Mathematics, normal distribution, 1,500 sample size* 134

Table 14. *Conditional Standard Error for Mathematics, normal distribution, 1,500 sample size* 150

Table 15. *Root-Mean-Squared-Difference for Vertical Scaling procedures for Mathematics, normal distribution, 1,500 sample size* 158

Table 16. *Correlations of item parameters fors Vertical Scaling for Mathematics, normal distribution, 1,500 sample size* 165

Table 17. *Correlations of item parameters for Vertical Scaling for Mathematics, normal distribution, 3,000 sample size* 168

Table 18. *Correlations of item parameters for Vertical Scaling for Mathematics, skewed distribution, 1,500 sample size* 171

Table 19. *Correlations of item parameters for Vertical Scaling for Mathematics, skewed distribution, 3,000 sample size* 174

Table B1. *Scale Score mean and standard deviations for Reading normal conditions* 224

Table B2. *Scale Score mean and standard deviations for Reading skewed conditions* 225

Table B3 *Scale Score mean and standard deviations for Mathematics normal conditions* 226

Table B4 *Scale Score mean and standard deviations for Mathematics skewed conditions* 227

Table B5. *Decision Accuracy and Consistency for Vertical Scaling for Reading, normal distribution, 3,000 sample size* 228

Table B6. *Decision Accuracy and Consistency for Vertical Scaling for Reading, skewed distribution, 1,500 sample size* 229

Table B7. *Decision Accuracy and Consistency for Vertical Scaling for Reading, skewed distribution, 3,000 sample size* 230

Table B8. *Conditional Standard Error for Reading, normal distribution, 3,000 sample size* 231

Table B9. *Conditional Standard Error for Reading, skewed distribution, 1,500 sample size* 232

Table B10. *Conditional Standard Error for Reading, skewed distribution, 3,000 sample size* 233

Table B11. *Root-Mean-Squared-Difference between Vertical Scaling procedures for Reading, normal distribution, 3,000 sample size* 234

Table B12. *Root-Mean-Squared-Difference between Vertical Scaling procedures for Reading, skewed distribution, 1,500 sample size* 235

Table B13. *Root-Mean-Squared-Difference between Vertical Scaling procedures for Reading, skewed distribution, 3,000 sample size* 236

Table B14. *Decision Accuracy and Consistency for Vertical Scaling for Mathematics, normal distribution, 3,000 sample size* 237

Table B15. *Decision Accuracy and Consistency for Vertical Scaling for Mathematics, skewed distribution, 1,500 sample size* 238

Table B16. *Decision Accuracy and Consistency for Vertical Scaling for Mathematics, skewed distribution, 3,000 sample size* 239

Table B17. *Conditional Standard Error for Mathematics normal distribution, 3,000 sample size* 240

Table B18. *Conditional Standard Error for Mathematics skewed distribution, 1,500 sample size* 241

Table B19. *Conditional Standard Error for Mathematics skewed distribution, 3,000 sample size* 242

Table B20. *Root-Mean-Squared-Difference between Vertical Scaling procedures for Mathematics, normal distribution, 3,000* 243

sample size 243

Table B21. *Root-Mean-Squared-Difference between Vertical Scaling procedures for Mathematics, skewed distribution, 1,500* 244

sample size 244

Table B22. *Root-Mean-Squared-Difference between Vertical Scaling procedures for Mathematics, skewed distribution, 3,000* 245

sample size 245

LIST OF FIGURES

Figure 1. Data collection design across grade levels.. 19

Figure 2. Example Item Characteristic Curve 22

Figure 3. Scaling design for Concurrent Calibration.. 26

Figure 4. Scaling design for Fixed Common Item Parameters 28

Figure 5. Scaling design for Test Characteristic Curve...................................... 34

Figure 6. Scaling design for Hybrid Characteristic Curve 36

Figure 7. Decision accuracy for Reading, normal distribution, 1,500 sample size 89

Figure 8. Decision consistency for Reading, normal distribution, 1,500 sample size 90

Figure 9. Decision accuracy for Reading, normal distribution, 3,000 sample size 92

Figure 10. Decision consistency for Reading, normal distribution, 3,000 sample size 93

Figure 11. Decision accuracy for Reading, skewed distribution, 1,500 sample size 94

Figure 12. Decision consistency for Reading, skewed distribution, 1,500 sample size 96

Figure 13. Decision accuracy for Reading, skewed distribution, 3,000 sample size 97

Figure 14. Decision consistency for Reading, skewed distribution, 3,000 sample size 99

Figure 15. Conditional Standard Error for Reading, normal distribution, 1,500 sample size 103

Figure 16. Conditional Standard Error for Reading, normal distribution, 3,000 sample size 105

Figure 17. Conditional Standard Error for Reading, skewed distribution, 1,500 sample size 106

Figure 18. Conditional Standard Error for Reading, skewed distribution, 3,000 sample size 108

Figure 19. Root-Mean-Squared-Difference for Reading, normal distribution, 1,500 sample size 113

Figure 20. Root-Mean-Squared-Difference for Reading, normal distribution, 3,000 sample size 114

Figure 21. Root-Mean-Squared-Difference for Reading, skewed distribution, 1,500 sample size 115

Figure 22. Root-Mean-Squared-Difference for Reading, skewed distribution, 3,000 sample size 117

Figure 23. Decision accuracy for Mathematics, normal distribution, 1,500 sample size 136

Figure 24. Decision consistency for Mathematics, normal distribution, 1,500 sample size 137

Figure 25. Decision accuracy for Mathematics, normal distribution, 3,000 sample size 139

Figure 26. Decision consistency for Mathematics, normal distribution, 3,000 sample size 140

Figure 27. Decision accuracy for Mathematics, skewed distribution, 1,500 sample size 142

Figure 28. Decision consistency for Mathematics, skewed distribution, 1,500 sample size 144

Figure 29. Decision accuracy for Mathematics, skewed distribution, 3,000 sample size 145

Figure 30. Decision consistency for Mathematics, skewed distribution, 3,000 sample size 147

Figure 31. Conditional Standard Error for Mathematics, normal distribution, 1,500 sample size 151

Figure 32. Conditional Standard Error for Mathematics, normal distribution, 3,000 sample size 152

Figure 33. Conditional Standard Error for Mathematics, skewed distribution, 1,500 sample size 154

Figure 34. Conditional Standard Error for Mathematics, skewed distribution, 3,000 sample size 155

Figure 35. Root-Mean-Squared-Difference for Mathematics, normal distribution, 1,500 sample size 160

Figure 36. Root-Mean-Squared-Difference for Mathematics, normal distribution, 3,000 sample size 161

Figure 37. Root-Mean-Squared-Difference for Mathematics, skewed distribution, 1,500 sample size 162

Figure 38. Root-Mean-Squared-Difference for Mathematics, skewed distribution, 3,000 sample size 163

Figure A1. BILOG-MG initial run sample syntax file 210

Figure A2. BILOG-MG Concurrent Calibration sample syntax file 211

Figure A3. BILOG-MG Fixed Common Item Parameter Reading grade 4 sample syntax file 212

Figure A4. BILOG-MG Fixed Common Item Parameter Reading grade 5 sample syntax file 213

Figure A5. BILOG-MG Fixed Common Item Parameter Reading grade 7 sample syntax file 214

Figure A6. BILOG-MG Fixed Common Item Parameter Reading grade 8 sample syntax file 215

Figure A7. BILOG-MG Fixed Common Item Parameter Mathematics grade 4 sample syntax file 216

Figure A8. BILOG-MG Fixed Common Item Parameter Mathematics grade 5 sample syntax file 217

Figure A9. BILOG-MG Fixed Common Item Parameter Mathematics grade 7 sample syntax file 218

Figure A10. BILOG-MG Fixed Common Item Parameter Mathematics grade 8 sample syntax file 219

Figure A11. BILOG-MG Test Characteristic Curve sample syntax file 220

Figure A12. BILOG-MG Hybrid Test Characteristic Curve sample syntax file 221

Figure A13. ICL theta scoring sample syntax file 222

CHAPTER 1 INTRODUCTION

One of the biggest challenges in education today is quantifying how much a student has learned over time. There are many methods teachers use to evaluate student performance, but testing a student's knowledge and skills through examinations is the most prevalent. Assessments throughout a school year can provide some evidence of learning for that year. However, measuring learning is more complex and difficult over longer periods of time, for example across grade levels. One way to accomplish this is to develop a vertical scale. Vertical scaling is defined as a process of placing scores on a common scale obtained from tests administered at different grade levels or ages that are intended to measure similar constructs but that differ in difficulty (Kolen & Brennan, 2004).

Vertical scaling is used to create scales across several grades for both test batteries and achievement tests (Hoover, Dunbar, & Frisbie, 2003; Ito, Sykes, & Yao, 2008; Karkee, Lewis, Hoskens, Yao, & Haug, 2003; Lohman & Hagen, 2002). However there are many different methods for creating vertical scales. While there is extensive use of vertical scales in practice, the "gold standard" for the development of vertical scales is still being pursued.

Early research on vertical scaling led to the development of Thurstone scaling, which was the first proposed method to create a vertical scale (Thurstone, 1925, 1938). As computers evolved in power new procedures were developed. One such development involved the use of item response theory (IRT; Lord, 1980). IRT was first used to equate single grade levels in the early 1980s. Subsequently, the single grade procedures were extended across grades, particularly for the one parameter model with the results compared to the results obtained from Thurstone scaling (Becker & Forsyth, 1992; Camilli, Yamamoto, & Wang, 1993; Cook & Douglass, 1982; Harris, 1991; Petersen, Cook, & Stocking, 1983; Phillips, 1986; Williams, Pommerich, & Thissen, 1998). More sophisticated models were developed where the two- and three-parameter models were used to establish vertical scales (Baker & Al-Karni, 1991; Becker & Forsyth, 1992; Briggs & Weeks, 2009; Camilli, et al., 1993; Chin, Kim, & Nering, 2006; Clemans, 1993; Cook & Douglass, 1982; Hanson & Béguin, 2002; Harris, 1991; Harris & Hooker, 1987; Ito, et al., 2008; Jodoin, Keller, & Swaminathan, 2003; Keller, Skorupski, Swaminathan, & Jodoin, 2004; J. Kim, 2007; S. H. Kim & Cohen, 1998; Kolen, 1981; Meng, 2007; Petersen, et al., 1983; Skaggs & Lissitz, 1988; Tong, 2005; Tong & Kolen, 2007; Williams, et al., 1998; Yen & Burket, 1997). These models were introduced as model-data fit issues resulted with the less complex models.

Concurrent calibration was the first method used for either the one-, two-, or three- parameter models (Briggs & Weeks, 2009; Camilli, et al., 1993; Chin, et al., 2006; Custer, Omar, & Pomplun, 2006; Hanson & Béguin, 2002; Harris, 1991; Harris & Hooker, 1987; Holmes, 1982; Ito, et al., 2008; Jodoin, et al., 2003; Karkee, et al., 2003; Keller, et al., 2004; J. Kim, 2007; S. H. Kim & Cohen, 1998; Meng, 2007; Petersen, et al., 1983; Pomplun, Omar, & Custer, 2004; Shen, 1993; Skaggs & Lissitz, 1988; Tong, 2005; Tong & Kolen, 2007; Williams, et al., 1998). Other methods included the fixed common item parameter procedure (Becker & Forsyth, 1992; Jodoin, et al., 2003; Keller, et al., 2004; J. Kim, 2007), mean-mean procedure (Baker & Al-Karni, 1991; Guskey, 1981; Hanson & Béguin, 2002; Skaggs & Lissitz, 1988), mean-sigma procedure (Chin, et al., 2006; Hanson & Béguin, 2002; Jodoin, et al., 2003; Keller, et al., 2004; Tong, 2005; Tong & Kolen, 2007), Haebara procedure (Hanson & Béguin, 2002), and the Stocking and Lord procedure (Baker & Al-Karni, 1991; Briggs & Weeks, 2009; Clemans, 1993; Hanson & Béguin, 2002; Karkee, et al., 2003; Keller, et al., 2004; J. Kim, 2007; S. H. Kim & Cohen, 1998; Meng, 2007; Tong, 2005; Tong & Kolen, 2007; Yen & Burket, 1997).

The different methods have been compared in approximately 20 studies using real or actual test data (Briggs & Weeks, 2009; Camilli, et al., 1993; Cook & Douglass, 1982; Guskey, 1981; Harris, 1991; Harris & Hooker, 1987; Holmes, 1982; Ito, et al., 2008; Jodoin, et al., 2003; Karkee, et al., 2003; J. Kim, 2007; Kolen, 1981; Petersen, et al., 1983; Phillips, 1986; Rentz & Bashaw, 1977; Shen, 1993; Slinde & Linn, 1978, 1979; Whitely & Dawis, 1974; Williams, et al., 1998). Taken together, these studies reveal that the different methods yield inconsistent results where different procedures in each study seem to be the best procedure. Consequently, it is not easy to recommend one vertical scaling method over the others.

In order to systematically determine which procedure might be best 13 simulation studies have been conducted (Baker & Al-Karni, 1991; Chin, et al., 2006; Clemans, 1993; Custer, et al., 2006; Gustafsson, 1979; Hanson & Béguin, 2002; Keller, et al., 2004; S. H. Kim & Cohen, 1998; Meng, 2007; Pomplun, et al., 2004; Skaggs & Lissitz, 1988; Tong, 2005; Tong & Kolen, 2007; Yen &

Burket, 1997). While these studies have evaluated vertical scaling more systematically than those using real data, many of the studies did not clearly delineate which scaling procedure produced the best results for all of the features that are important operationally. For example, all of the studies based their simulated data on only one content area and many included only two grade levels.

There are two aspects of the previous research on vertical scaling that have not been systematically evaluated in one study. One aspect is the properties of the data from which the real data has been simulated. For example, several of the studies were based on response data from a content area such as Reading or Mathematics. Only three studies compared more than one content area at the same time, and they were conducted with real data (Ito, et al., 2008; J. Kim, 2007; Tong & Kolen, 2007). Furthermore, in the case of simulation studies, a normal distribution of scores was assumed in all of the studies except the Custer et al.'s (2006) study in which "slightly" skewed distributions were simulated.

The second aspect not considered in the previous studies is the use of evaluation methods that would be most useful to practitioners in the field (e.g., state testing officials). For example, Jodoin et al. (2003) placed examinees in proficiency categories based on cut-scores, but the results across the methods were different and, since the data were real, it is unclear which procedure provided the most accurate and consistent results. In another study, Meng (2007) evaluated the absolute bias, standard error (SE), and root-mean-squared-error (RMSE) between the true versus estimated proportion classification values, which, while important indicators of success, are not easily understood by practitioners. Measures such as the accuracy and consistency of the decisions made using the cut-scores were not evaluated in the context of vertical scaling procedures in either of these two studies.

Purpose of the Study

To address these gaps in the literature the purpose of this study was to evaluate four different vertical scaling procedures in terms of classification accuracy and consistency, conditional standard errors of the cut scores, root-mean-squared-differences (RMSD) between scale scores, and correlations between scaled item parameters. Therefore, a simulation study based on real data acquired during the implementation of a vertical scale was conducted. The real data, together with cut-scores were used to develop a simulation study that mirrored actual practice.

The following four scaling procedures were examined:

1. Concurrent calibration (CC). This method simultaneously calibrates all of the item parameters and theta values for all grades at the same time. No linking or equating methods are used to place the item parameters or theta values on the same scale.
2. Fixed common item parameters (FCIP). This method sequentially calibrates the item parameters for two test forms for each grade level. The base level grade 6 item parameters are estimated first. Then the separate grade item parameters are used to calibrate the upper and lower grade levels (e.g., start at grade 6 and move down to grade 5), where the known parameters are fixed (e.g., grade 6 item parameters) and the unknown

parameters (e.g., grade 5 item parameters) are estimated. Once all grade combinations are estimated the final theta values are estimated based on the final item parameters.

3. Test characteristic curve (TCC). Similar to the FCIP procedure, the TCC procedure is sequenced. First, the item parameters for each grade are separately estimated. Then the item parameters for each grade other than the base grade level are transformed onto the scale of the base grade level using an IRT equating method. In this study that method will be the Stocking and Lord (1983) IRT equating procedure.
4. Hybrid characteristic curve (HCC). This method is similar to both the CC and TCC methods. Instead of calibrating the separate grade levels as is done for TCC, two or more grade consecutive levels are combined to conduct a concurrent calibration (e.g., grade 3 and 4, grade 5 and 6, and grade 7 and 8) and then the three separate grade grouping item parameters are transformed onto the same scale. In this example the concurrent calibration is smaller as only two grades levels are included instead of six to prevent difficulties in estimating the parameters. The TCC equating procedure is the method used to place the item parameters onto the scale of the base grade level.

Four factors were evaluated: scaling method (four procedures), distribution shape (normal, negatively skewed), content area (Reading, Mathematics), and sample size (1,500, 3,000). The four factors resulted in a fully crossed 4 x 2 x 2 x 2 design with 32 conditions.

Evaluation criteria

Five statistical procedures were used to evaluate the results: decision consistency and accuracy, conditional standard error at each cut-score, root-mean-squared-differences of the final scale scores, and correlations of the final parameters across vertical scaling methods. The purpose of the first three of the outcome measures was to evaluate the four scaling methods across the conditions in a criterion-referenced score interpretation context. The first method was decision accuracy, which measured how accurate the decisions were in placing students into one of three proficiency categories. The second method was decision consistency, which measured how consistent the decisions were in placing students in these categories (Crocker & Algina, 1986; Livingston & Lewis, 1995). The third measure was the conditional standard error at each cut-score, which measured the relative error for the theta score at each cut-score.

The last two evaluation criteria were used to evaluate the vertical scaling procedures commonly used in other vertical scaling research. The fourth criteria, the root-mean-squared-difference (RMSD) measured the agreement between the scale scores for the pairs of each vertical scaling procedure. The fifth criteria, the correlation between the item parameters obtained for each scaling method, measured the agreement between the item parameters found by the pairs of vertical scaling procedures.

Research Questions

The following four research questions were addressed using simulated data based on real data:

1. Do vertical scaling methods perform the same for the five evaluation criteria?
2. Does distribution shape have an effect on the five evaluation criteria?
3. Does content area have an effect on the five evaluation criteria?
4. Does sample size have an effect on the five evaluation criteria?

Organization of Dissertation

An introduction to the research on different vertical scaling procedures, purpose of the study, brief description of the procedures and the evaluation criteria used and the presentation of the simulation conditions and research questions were presented in Chapter 1. Chapter 2 contains the context in which vertical scaling occurs followed by an explanation of the vertical scaling procedures, an extensive literature review, and de-limitations of the current study. Chapter 3 describes the methods used in the present study. This includes a description of the vertical scaling procedures, simulation conditions, and evaluation procedures used. The results for the Reading conditions are presented in Chapter 4. The results of the Mathematics conditions are presented in Chapter 5. The discussion, conclusions, recommendations for practice, and recommendations for future research are presented in Chapter 6.

CHAPTER 2 REVIEW OF THE LITERATURE

Chapter 2 is organized into three main sections. Section 1 provides a description of vertical scales used in practice and the rationale for why vertical scales are created. In Section 2 the data collection design used in this study is described first. Then a brief description of traditional vertical scaling procedures is presented. As well, the item response models and estimation methods are described. The four vertical scaling procedures evaluated in this study are then described in greater detail: Concurrent Calibration (CC), Fixed Common Item Parameters (FCIP), Stocking and Lord test characteristic curve (TCC), and Hybrid test characteristic curve (HCC). Within the subsection on the four scaling procedures, the mean/mean and mean/sigma equating methods are briefly described as the test characteristic curve procedures are similar to these methods. Finally, the last section provides a summary and critical analysis of the research literature on vertical scaling methods. Real data studies are reviewed first followed by simulation studies. Studies that compared the IRT methods to traditional methods are briefly presented, followed by studies that compare IRT vertical scaling methods.

Description of Vertical Scaling

Educational research is focused on many areas that involve student learning. One method to evaluate some aspects of this learning is through assessments. In many cases learning is defined as student growth. One purpose of school achievement or aptitude tests is to estimate individual student growth, either within a year, year-to-year or over multiple years.

To adequately measure growth it is necessary to administer achievement tests across time. There are different methods to estimate growth. One method is to administer the same test each year and chart growth in test scores over multi-year periods (Kolen & Brennan, 2004). However, administering the same test items over a wide range of educational levels is problematic as many items are likely too difficult for students at early educational levels and too easy for students at higher educational levels (Kolen & Brennan, 2004). In addition, pre- and post-test gain scores are difficult to measure due to regression (Glass & Hopkins, 1996). One alternative method to overcome these two problems is to use similar test forms that measure the same construct and the process of vertical scaling.

The purpose of vertical scaling is to allow comparisons of students or cohorts at different grade levels by administering different test forms that measure the same general construct and that are constructed based on the range of item difficulty for each group of students for each testing occasion. A vertical scale consists of separate test forms that differ in difficulty to reflect the differences in learning by age or grade, but which are intended to measure the same constructs; vertical scaling involves placing the scores obtained from these tests on the same scale (Kolen & Brennan, 2004). Vertical scales are created for constructs that remain similar across educational levels, such as the constructs of Reading, Writing, and Mathematics, especially those created in the context of No Child Left Behind legislation. They are not typically used for constructs which change across increasing educational levels, such as the constructs of science and social studies. In the case of educational achievement assessments such an overlap in content may be very different (e.g., Social Studies grade 3 – California history and Social Studies grade 4 – United States government). Vertical scales provide the ability to measure growth on a common metric so that the interpretation of educational growth is easier.

Two types of tests that can use vertical scales are test batteries and educational achievement tests. For example, the *Iowa Test of Basic Skills* (ITSB; Hoover, et al., 2003) and the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2002) are test batteries that are typically administered to students in consecutive grades. However, test batteries are sometimes created without using a vertical scale. Norms are created using representative samples of students for the population of students at each of the identified grade levels. In this case, the purpose of the vertical scale is not to measure growth per se, but to compare students or schools to the appropriate norms developed from the responses of the students in the different grade levels. For example, results from students in a particular school can be compared to the national population using national school norms at each grade level. These batteries include tests in several content areas. An example of a normed test battery is the *TerraNova, The Second Edition (CAT/6)*, which includes two different forms (C and D), K-12 educational levels, and multiple content areas (e.g., Reading/Language Arts, Mathematics, Science, Social Studies, Word Analysis, Vocabulary, Language Mechanics, Spelling, and Mathematics Computations; CTB/McGraw-Hill, 2001). In this case the vertical scale is constructed for norming information but in educational achievement testing, but vertical scaling is not constructed for some of these content areas (i.e., Science and Social Studies). This is primarily due to the content specifications of test batteries not being based on general knowledge where an overlap across grades is measured and not for a specific state assessment.

Educational achievement tests have not been necessarily administered at adjacent grade levels like many test batteries. For example, many state testing programs included non adjacent grade levels, such as grades 2, 5, 7, and 9. More recently vertical scaling methods have been used to scale educational achievement tests that are constructed for multiple adjacent grade levels so that growth across grades can be measured.

The No Child Left Behind Act of 2001 (NCLB, 2001) changed the pattern of nonadjacent grades for educational achievement tests in the United States, where students in consecutive grades 3 through 8 are currently tested in Reading and Mathematics, and will be required to be tested in Science and Social Studies in the near future. However, vertical scaling will not likely be used in the content areas of Science and Social Studies as the constructs in these areas may be too different. However, vertical scaling is still used when creating national schools norms for these content areas. As of early 2009, 39 states were fully compliant and had passed Federal Peer Review, 11 states were not fully compliant (10 states did not test all the required grades in Science and one state had not passed the Federal review due to using inappropriate content standards for Mathematics; U.S. Department of Education, January, 2009). To become fully compliant each state must pass Federal peer review that indicates that their state assessment that adheres to the NCLB legislation (a full description can be found on the U.S. Department of Education website; Lead and manage my school: The standards and assessments peer review Program Overview, U.S. Department of Education, n.d.). The NCLB legislation also requires that states achieve 100% proficiency by 2014 (Linn, Baker, & Betebenner, 2002). To measure whether states achieve this growth, Adequate Yearly Progress (AYP) measures are used to evaluate the magnitude of growth which is used as part of the accountability program within each state. There is no clear definition of how AYPs are to be calculated, but a measure of learning over time is necessary, which could include year to year comparisons of cohorts or a longitudinal comparison of the same cohort over time (see the U.S. Department of Education summary by the National Title I Directors' Conference for a full description; U.S. Department of Education, 2003). Although AYPs have been defined differently in different states, it is common for states to use vertical scales to measure the states' AYP (e.g., Arizona, Iowa; Lead & manage my school. Letters to chief of state school officiers regarding an update on several NCLB cornerstones, U.S. Department of Education, n.d.).

Given several vertical scaling procedures are available, different states employ different procedures, albeit with Federal oversight by the Department of Education (i.e., Federal Peer Review). Unfortunately, different vertical scaling procedures do not always provide the same results. Consequently, it may be difficult to compare validly the AYPs across states. Therefore, there is a need for additional research to evaluate different methods for vertical scaling educational achievement tests and to determine if one procedure is clearly superior.

Before describing vertical scaling, three types of linking are described: (1) equating, (2) scaling and (3) predicting (sometimes called linking). Linking is used to derive a metric for which student scores are "comparable" (Holland, 2007). The statistical mechanisms used to link two or more tests are generally similar. However, the requirements for and strength of interpretations are different for each type of linking. Equating is the most demanding and strongest form of linking (Linn, 1993; Mislevy, 1992), and "a direct link is one that functionally connects the scores on one test directly to those of another" (Holland & Dorans, 2006, p. 188). According to Kolen and Brennan (2004) equating adjusts for difference in difficulty among forms that are built to be similar in difficulty and content (p. 2). Further, there are four properties of equating that must be met (1) the *symmetry property* – which requires that equating transformations be symmetric (i.e., where the function used to transform a score on one form to the second form be the inverse of the function used to transform a score on the second form), (2) the *same specifications property* – the different test forms must be built to the same content and statistical specifications, (3) the *equity property* – Lord (1980) stipulated that it must be a matter of indifference to each examinee whether one form or another is administered or that the distribution of true scores has the same distribution of converted scores on either form, and (4) the *group invariance property* – the equating relationship is the same regardless of the group of examinees used to conduct the equating (Kolen & Brennan, 2004). The strongest interpretations can be made with equating as the different forms or scores are considered the most comparable when the properties are met. The resulting scores are considered interchangeable in the sense that the students' score on one form would be the same on the other form.

Scaling is to align or transform the scores from two different tests onto a common scale to create comparable scores (Holland & Dorans, 2006; Kolen, 2006; Linn, 1993; Mislevy, 1992). Scaling is used in different types of situations: a) battery scaling - different constructs and a common population of examinees, b) anchor scaling - different constructs and different populations of examinees, c) vertical scaling - similar constructs and similar reliability but different difficulties and populations of examinees, d) calibration - same construct, different reliabilities, and the same population of examinees, and e) concordance - similar constructs, difficulty, reliability and the same population of examinees (Holland & Dorans, 2006). Scaling does not strictly meet any of the properties of equating, such as symmetry, same specification, equity, and group invariance. While the statistical mechanisms used for equating can be used for scaling, the scores are not as closely comparable as with equating. Therefore, the strong interpretations of interchangeable scores that are made with equating cannot be made with scaling. But the scores are considered comparable.

The final linking type is called predicting (sometimes called linking) where assessments are constructed around different types of tasks, administered under different conditions, or for different purposes related to students' affect and motivation, where the equating or scaling application could be misleading (Kolen, 2006; Linn, 1993; Mislevy, 1992). In a prediction study, a relationship is determined between the scores to be predicted from one test (representative of the dependent variable) from the scores obtained on the second test (representative of the independent variable; Mislevy, 1992). Predicting is achieved by matching a set of similar items to provide a “link” between two sets of test items (predicting is sometimes called linking by practitioners but I will use the term predicting throughout this paper). Kolen and Brenan (2004) provide an example of a predicting of the ACT Assessment composite scores and the SAT I Verbal-plus-Mathematics (V+M) scores. Both of these testing programs are used for college admissions in the U.S. While there are similarities in the content tested and the correlations between the scores are relatively high (usually in the low 0.90s), the forms are developed with a different table of specifications and do not provide interchangeable scores. The only requirement is that a relation between the two measures can be created. The interpretation of the “comparable” scores in predicting is the weakest, as compared to equating and scaling. All three linking procedures use the same statistical machinery to establish the link between test forms, the interpretation of the scores is qualitatively different, with the strongest interpretation with equating, followed by scaling and then predicting.

In the context of NCLB, while the general constructs across the full set of grades are similar, and the constructs between grades are similar, the conditions necessary for equating are not met. But the degree of comparability of the constructs is not so different to be a predicting link. Consequently, scaling for comparability is the appropriate form of linking for vertical scales. As such, the statistical procedure used to create a vertical scale is important to ensure the interpretations of scaled scores, while not as comparable as equated scores, are nevertheless comparable and can be interpreted across different grade levels.

Vertical Scaling Designs

Vertical Scaling Data Collection Design.

There are two basic data collection designs for vertical scaling, one based on randomly equivalent groups and the other based on non-equivalent groups. The randomly equivalent groups design requires that two or more test forms be randomly assigned to students. Given the ability of the students in the two groups is randomly equivalent, the performance on the test forms is linked through scaling procedures developed for a random equating design. In contrast, the students in a non-equivalent groups design are intact groups such as classrooms and schools. To overcome the non-randomness, the non-equivalent groups design requires a common set of items, presented in approximately the same location on the different test forms administered to the different groups. While the ability of the students on the common-items may differ between the two test forms administered to these groups, the common items are used to link the performance on the different tests using a common-item equating process (Kolen & Brennan, 2004).

In the context of creating a vertical scale both the randomly equivalent groups design and the common-item non-equivalent groups design can be used (Kolen & Brennan, 2004). However, when using different vertical scaling procedures the data collection design must conform to the requirements of the scaling procedures, and must be the first decision made when determining how to create the vertical scale. For the randomly equivalent groups design the on-level form(s) and off-level test form(s) are randomly assigned to random samples of students at different grade levels. The students' scores can then be placed on the same scale using concurrent calibration procedures. However, scaling across many grade levels would require that all test forms for each grade level are randomly assigned to all examinees. But this is not reasonable in most contexts. The common-item non-equivalent groups design can be used to overcome this problem. But what is now required is that the on- and off-level test forms contain a common set of items. The students' scores may then be linked using the common items with one of the CC, FCIP, TCC, and HCC procedures. Most vertical scaling designs involve both randomly equivalent and non-equivalent groups.

Figure 1 shows such a design. The design being used in this study involves administering the on-grade level and the off-grade level test forms randomly to grade level students. The solid horizontal lines indicate that two test levels are administered to two randomly assigned groups of students at the same grade level. For example, the grade 4 students were randomly assigned to either the grade 4 test or the below grade 3 test. The dashed diagonal lines indicate the same grade level tests (i.e., common items) but administered to two different grade level students. For example, two samples of grade 3 and grade 4 students are administered the two grade 3 level test forms. Thus both concurrent calibration within grade and/or common-item equating procedures between grades are needed to place the scores of the students in all grades on a common metric.

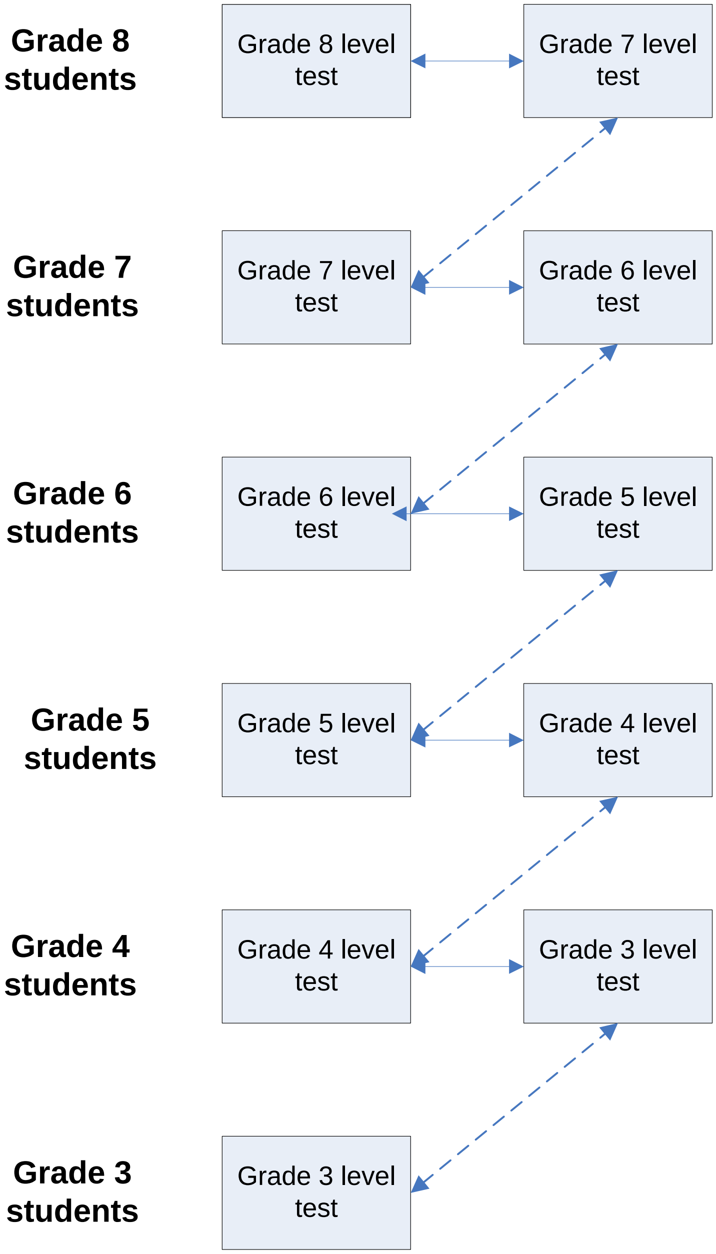

Figure 1. Data collection design across grade levels

Traditional Vertical Scaling Methods.

One of the first vertical scaling methods was proposed by Thurstone (1925). Thurstone’s Absolute Scaling method involves test items that are dichotomously scored and separate norms are constructed for successive age or grade levels. This method has an assumption that the scores are normally distributed within each age or grade level. In 1938, Thurstone suggested using transformed percentile ranks of examinees to create a vertical scale (as cited in Williams, et al., 1998). While Thurstone Absolute Scaling is computationally simpler than some other methods discussed below, the requirement of a normal distribution for each grade level might not be practical in the case of the tests used for NCLB. This legislation requires 100% proficiency, which will likely lead to negatively skewed distributions.

There are additional traditional equating methods, including linear (e.g., Tucker Observed Score; Levine Observed Score, and Tucker True Score) and equipercentile equating procedures. However, since these methods are more appropriate for horizontal equating, they are not included in the methods identified by Kolen and Brennan (2004) to establish vertical scales. Consequently, they are not discussed here.

IRT Logistic Models.

Item response theory (IRT) can be used to create vertical scales using the one- two- or three- parameter logistic IRT models. The one-parameter logistic model (1-PL) includes one item parameter, namely the item difficulty. The item characteristic curves (ICCs) for this model are given by the equation:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1,2, \ldots, n,$$

where $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly, b_i is the item i difficulty parameter, n is the number of items in the test, and e is a transcendental number whose value is 2.718 (Hambleton, Swaminathan, & Rogers, 1991, pp. 12-13). The two-parameter logistic model (2-PL) includes two item parameters: item discrimination and item difficulty. The ICCs for the (2-PL) are given by the equation:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1,2, \dots, n,$$

where $P_i(\theta)$, b_i, n, and e are defined as above, a_i, which is the discrimination parameter, and D is the scaling factor of 1.7 for item i (Hambleton, et al., 1991, p. 15). The three-parameter logistic model (3-PL) includes three parameters: item discrimination, item difficulty, and pseudo-guessing. The ICCs for the (3-PL) are given by equation:

$$P_i = P(X_i = 1|\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{[Da_i(\theta - b_i)]}} \qquad i = 1,2, \dots, n,$$

where $P_i(\theta)$, b_i, n, e, and a_i are defined as above and c_i is the pseudo-guessing parameter for item i (Hambleton, et al., 1991, p. 17). An item characteristic curve (ICC) can be used to represent the relationship to $P_i(\theta)$ and θ for each item. An example is provided in Figure 2 for an item with the *a*- parameter = 0.9 the *b*-parameter = 1.25 and the *c*- parameter = 0.18.

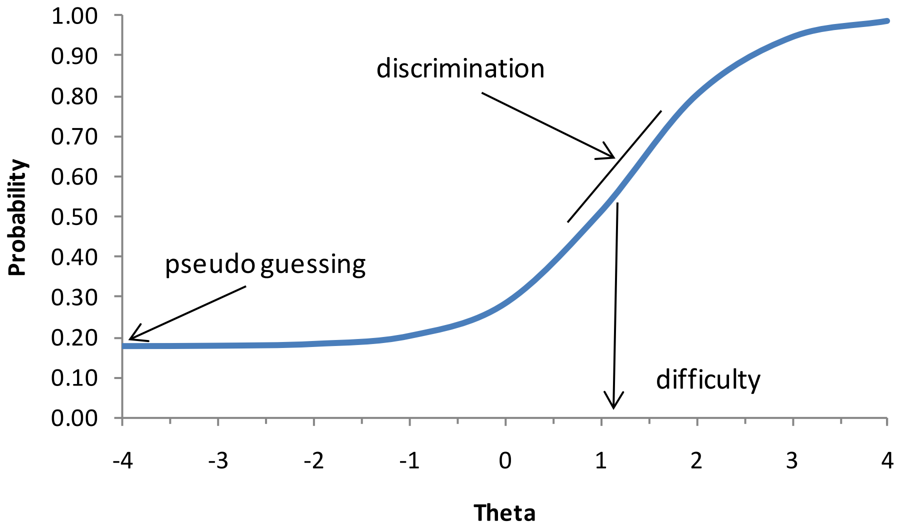

Figure 2. Example Item Characteristic Curve

There are three assumptions underlying the use of the 1-PL, 2-PL, and 3-PL logistic models. The first assumption is unidimensionality, where only one domain is measured by the items on a test (Hambleton, et al., 1991). According to Hambleton et al. (1991) this assumption cannot be strictly met because several cognitive, personality, and test-taking factors always affect test performance. However, it is expected that for the set of items on a test there is one *dominant* factor, for example, Mathematics ability. The second assumption is local independence – an examinee's responses to any pair of items are statistically independent (Hambleton, et al., 1991). After taking student's overall ability into account, there is no relationship between examinees' responses to different items. The third assumption is nonspeededness, that is, the test is a power test. A power test is a test for which a person's total correct score is equal to the number of items for which the person is not pressured by time and simply responds incorrectly (Oshima, 1994). One measure of speededness is the percentage of examinees completing the test, where the percentage of students completing 75% and 80% of the items are reviewed (Hambleton, et al., 1991).

IRT Estimation Methods.

Each of the logistic models can be used to simultaneously estimate the person latent abilities and each item parameter in the IRT model used. Three different estimation methods can be used: *maximum likelihood estimation* (MLE), Bayesian modal estimation or *maximum a posteriori* (MAP), or *expected a posteriori* estimation (EAP). Each estimation procedure has benefits and drawbacks. For the MLE procedure the data may not contain perfect and imperfect scores (i.e., 0% and 100%). The MAP estimator is similar to the EAP estimator but has a somewhat larger average total error. According to the authors of BILOG-MG "the EAP estimator exists for any answer pattern and has a smaller average error in the population than the other two estimators" (Zimowski, Muraki, Mislevy, & Bock, 1996, p. 17). Therefore, the EAP procedure will be used in this study.

EAP is conducted where the Bayes estimate is the mean of the posterior distribution of θ, given the observed response pattern x_i (Bock & Mislevy, 1982; as cited in Zimowski, et al., 1996, p. 16). This procedure is approximated by the Gaussian quadrature using the following function:

$$\overline{\theta}_i \cong \frac{\sum_{k=1}^{q} X_k P(x_i | X_k) A(X_k)}{\sum_{k=1}^{q} P(x_i | X_k) A(X_k)},$$

where X_k is one of q quadrature points (the number of quadrature points are indicated in the program syntax), $P(x_i|X_k)$ is the probability that $x = 1$ at the point θ on the ability continuum, and $A(X_k)$ are weights depending on the assumed distributions of θ (Bock & Mislevy, 1982; Zimowski, et al., 1996). A measure of the precision of $\bar{\theta}_i$ is the posterior standard deviation (PSD), which is approximated by

$$PSD(\overline{\theta}_i) \cong \frac{\sum_{k=1}^{q}(X_k - \overline{\theta}_i)^2 P(x_i|X_k)A(X_k)}{\sum_{k=1}^{q} P(x_i|X_k)A(X_k)},$$

where X_k, $P(x_i|X_k)$, and $A(X_k)$, and $\bar{\theta}_i$ are as defined above (Bock & Mislevy, 1982; Zimowski, et al., 1996).

IRT Scaling Methods.

Concurrent Calibration.

An IRT logistic model is typically used to calibrate a single test form administered at a single grade level. However, practitioners sometimes administer several forms at the same grade level or the same form is administered to multiple grade levels. The scores for these two situations can be linked simultaneously using concurrent calibration. Only one analysis is conducted, and separate linking procedures are not needed. A simultaneous calibration of multiple grade levels and/or forms can be conducted to estimate the latent ability (θ) of the students and the item parameters, providing the following requirement is met: the test forms administered must be randomly assigned to students at each grade level. The requirement of random assignment applies to both multiple forms and multiple grade levels. However, the random assignment of test forms is not necessarily inclusive of the entire grade span. Instead adjacent grade level test forms are randomly assigned.

An example of the scaling design is shown in Figure 3 where an ellipse indicates that all of the grade level tests are simultaneously estimated. Any one of the three logistic models can be used to conduct concurrent calibration (CC).

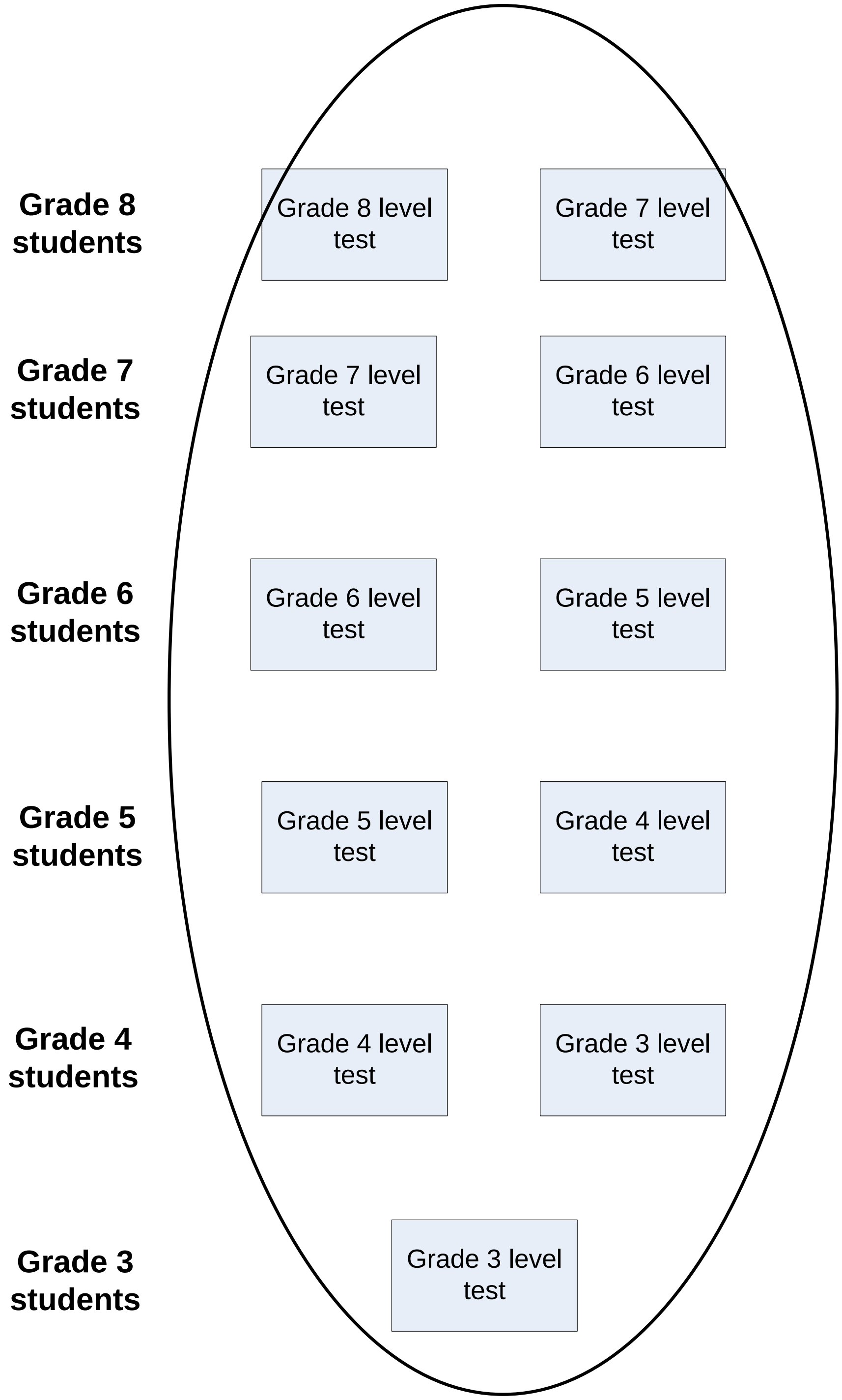

Figure 3. Scaling design for Concurrent Calibration

One data set is constructed that contains all of the item vectors for each person at each grade level to be scaled. In the case of a single grade concurrent calibration, BILOG-MG can be used to complete the estimation. Either BILOG-MG (Zimowski, et al., 1996) or MULTILOG (Thissen, 1991) can be used to perform the needed estimation when multiple grade levels are involved.

Fixed Common Item Parameters.

Fixed common item parameter estimation (FCIP) occurs when common items are used to link items in two test forms administered either at the same grade level or at different grade levels. In contrast to CC, FCIP estimation begins with identification of the base grade level. CC is first used to place the base grade level test forms on a common scale. For example, at the base grade 6 level the responses of the random sample of students for the grade 6 items and the responses of the random sample that responded to the adjacent grade 5 items are calibrated concurrently. This calibration is represented by the ellipse at the base level seen in Figure 4. The second step occurs when the parameters for the items for the base level are used in the other grade group calibrations.

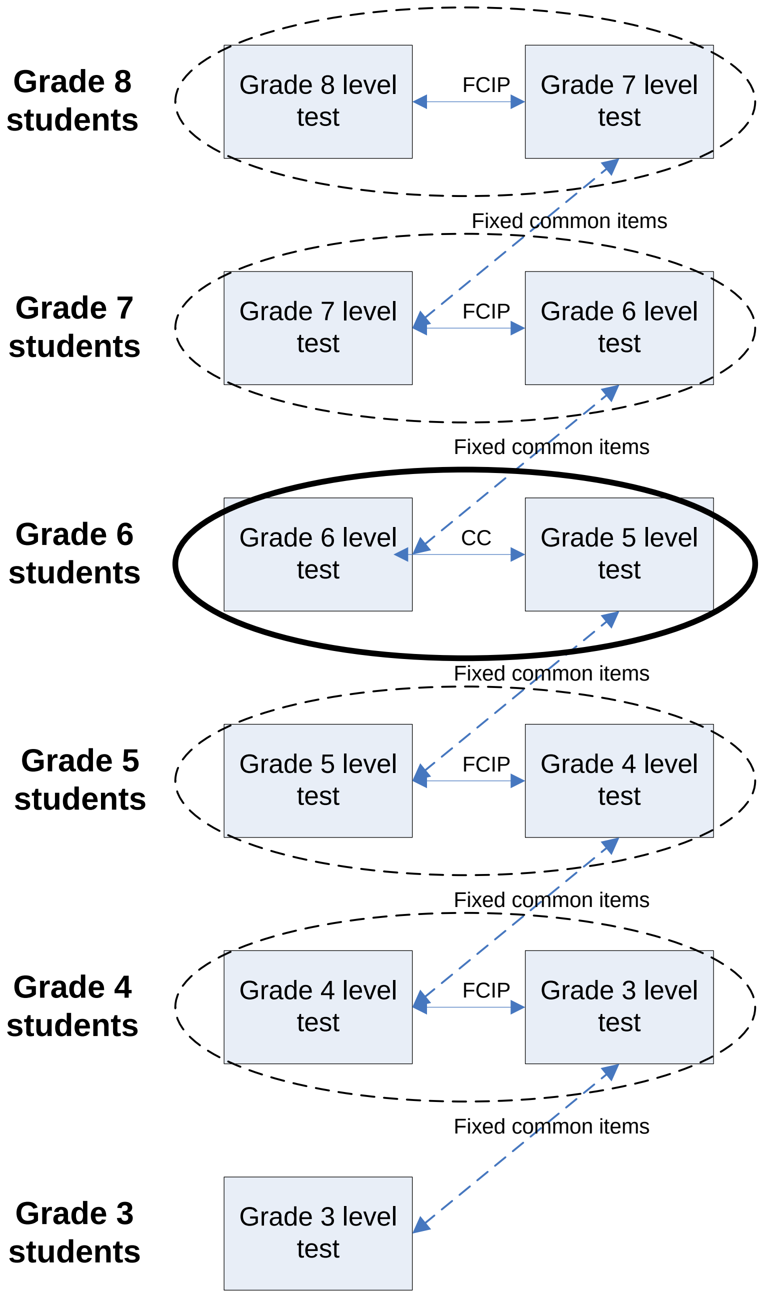

Figure 4. Scaling design for Fixed Common Item Parameters

The item parameters of the common items in the common form are fixed to their base values to establish a scale for the estimation of item parameters for the upper and lower level grade items. As shown in Figure 4, the item parameter values obtained from the previous calibration of the grade 6 test form with the grade 6 students are fixed for the grade 6 level test form(s) administered to a random sample of students in grade 7. FCIP is then conducted by fixing the item parameters for the grade 6 level test form(s) for the calibration of the grade 7 level test form(s). Therefore, the fixed item parameters are read into the computer program, and the syntax indicates that the non-fixed item parameters are to be estimated while at the same time fixing the item parameters indicated (the data for both test forms are included in the estimation process). The solid lines indicate a single calibration within a grade (e.g., the grade 5 and 6 tests administered to random sample of the grade 6 level students). The dashed lines and dashed ellipses indicate that the FCIP scaling employed the fixing of the parameters indicated for the subsequent calibration. This process occurs up and down the scale until all item parameters are estimated for each of the grade level students. In summary, the vertical scale in FCIP estimation fixing item parameters for grade levels with common items for each subsequent calibration.

IRT Linking Methods.

In addition to the CC and FCIP calibration methods, item parameters can be linked using a separate IRT calibration for each test and a linking procedure. Using IRT linking methods, the two grade level test forms for each of randomly different student samples are concurrently calibrated first (e.g., at grade 6, the grade 6 and grade 5 level test items are concurrently calibrated). This process is repeated separately for all of the grade level student groups (e.g., all grade level student groups are separately calibrated). Second, the item parameters between grade levels are linked using an IRT linking procedure. In this step the item parameters for the common test forms are used to transform the item parameters of the adjacent grade level items so that the same scale is established. Four different IRT equating procedures are commonly used in vertical scaling: Mean-Mean, Mean-Sigma, Stocking and Lord test characteristic curve and Hybrid test characteristic curve procedures.

Mean-Mean.

The Mean-Mean (MM) IRT equating method was first described by Loyd and Hoover (1980). The ability and item parameter estimates for the two groups (either two test forms or one form at two grade levels) are placed on a common scale using the item difficulty and discrimination estimates of the common items for each set of items (Loyd & Hoover, 1980). The θ values for the two scales are related as follows:

$$\theta_{2i} = A\theta_{1i} + B$$

where, *A* and *B* are the constants in the linear equation and θ_{1i} and θ_{2i} are values of θ for common item *i* on form 1 and form 2, respectively (Kolen & Brennan, 2004). The *A* constant is calculated using the mean of the *a*- discrimination - parameter estimates for the common items by the following equation:

$$A = \frac{\mu(a_1)}{\mu(a_2)},$$

where, $\mu(a_1)$ is the mean of the *a*- parameter estimates for the common items on form 1, and $\mu(a_2)$ is the mean of the *a*- parameter estimates for the common items on form 2 (Kolen & Brennan, 2004). The *B* constant is calculated using the mean of the *b*-difficulty - parameter estimates of the common items by the following equation:

$$B = \mu(b_2) - A\mu(b_1),$$

where, *A* is the constant defined above, $\mu(b_2)$ is the mean of the *b*- parameter estimates for the common items on form 2, and $\mu(b_1)$ is the mean of the *b*-parameter estimates for the common items on form 1 (Kolen & Brennan, 2004). Given the values for *A* and *B* the unique item parameters are transformed via the following equations:

$$a_{2i} = \frac{a_{1i}}{A},$$

where, a_{2i} is the *a*- parameter estimate for unique item *i* on form 2, and a_{1i} is the *a*- parameter estimate for unique item *i* on form 1,

$$b_{2i} = Ab_{1i} + B,$$

where, b_{2i} is the *b*- parameter estimate for unique item *i* on form 2, and b_{1i} is the *b*- parameter estimates for unique item *i* on form 1, and

$$c_{2i} = c_{1i},$$

where, c_{2i} is the *c*- pseudo guessing - parameter for item *i* on form 2, and c_{1i} is the *c*- parameter estimate for item *i* on form 1 (Kolen & Brennan, 2004).

Mean-Sigma.

The Mean-Sigma (MS) IRT equating method was first described by Marco (1977). This method is similar to the MM method where the *A* and *B* constants are estimated to transform the unique items on form 2. However, the calculation of the *A* constant is different. The *A* constant is calculated as follows:

$$A = \frac{\sigma(b_2)}{\sigma(b_1)},$$

where, $\sigma(b_2)$ is the standard deviation of the *b*- parameters on form 2, and $\sigma(b_1)$ is the standard deviation of the *b*- parameters on form 1 (Kolen & Brennan, 2004). The *B* constant is calculated using the formula for *B* for the MM method (Kolen & Brennan, 2004).

Stocking and Lord Test Characteristic Curve Method.

In contrast to the MM and MS procedures, a more precise estimation that takes into account more information for each item parameter can be conducted by matching test characteristics curves (TCC method; Stocking & Lord, 1983), which is the sum of the ICCs on a test form, or a subset of items such as common items. The MM and MS methods may be overly influenced by the differences in the *b*- parameter estimates as the MM and MS procedures does not consider all of the common item parameter estimates simultaneously unlike the TCC procedure (Kolen & Brennan, 2004). The TCC method overcomes this problem.

The TCC method minimizes the differences between the TCCs. The function in which the sums over all items of the squared differences between the two ICCs for each form being equated is calculated and given by:

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}\left(\theta_{2i}; \hat{a}_{2j}, \hat{b}_{2j}, \hat{c}_{2j}\right) - \sum_{j:V} p_{ij}\left(\theta_{2i}; \frac{\hat{a}_{1j}}{A}, A\hat{b}_{1j} + B, \hat{c}_{1j}\right)\right]^2.$$

The $SLdiff(\theta_i)$ is the function that defines the differences between the ICCs. The summation is taken over common items for each set of parameter estimates before squaring (Kolen & Brennan, 2004). The function

$$\tau(\theta_i) = \sum_j p_{ij}(\theta_i)$$

is referred to as the *test characteristic curve* (Kolen & Brennan, 2004), where *j* represents the common items (j:V in the above equation calls for summation over the common items). Therefore, the test characteristic curve is the sum of the item probabilities across items for that test form. The expression $SLdiff(\theta_i)$ is the squared difference between the test characteristic curves for a given θ_i. These differences are then summed over examinees. The estimation proceeds by finding the combination of *A* and *B* that minimizes the differences between test characteristic curves on two forms by the following equation:

$$SLdiff = \sum_{i} SLdiff(\theta_i).$$

An example of the linking design and TCC procedure is shown in Figure 5; this design also allows the use of the MM and MS linking procedures in the context of vertical scaling.

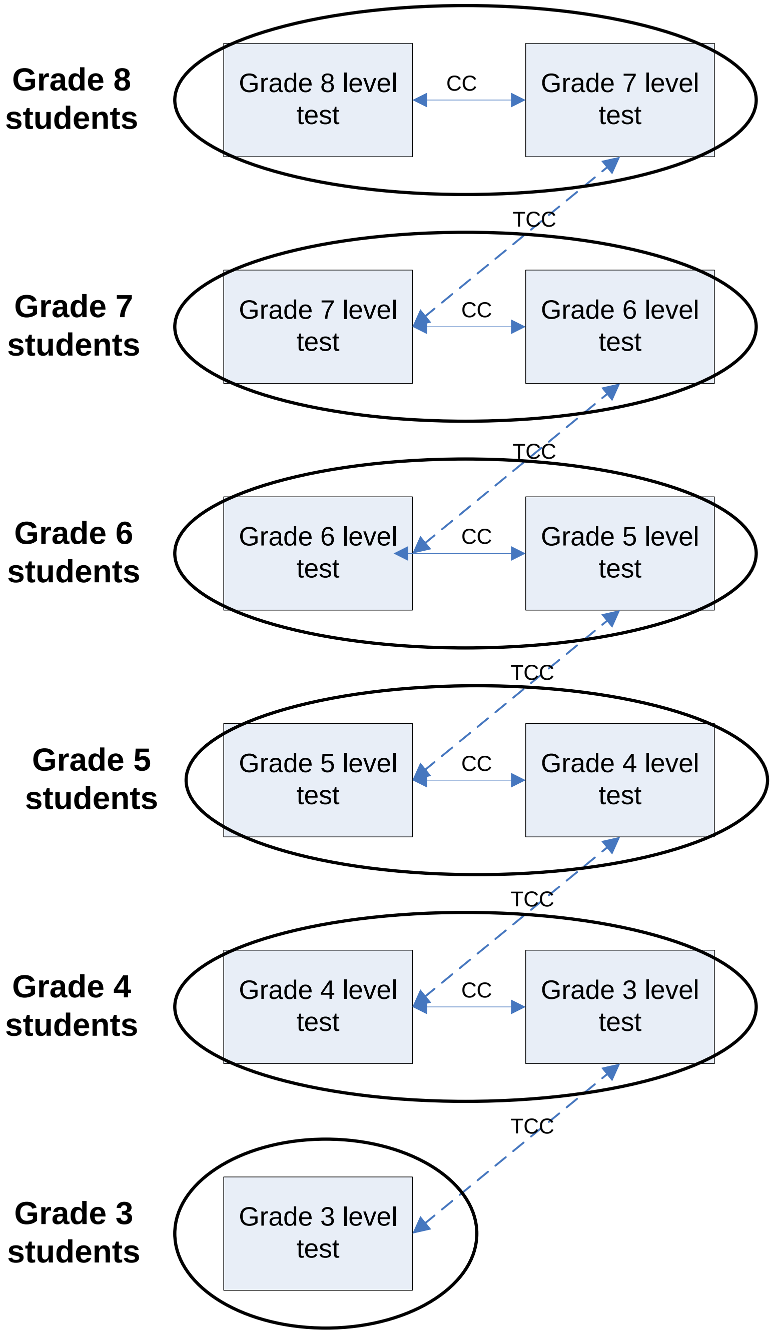

Figure 5. Scaling design for Test Characteristic Curve

As shown in Figure 5, the grade level students are concurrently calibrated for items that are both on-level and off-level for grade level students. The item parameters are placed on the same scale using the common items between grade level students using the TCC method. For example, the item parameters from the concurrent calibration of the grade 6 level student test forms are used as the common-item anchors to link the grade 7 level student items (i.e., grade 6 and 7 test items). That is, the item parameters for the grade 7 student test forms (grade 7 and 6 test items) are transformed using TCC scaling. Subsequently, the transformed grade 7 item parameters for the grade 7 students are used to transform the grade 8 level student items (i.e., grade 7 and 8 items). This process is continued for the grade 5, 4, and 3 student groups. The TCC procedure is a commonly used IRT equating procedure for form-to-form equating.

Hybrid Test Characteristic Curve Method.

The Hybrid test characteristic curve method (HCC) uses both concurrent calibration and an IRT equating procedure, such as TCC procedure, mean/mean, mean/sigma or Haebra. However, using FCIP in this context would be more difficult as several grade level forms would need to be fixed and the concurrent calibration of some grade levels is not possible (i.e., upper or lower grade groupings could not be concurrently calibrated). In this method, groups of continuous grades are concurrently calibrated instead of the entire grade span at a time, and then groups of concurrent calibration results are linked using IRT equating procedures such as TCC. The grade groups can be as small as two grades but less than the full grade span, as shown in Figure 6.

The grade 8 and 7 students that take both the on-level test and off-level exams are concurrently calibrated. In addition, the grade 5 and 6 students that take both on-level and off-level tests are concurrently calibrated and the grade 4 and 3 students are concurrently calibrated. After the three separate concurrent calibrations, the item parameters are linked via the TCC method (or the MM, MS, and Haebra procedures). However, only four studies to date have used the HCC method and all studies used the TCC method (Briggs & Weeks, 2009; Ito, et al., 2008; Karkee, et al., 2003; Meng, 2007).

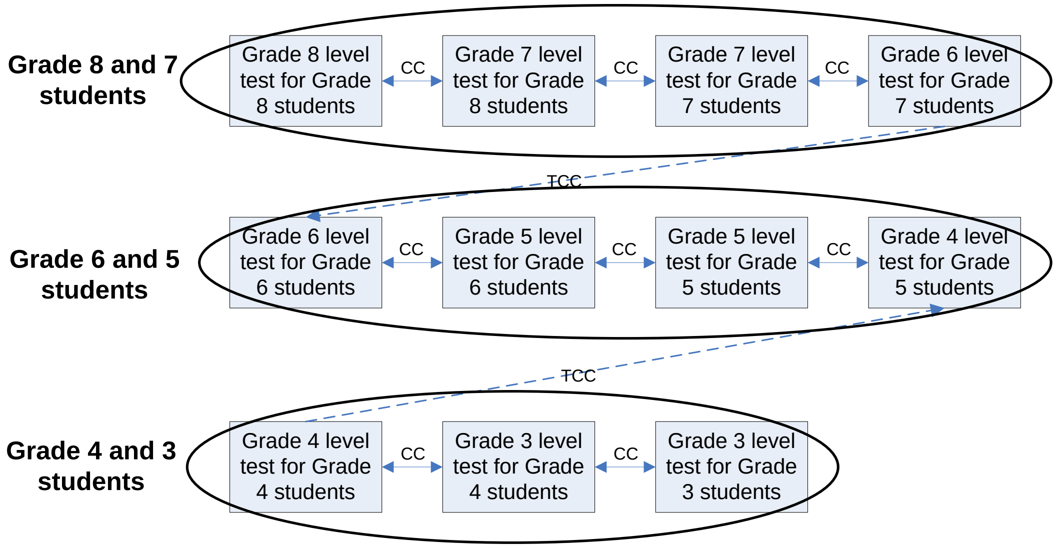
Grade 8 and 7 students
Grade 8 level test for Grade 8 students
CC
Grade 7 level test for Grade 8 students
CC
Grade 7 level test for Grade 7 students
CC
Grade 6 level test for Grade 7 students
TCC
Grade 6 and 5 students
Grade 6 level test for Grade 6 students
CC
Grade 5 level test for Grade 6 students
CC
Grade 5 level test for Grade 5 students
CC
Grade 4 level test for Grade 5 students
TCC
Grade 4 and 3 students
Grade 4 level test for Grade 4 students
CC
Grade 3 level test for Grade 4 students
CC
Grade 3 level test for Grade 3 students

Figure 6. Scaling design for Hybrid Characteristic Curve

Literature Review

The literature review is organized in two sections. Research studies in which the different vertical scaling methods were compared using real data are described and analyzed in the first section. Research studies in which the vertical scaling methods were compared using simulated data are described and analyzed in the second section. In each section, comparisons of traditional vertical scaling methods to IRT vertical scaling methods are briefly summarized followed by studies in which IRT vertical scaling methods were used. The traditional vertical scaling studies are not discussed in detail since these methods are typically not currently used to develop vertical scales. The primary focus of this literature review is to highlight the differences between currently used vertical scaling procedures and to identify what methods were used to evaluate the results. For the simulation studies, the conditions simulated are also highlighted in addition to comparing vertical scaling procedures and identifying the methods of evaluation. For the more recent studies the content areas, grades, sample sizes, and vertical scaling methods are listed. In addition, the evaluation methods to compare the different scaling procedures are listed and summarized. Of particular concern is whether the methods of evaluation and conclusions definitively indicate which procedure is best.

Real Data Studies.

Traditional and IRT Scaling Comparisons.

Although Thurstone Absolute Scaling is not used with many current real data applications, it has been compared to IRT methods in terms of scale shrinkage across grades. Scale shrinkage occurs when the variances of the scale scores decrease as grade levels increase. In the first of three studies, Becker and Forsyth (1992) compared Thurstone Absolute Scaling to two IRT vertical scaling methods, 1-PL FCIP and 3-PL FCIP. They noted that although the Thurstone and IRT methods yielded similar results in terms of average growth grade to grade, differences were found in the tails of the distribution (at and above the 90th percentile and at or below the 10th percentile). In addition, given the average variability was consistent across grades scale, shrinkage did not occur with this data. In the second study, Williams et al. (1998) compared three different versions of Thurstone Absolute Scaling to a 3-PL CC with two different computer programs, BIMAIN and MULTILOG. The authors reported that the scaling methods produced similar trends in mean growth, with somewhat different trends in terms of standard deviations, where two of the Thurstone methods showed increased variability across grades, while the Thurstone (1925) method and the IRT scales showed variances that did not consistently increase or decrease. In the third study, Camilli et al. (1993) evaluated the use of the 3-PL model (CC) as an alternative to equipercentile equating in vertical scaling and primarily focused on scale shrinkage. They found that the IRT method did not produce equal interval scales, and suggested that the "criteria for determining useful vertical scales constitute a controversial topic for debate and research" (p. 387).

Several researchers compared vertical scaling using linear or equipercentile methods to IRT methods (Cook & Douglass, 1982; Harris, 1991; Kolen, 1981; Petersen, et al., 1983; Phillips, 1986). These studies focused on meeting the requirements of IRT adequately (i.e., misfitting persons, misfitting items, guessing and in the case of the one parameter model equal discrimination). Harris (1991) examined different data collection designs in addition to comparing equipercentile and IRT true score equating. These studies indicated that data fit issues can impact the choice of the IRT model to be employed, and that the 3-PL model performed the best among the three IRT estimation models in the vertical scaling context.

IRT Scaling Comparisons.

Several early researchers evaluated IRT scaling methods to create vertical scales, where the adequacy of the model was the primary focus and where the 1-PL logistic model CC was commonly used (Guskey, 1981; Holmes, 1982; Rentz & Bashaw, 1977; Shen, 1993; Slinde & Linn, 1978, 1979; Whitely & Dawis, 1974). The results generally indicated that the 1-PL model was inadequate for use in vertical scaling due to model-data fit issues and assumption violations. In a later study to evaluate the use of the 3-PL model, Harris and Hooker (1987) evaluated vertical scaling for a Mathematics computation test from the *Iowa Tests of Basic Skills* for grades 4 through 8 with approximately 200 examinees per test form (though two forms only had 79 and 89 examinees) using the LOGIST 5 computer program. In their study, the model's properties, such as item-free measurement and person-free calibration, were evaluated and found to not hold. In summary, the use of IRT equating methods in vertical scaling initially produced inferior results due to not meeting assumptions such as unidimensionality and/or poor model-data-fit. The assumption of normal distributions was met in all of these studies. However, as pointed out above, the condition of normality may not hold with the need to meet NCLB proficiency requirements.

Several IRT vertical scaling methods have been compared in more recent research. Jodoin et al. (2003) compared CC, FCIP, and MS , each conducted using MLE and EAP, to assess academic growth year-to-year for a single grade level in Mathematics (grade level not specified) from a high-stakes state-wide testing program, with a sample of about 60,000 examinees. This study focused on year-to-year growth across three or more years and did not evaluate the creation of a vertical scale across grades. However, the scaling procedures they used are also used in the context of vertical scaling. Jodoin et al. (2003) used two different types of comparisons: (1) the average, standard deviation, minimum and maximum theta values obtained from the CC, FCIP and MS procedures, and (2) the theta values yielded by the MLE and EAP estimation procedures.

For both estimation procedures the largest growth was found in the 2000 year using the CC method (i.e., 0.09, 0.07, and 0.16 for EAP for the 1998, 1999, and 2000 years respectively; and 0.12, 0.09, and 0.21 for MLE), followed by the FCIP procedure which estimated a moderate amount of growth (i.e., 0.06, 0.08, 0.13 for EAP, and 0.08, 0.10, and 0.18 for MLE), and the smallest amount of growth was found using the MS procedure (i.e., 0.02, 0.03, and 0.05 for EAP, and 0.04, 0.06, and 0.09 for MLE). Further, the mean growth was found to be consistently higher with the MLE estimation method than the EAP estimation method. According to Jodoin et al. (2003) “the results suggest that the choice of equating methodology and ability estimator have important consequences in measuring academic growth” (p. 241). In addition, this study was the only real data study that compared the multiple scaling methods in terms of the agreement of the classification of examinees into proficiency categories. Each examinee was placed into one of the four proficiency categories (i.e., inadequate, adequate, proficient, and advanced). Whereas, all three scaling methods had a high level of consistent classifications, the MS and FCIP methods produced a slightly higher degree of agreement for proficiency categories than the CC procedure, and the FCIP method placed examinees in the next higher proficiency category more often than the MS and CC procedures for MLE, while inconsistent placement occurred for all scaling procedures for EAP. However, the percentage agreement between the two estimation methods (i.e., MLE vs. EAP) was not provided.

Karkee et al. (2003) studied the CC, TCC, and HCC methods used to vertically scale student performance on a series of Mathematics test forms that spanned six grade levels (grades 5 through 10). The sample size was approximately 10,000 examinees per grade level. The HCC procedure was conducted by using the CC procedure to link pairs of adjacent grade level items (grades 5 and 6, grades 7 and 8, and grades 9 and 10). Then the results of the grade 5 and 6 concurrent calibration were linked to the CC results of grade 7 and 8 using the TCC method. The grade 9 and 10 CC results were linked to the results of the grade 7 and 8 results in a similar fashion. Four different types of comparisons were conducted to evaluate the results: (1) number and average of item model-data fit indices, convergent items, and differential item functioning (DIF; a statistical flag for item bias where the probability of answering an item correctly that is compared for two groups), (2) means and standard deviations of the scale scores, (3) residuals between the observed and predicted scores for items and tests, and (4) mean and standardized difference between means (SDMs) of the item parameters.

Karkee et al. (2003) indicated that the TCC method produced consistently better results than the CC or HCC methods. The MLE estimation procedure failed to converge for two items for the CC method and for one item each for the HCC and TCC methods. For model-fit indices, 20 items were flagged as misfitting for the CC method, 18 items for the HCC procedure, and 10 for the TCC procedure. For DIF indices, the CC procedure identified 30 items, HCC identified 27 items, and the TCC procedure identified 13 items. The CC procedure had the smallest range of item residuals, which ranged from 2 to 24, for the HCC procedure from 0 to 44, and for the TCC procedure from 2 to 30. The test score residuals were generally consistent across procedures. The Pearson correlations between the item parameters across procedures were all high ($\rho >$ 0.90) for the *a*- and *b*- item parameters, and were lower for the *c*- parameters which ranged from 0.69 to 0.88. In summary, the TCC procedure seemed to outperform the HCC and CC procedures in this study.

Kim (2007) examined the CC, FCIP, and TCC methods when scaling grades 3 through 8 for Vocabulary, Reading, Mathematics, and interestingly Science assessments. The data used were based on the *Iowa Test of Basic Skills* (ITBS) which is not referenced to any particular curriculum and where the focus is on national outcomes and comparisons. The number of examinees per grade ranged from approximately 500 to 1,800 examinees. Four different comparisons were made: (1) grade-to-grade growth, (2) grade-to-grade variability, (3) effect size of grade separation, and (4) horizontal distances. In addition, Kim examined five different estimation procedures: MLE-pattern scores, MLE-number scores, EAP-pattern scores, EAP-number scores, and posterior distribution of “true” underlying proficiency (QD). Kim used the BILOG-MG computer program for the analyses. The results of the grade-to-grade variability and horizontal distances will not be presented here as these results are not relevant to the current study.

The mean theta differences between consecutive grades across the five estimation procedures for Vocabulary ranged from 0.19 to 0.77 for CC, from 0.19 to 0.86 for FCIP, and from 0.19 to 0.78 for TCC. The mean theta differences between consecutive grades across the five estimation procedures for Reading ranged from 0.53 to 1.06 for CC, from 0.44 to 1.61 for FCIP, and from 0.47 to 1.66 for TCC. The mean theta differences between consecutive grades across the five estimation procedures for Mathematics ranged from 0.24 to 0.87 for CC, from 0.22 to 1.01 for FCIP, and from 0.25 to 0.94 for TCC. The mean theta differences between consecutive grades across the five estimation procedures for Science ranged from 0.40 to 0.83 for CC, from 0.38 to 1.62 for FCIP, and from 0.40 to 1.02 for TCC. Kim (2007) noted that the five estimators provided almost the same results and that the differences between scaling procedures were very small for Vocabulary and Mathematics. However, for Reading and Science the grade level mean difference for the scaling methods were different depending on the proficiency estimator and grade level. CC provided smaller mean differences at the lower grades (grades 3-5) than did FCIP or TCC, and larger mean differences between grades 3 and 4 when MLE (pseudo and pattern) was used.

Kim (2007) used Yen’s (1986) effect size measure to measure the separation of grade distributions. The effect sizes for pairs of consecutive grades across the five estimation procedures for Vocabulary ranged from 0.17 to 0.74 for CC, from 0.18 to 0.73 for FCIP, and 0.19 to 0.75 for TCC. The effect sizes for pairs of consecutive grades across the five estimation procedures for Reading ranged from 0.45 to 0.80 for CC, from 0.47 to 0.81 for FCIP, and 0.52 to 0.81 for TCC. The effect sizes for pairs of consecutive grades across the five estimation procedures for Mathematics ranged from 0.20 to 0.76 for CC, from 0.20 to 0.75 for FCIP, and 0.22 to 0.78 for TCC. The effect sizes for pairs of consecutive grades across the five estimation procedures for Science ranged from 0.27 to 0.63 for CC, from 0.33 to 0.62 for FCIP, and 0.29 to 0.68 for TCC. The effect sizes for the different methods were essentially the same. The effects sizes for Mathematics were the smallest followed by Science, Vocabulary, and Reading. The effect sizes decreased as grade increased for Vocabulary and Mathematics and fluctuated for the Reading and increased for the Science test as grade increased.

The range of theta differences was the smallest for CC and TCC (except for TCC for Reading which was similar to FCIP), and the FCIP had the largest range of theta differences. The range of effects sizes for pairs of consecutive grades was generally similar across the CC, FCIP and TCC procedures. In summary, it is difficult to know which procedure was best since some of the results were similar and the evaluation methods only looked at the differences between grades and not any type of error rate.

Ito et al. (2008) examined the CC and HCC methods when scaling Kindergarten through grade 9 Reading and Mathematics, with approximately 1,700 examinees per grade. Three different comparisons were made: (1) means, standard deviations, and correlations of the item difficulties, and item discriminations and (2) root-mean-square-differences (RMSDs) between the sets of scales scores. Ito et al. (2008) used the 3-PL model with the proprietary software PARDUX (Burket, 1991) and BMIRT programs, which implement unidimensional and compensatory multidimensional multi-group IRT models using Markov chain Monte Carlo (MCMC) methodology.

The scale scores were transformed from the theta metric using a multiplier of 30 and an additive constant of 550 for each content area. The mean difficulties for Reading ranged from 367 to 586 for the CC method and 391 to 586 for the HCC procedure. The mean difficulties for Mathematics ranged from 355 to 606 for the CC method and 400 to 606 for the HCC procedure. The mean item discriminations for Reading ranged from 0.02 to 0.04 for both the CC and HCC procedures. The mean item discriminations for Mathematics ranged from 0.01 to 0.03 for the CC, and 0.02 to 0.04 for the HCC procedure. The discrimination estimates from the CC method for extreme grades (upper or lower grades) were on average higher than those from the HCC method.

Strong linear relationships between ability estimates for the two procedures were observed ($\rho > 0.99$). The RMSDs of the scale scores for Reading ranged from 0.81 to 8.12 with grades 1, 2, and 9 having the largest differences. The RMSDs of the scale scores for Mathematics ranged from 1.19 to 21.58 with grades 1, 2, 8 and 9 having the largest differences.   Ito et al. (2008) noted the two procedures provided similar results for many grades in terms of correlations and RMSDs. However, there were larger RMSDs for Mathematics and smaller RMSDs for Reading. According to Ito et al. (2008), the results indicate that it matters more for Mathematics than for Reading which scaling method is used and suggested that the dimensionality of the construct could be a contributing factor. However, Ito et al. (2008) noted that determining dimensionality with multiple grades is complex and not easily defined. These results indicate that there may be scaling differences attributable to content area; however, until this is systematically evaluated using simulation methods this is only a hypothesis.

The most recent study was conducted by Briggs and Weeks (2009). Briggs and Weeks (2009) compared three factors: (1) 1-PL and 3-PL IRT models, (2) TCC and a variant of the HCC method, and (3) EAP and MLE estimation procedures. They used Reading data in grades 3 through 7, with an average of 55,681 students for each grade. The data for grades 3 and 4 were collected in 2003, for grades 4 and 5 in 2004, for grades 5 and 6 in 2005, and for grades 6 and 7 in 2006. In this study, the HCC method was conducted in which the CC procedure was used within the same grades across years (e.g., grade 4 in 2003 and 2004), and the TCC procedure was used across different grades within a test year (e.g., grade 3 and 4 in 2003). Two outcomes were compared for the eight comparisons (1) mean and SDs of the logit scale scores across grades and (2) an effect size of the growth.

The mean logist scale scores ranged from approximately 0.02 to 0.50 for the 1-PL model and 0.02 to 1.00 for the 3-PL model (values are estimated from figures since specific values were not provided). The effect sizes of growth for the 1-PL models ranged from approximately 0.45 for the grade 3-4 group to 0.30

for the grade 5-6 group. The effect sizes of growth for the 3-PL models ranged from approximately 0.50 for the grade 3-4 group to 0.25 for the grade 5-6 group. Depending on the IRT model, linking method, and estimation approach used, the vertical scale was either stretched or compressed. Briggs and Weeks indicated that no single method adequately measured growth.

Taken together, the studies in which real data were used revealed that differences in the outcome measures were not consistent across studies. Real data studies in which the 1-PL model with the CC procedure was used produced inferior model-fit in vertical scaling. The 3-PL model using the CC, FCIP, and TCC procedures produced mixed results with some vertical scaling methods producing better results for some studies, but no particular procedure providing consistently better results across all the studies reviewed. The primary evaluation method used in these studies was to examine model-data-fit in terms of convergence of items and model-data fit indices. These studies also used different content areas to evaluate the vertical scaling methods (i.e., Vocabulary, Reading, Mathematics, and Science) and varying sample sizes (1,700 to 60,000). Only three studies evaluated content areas comparatively and they indicated differences in results between the different vertical scaling methods considered (Ito, et al., 2008; J. Kim, 2007; Tong & Kolen, 2007). The first study compared the size of the RMSDs (Ito, et al., 2008) and the second study compared the grade-to-grade growth and effect sizes (J. Kim, 2007). The third study is summarized in the next section compared the effect size and horizontal distances of score distributions (Tong, 2005; Tong & Kolen, 2007). Lastly, different evaluative procedures were used to assess and compare the results obtained by the different vertical scaling procedures, but the evaluative procedures did not indicate which procedure produced the best results. However, since all of these studies are based on real data, comparisons across methods might be better analyzed using simulated data with known parameters.

Simulation Studies.

Traditional and IRT Scaling Comparisons.

Skaggs and Lissitz (1988) simulated vertical equating of two tests for examinee samples of low, medium, and high ability, and ability matched to the difficulty level of an unspecified test (unknown if simulated or taken from an existing test). The 1-PL MM, 3-PL CC, and equipercentile methods were compared. The sample size for each test form was 2,000. Four levels of examinee ability were considered: -0.5 mean logits for both samples, 0.0 mean logits for both samples, 0.5 mean logits for both samples, and -0.5 mean logits for one sample and 0.5 mean logits for the second sample. The focus of the study was to evaluate the effect of examinee ability on the equating results. Two statistics were used to evaluate the results: (1) unweighted mean square error and (2) weighted mean square error. The results indicated that all three methods were generally invariant with respect to examinee ability level for all conditions. They found that the 3-PL model seemed to work better than the 1-PL model. The authors indicated that:

> Multidimensionality might account for a lack of test equating invariance has profound implications for vertical equating. It may not be meaningful

> to vertically equate certain kinds of tests. The Harris and Hoover, Harris and Kolen, and Loyd and Hoover studies all used mathematics tests, the content of which is known to vary considerably across grade levels. Reading and vocabulary test, on the other hand, might be more unidimensional across grades and may provide more invariant equating results. (Skaggs & Lissitz, 1988, p. 80).

Similar to the real data studies, early research in evaluating vertical scaling was to ensure model-data fit, an important first step, in the context of simulated conditions.

One identified problem in earlier research in creating vertical scales was scale shrinkage. While the current focus of the present study was not on scale shrinkage, the next two studies provide additional evidence that agreement is problematic on which vertical scaling method is most appropriate. Clemans (1993) simulated data based on body weights in pounds for each of 12 age groups from age 6 to 17. The two scaling procedures used were Thurstone Absolute Scaling and TCC (Clemans did not indicate which IRT model was used). Clemans claimed that the TCC method produced scale shrinkage. He also indicated that the Thurstone procedure did not result in a systematic decrease in variance, and the results for the Thurstone method were more precise.

In the second study, Yen and Burket (1997) examined the Thurstone Absolute Scale and the 3-PL TCC methods in a response to Clemans' (1993) article. One criticism of the Clemans article was that the simulated data were favourable to the Thurstone Absolute Scaling method. Yen and Burket (1997) simulated data that was consistent with both the Thurstone and 3-PL models. Their results indicated that realistic simulation conditions and analysis procedures did not produce scale shrinkage for either procedure.

Tong and Kolen (2007; see Tong, 2005 for full details) evaluated Thurstone Absolute Scaling method and several IRT scaling methods using both real and simulated data. Four content areas were considered in the real data comparison: Vocabulary, Mathematics, Language and Reading Comprehension for grades 3 through 8 with approximately 600 to 1,800 examinees per grade. The simulated data were only based on item parameters from the Vocabulary tests. The factors modeled in the simulation included: (1) scaling method (Thurstone, 3-PL scaling test design, 3-PL common item design), (2) SDs by grade (increasing, decreasing, and constant), and (3) sample size (500, 2,000, and 8,000). For the 3-PL IRT scaling model two designs were considered: a scaling design in which a scaling test is constructed to cover the content area across all levels being placed on the same developmental scale (Kolen & Brennan, 2004; Tong, 2005; Tong & Kolen, 2007) and a common anchor item design. Five scores were produced for each design: MLE-pattern scores, Quadrature Distribution (QD)-pattern scores, EAP-pattern scores, MAP-pattern scores, and summed scores based on EAP. The evaluation criteria included: (1) means and standard deviations of the scale score distributions, (2) effect size of the scale score distributions, and (3) horizontal distance of the scale score distributions. Horizontal distance is defined as a percentile difference on the score scale for the same percentage between two distributions (Holland, 2002; as cited in Tong & Kolen, 2005, p.237). The means and SDs and horizontal distances of the scale score distributions are not presented here as they are similar to the effect size measure.

The effect sizes for the real data are presented in this paragraph and the simulated condition results are presented in the next paragraph. This study is the third study that compared content areas. The effect sizes for the real data Vocabulary ranged from 0.41 to 0.80 for the Thurstone method, from 0.20 to 0.62 for scaling test design, and from 0.26 to 0.69 for common-item design. The effect sizes for the real data Mathematics ranged from 0.11 to 0.71 for the Thurstone method, from 0.09 to 0.67 for scaling test design, and from 0.31 to 0.69 for common-item design. The effect sizes for the real data Language ranged from 0.16 to 0.48 for the Thurstone method, from 0.13 to 0.48 for scaling test design, and from 0.28 to 0.76 for common-item design. The effect sizes for the real data Reading ranged from 0.26 to 0.45 for the Thurstone method, from 0.20 to 0.42 for scaling test design, and from 0.61 to 0.76 for common-item design. The effect size differences were on average the highest for Reading for all three procedures. The second highest effect size differences were found with Language at grades 3/4, Vocabulary for grades 4/5 and grades 6/7, and Mathematics for grades 7/8 (grades 5/6 were tied between Vocabulary and Language). The lowest effect size differences were found in Mathematics for all grade groups except between grade 7 and 8. Similar to the other two studies that compared content areas, this study found differences between content areas for some of the outcome measures.

This paragraph summarizes the simulated portion of the study. The effect sizes for the increasing SD conditions ranged from 0.09 to 0.52 for Thurstone, from 0.11 to 0.64 for scaling test design, and from 0.11 to 0.63 for the common-item design. The effect sizes for the decreasing SD conditions ranged from 0.12 to 0.62 for Thurstone, from 0.18 to 0.76 for scaling test design, and from 0.18 to 0.80 for the common-item design. The effect sizes for the constant SD conditions ranged from 0.11 to 0.55 for Thurstone, from 0.02 to 0.70 for scaling test design, and from 0.10 to 0.70 for common-item design. According to Tong and Kolen (2007) the effect sizes "often yielded higher estimates for the common-item design than the scaling test design, and the general trend was the higher the grade level, the larger the effect size difference between the two designs" (p. 240). But Tong and Kolen recognized that other models (i.e., random-equivalent group design, multidimensional IRT) should be examined to determine which IRT model is most appropriate. They indicated that if the items on the tests across grades were reasonably unidimensional, then IRT methods might be preferred due to the assumption of normality underlying the use of Thurstone Scaling. In summary, Tong and Kolen (2007) indicated that "clearly, the results of this study show that the choice of design can have important practical effects on the results" (p. 248).

In summary, the IRT vertical scaling methods appear to produce better results than Thurstone Absolute Scaling as long as the data fit and met the assumptions of the model. However, the studies summarized suggest that using the 3-PL model either concurrently or with a linking method like MS and TCC is superior to the Thurstone scaling method or the 1-PL model, but they did not suggest which of the 3-PL methods was superior.

IRT Scaling Comparisons.

Similar to the early real data studies, one of the earliest simulation studies examined model-data-fit indices to evaluate the appropriateness of IRT vertical scaling procedures. Gustafsson (1979) simulated data (1,300 examinees with normal distribution) to fit the 1-PL CC model with two levels of difficulty corresponding to an "easy" and "hard" test. He used model fit indices to evaluate these procedures for both horizontal and vertical scaling. He indicated that the lack of applicability of the 1-PL model for use in vertical scaling shown in a previous study (Slinde & Linn, 1978) was due model data misfit.

Baker and Al-Karni (1991) compared the MM and TCC procedures using the 3-PL model. Their simulation was based on three different ability levels (low, medium and high) within a normal distribution. Three different comparisons were conducted: (1) average equating coefficients, (2) loss-function values, and (3) root-mean-squared-difference values (RMSD) for the *a*-, *b*-, and theta values between the two scaling procedures. The two linking methods produced similar linking coefficients for all three ability levels. The loss function values ranged from 0.02 to 6.15 for the MM procedure and from 0.00 to 1.28 for the TCC procedure. The RMSD values varied from 0.03 to 0.38 for the theta estimates, from 0.01 to 0.19 for the *a*- parameters, and from 0.02 to 0.23 for the *b*-parameters. Baker and Al-Karni concluded that the TCC method was superior to the MM method and was less sensitive to atypical test characteristics.

Kim and Cohen (1998) compared the CC and TCC procedures and the BILOG and MULTILOG computer programs in their simulation study. Two grade levels were simulated using a 2-PL model with a sample of 500 normally distributed scores for each grade. In addition, four different numbers of common items were simulated (5, 10, 25, and 50). Two comparisons were conducted: (1) root-mean-squared-differences (RMSD) for the *a*- and *b*- parameters and (2) mean Euclidean differences based on both parameters. The RMSDs for the item discriminations ranged from 0.09 to 0.21 for the CC and from 0.10 to 0.15 for the TCC method. The RMSDs for the item difficulties ranged from 0.07 to 0.25 for the CC and 0.07 to 0.11 for the TCC method. The mean Euclidean differences ranged from 0.10 to 0.27 for the CC, and from 0.10 to 0.15 for the TCC method. The TCC method yielded smaller root-mean-square-differences for the item discrimination and difficulty parameters with the two smaller number of common item conditions (i.e., 5 and 10) than the CC method, particularly for the smaller sample sizes. For the conditions with the largest number of common items (i.e., 50) the two methods produced similar results. The RMSD differences for both item discrimination and difficulty were similar using either BILOG or MULTILOG for TCC, but the RMSD values were smaller for item discriminations and higher for item difficulties for the CC and BILOG compared to CC and MULTILOG. The mean Euclidean differences indicated that the CC and MULTILOG values were smaller than CC and BILOG, but for TCC either computer program produced similar results.

Hanson and Béguin (2002) evaluated five scaling procedures used to establish a vertical scale across two grade levels - CC, MM, MS, TCC, and Haebara (Haebara, 1980). The Haebara procedure is an alternative item characteristic curve equating procedure. Five additional factors were examined: (1) computer program (MULTILOG and BILOG-MG), (2) sample size (1,000 and 3,000) based on a normal distribution, (3) number of common items (10 and 20), (4) equivalent and non-equivalent groups with two forms (no mean difference and one SD difference), and (5) score type (weighted true score and weighted ICC score). The equivalent group conditions were simulated to mimic horizontal equating and the non-equivalent group conditions were simulated to mimic vertical scaling. Two comparisons were conducted to evaluate the results: (1) squared bias for the weighted true score equating and weighted ICC criterion, and (2) mean square error for the weighted true score equating and weighted ICC criterion.

The summary of Hanson and Béguin’s (2002) study results only include the non-equivalent group conditions because this portion of the study simulated a creation of a vertical scale and the equivalent group conditions are similar to a general equating within a grade level with alternate forms. Graphical representations of the results were provided by Hanson and Béguin (2002). For the weighted true scores using BILOG-MG the smallest amount of squared error occurred for the CC procedure, followed in order of increasing error by the Haebara, TCC, MS, and MM procedures. The pattern was not as consistent for the MULTILOG procedure where squared error values for the TCC procedures were the smallest for the smaller number of common item conditions and the square error for the MS procedure was the smallest for the larger number of common items. The average bias values for all conditions were overall smaller for MULTILOG than for BILOG-MG.

In contrast, a consistent pattern of results was found for the weighted ICC criterion. The MS and MM had the highest squared bias and mean square error across conditions and the CC condition had the smallest error with the exception of the 3,000 sample size conditions for squared bias. In addition, the values were higher for BILOG-MG conditions than for MULTILOG. Hanson and Béguin (2002) indicated that the CC estimation generally resulted in lower error than the separate estimation methods for the smaller sample sizes, but higher error for the 3,000 sample size conditions.

Keller et al. (2004) evaluated the CC, FCIP, MS, and TCC scaling procedures to evaluate student year-to-year growth. Three additional factors were manipulated: (1) sample size (250 and 5,000 based on a normal distribution), (2) number of common items (5 and 9), and (3) seven levels of growth (0.00, 0.10, 0.25, 0.50, -0.10, -0.25, and -0.50). The mean growth values recovered were used to evaluate the different scaling procedures between the estimated and true theta values. The differences between estimated and true growth for the smaller number of anchor items across sample size conditions ranged from 0.04 to 0.14 for the CC procedure, from 0.00 to 0.30 for the FCIP procedure, from 0.00 to 0.05 for the MS procedure, and from 0.01 to 0.24 for the TCC procedure. The differences between the estimated and true growth for the larger number of anchor items across sample size conditions ranged from 0.01 to 0.10 for the CC procedure, from 0.00 to 0.26 for the FCIP procedure, from 0.00 to 0.09 for the

MS procedure, and from 0.03 to 0.06 for the TCC procedure. Keller et al. (2004) found that there was a lack of consistency across the scaling procedures, but that the MS method performed the best and the FCIP method performed the worst, consistently underestimating the amount of growth.

Pomplun et al. (2004) compared the BILOG-MG and WINSTEPS computer programs with the 1-PL CC vertical scaling procedure. Real and simulated data based on a Mathematics test were scaled across 5 grade levels (grades 2 through 6) with a sample of 2,500 normally distributed examinees per grade. Three comparisons were used to evaluate the results: (1) mean differences between true versus estimates theta values, (2) correlations between true versus estimated values for both item difficulties, and (3) root-mean-squared-error values for the true versus estimated values for the item difficulties and thetas. The correlations between the true and estimated item difficulties were perfect at 1.00 and the mean differences ranged from -0.18 to 0.18 for all grades and for both computer programs. The true versus estimated theta estimates were also perfectly correlated for all grades and the mean differences ranged from -0.17 to 0.21 for both computer program estimates. The RMSE values for the item difficulties ranged from 0.05 to 0.10 for WINSTEPS, and 0.06 to 0.24 for BILOG-MG. The RMSE values for the theta estimates ranged from 0.49 to 0.55 for WINSTEPS, and 0.49 to 0.57 for BILOG-MG. The authors indicated that while the WINSTEPS program captured the individual and mean estimates more accurately, BILOG-MG captured the standard deviations more accurately. They also suggested that the choice of software seemed to influence the outcome for the 1-PL CC estimation.

Custer et al. (2006) examined the 1-PL CC vertical scaling procedure using the same two computer programs but with the real and simulated data based on a vocabulary test across 11 grades (K-10, simulated 7,500 examinees per grade). Three factors were manipulated: (1) distribution (normal and skewed distributions-positively skewed for K-1, and negatively skewed for grades 2 through 10), (2) computer program (WINSTEPS and BILOG-MG), and (3) convergence settings (default settings 0.01 for BILOG-MG and WINSTEPS, and tighter settings of 0.003, 0.001, and 0.0005 of the threshold value of the logit change). The following comparisons were made: (1) mean and standard deviation for item and theta estimates, and (2) effect size for item and theta estimates. The effect sizes were calculated by obtaining the difference at each grade between the estimated and simulated mean and dividing this difference by the simulated standard deviation (Custer, et al., 2006).

The effect size values for normal distribution conditions ranged from 0.00 to 0.86 for WINSTEPS, and 0.00 to 0.49 for BILOG-MG. Likewise, the effect size values for the skewed distribution conditions ranged from 0.01 to 0.86 for WINSTEPS, and 0.01 to 0.49 for BILOG-MG. The results indicated that BILOG-MG captured the individual and mean estimates more accurately, but with tighter convergence settings both programs provided similar results. There were not large differences between the normal distribution conditions and the skewed distribution conditions. Although the simulated skewed distributions were based on the real skewed distributions in the data, the skewed distributions were not markedly different from the normal distributions.

Chin et al. (2006) simulated four factors and evaluated two IRT scaling methods (MS and CC) with the 3-PL model. The four factors were: (1) amount of grade-to-grade growth (0.5 and 1.0), (2) number of grade levels (3, 4, and 5), (3) number of common items (12, 18, 24), and (4) difficulty range of linking items (0.0 SD, 1.0 SD, and 2.0 SD). The data were simulated with a sample of 10,000 students per grade level with a normal distribution. Three criteria were used to evaluate the results: (1) the number of estimation cycles for convergence, (2) mean and standard deviation of the thetas, and (3) root-mean-squared-error (RMSE) for the item parameters and ability estimates. The total number of estimation cycles ranged from 32.20 to 102.60 for the CC method and from 6.50 to 9.20 for the MS method. The RMSE values for the item discrimination values ranged from 0.07 to 0.18 for the CC method and from 0.08 to 0.60 for the MS method. The RMSE values for the item difficulty values ranged from 0.16 to 0.41 for the CC method and from 0.18 to 0.42 for the MS method. The RMSE values for the theta values ranged from 0.35 to 0.51 for the CC method and from 0.36 to 0.60 for the MS method. The results indicated that MS is vulnerable to restriction of the common item difficulty range, whereas CC was generally less affected by number of common items or the range of difficulty of the items. However, CC estimation was more likely than the MS procedure to have items converge when the number of forms to be linked was larger and grade-to-grade growth was larger.

Meng (2007) evaluated four different vertical scaling procedures using simulated data derived from a Reading assessment in grades 3 through 8. The four scaling procedures were CC, TCC, and two versions of the HCC procedure (pairwise HCC - two adjacent grade pairs with CC, and semi-concurrent HCC - three adjacent grades with CC). The 3-PL model was used for the dichotomous items and the generalized partial credit model (Muraki, 1992) was used for the polytomous items. Four additional factors were varied: (1) sample size (500, 1,000, and 5,000), (2) number of common items (10 and 20), (3) type of common item (dichotomous, dichotomous and polytomous), and (4) number of constructed response items (6 and 12). All of the data were simulated based on a normal distribution. The computer program used to simulate and calibrate the data was the IRT Command Language program using 500 cycles and convergence of 0.001 with MLE estimation. There were four classification proficiency categories labelled Level 1 through 4. The results were evaluated based on the following outcome measures comparing true parameters to estimated parameters: (1) proficiency score mean comparisons using absolute bias, RMSE, and SE, (2) proficiency score SD comparisons using absolute bias, RMSE, and SE, (3) effect size between adjacent scores for grades using absolute bias, RMSE, and SE, and (4) proficiency score classification proportions comparisons using absolute bias, RMSE and SE. The results for the proficiency classification proportions are the focus of this review since it is the most relevant information to this proposal. The results for the means, SDs, and effect size can be found in the full paper (Meng, 2007).

Meng summed the results for each outcome measure (i.e., absolute bias, SE, and RMSE) across grade levels and then averaged across conditions to compare the vertical scaling procedures (see Tables 4.4 through 4.7 in Meng (2007) for full results by condition). The results were summed across grade levels and averaged across conditions were summarized here. There were three cut-scores used for this study that resulted in four levels labelled Level 1 through Level 4 for each grade. For Level 1, the CC procedure had the lowest value and the TCC procedures had the largest value for absolute bias, SE and RMSE. For Level 2, the pairwise HCC had the lowest value and the CC procedure had the largest value for absolute bias, SE and RMSE. For Level 3, the semi-concurrent HCC procedure had the lowest value for absolute bias and RMSE, but the CC procedure had the lowest value for SE. The TCC procedure had the highest value for all three outcome measures. For Level 4, the pairwise HCC procedure had the lowest value for absolute bias and RMSE, but the semi-concurrent HCC procedure had the lowest value for SE. The CC procedure had the highest value for absolute bias and RMSE, and the TCC procedure had the highest value for SE. The results when examined more closely showed inconsistencies across grade levels (i.e., summarized results versus tables that did not summed across grades and averaged across conditions). The summarized results may be consistent for one level (i.e., Level 1) by condition and grade, but the raw results showed inconsistencies, where some conditions show higher error by scaling method with no clear picture as to which procedure is best. For example, for Test 1 results in Table B28 (Meng, 2007, p. 245) show the largest error for grades 3, 4 and 8. It is difficult to interpret the resulting summed values when the values across grades are different and when these results are averaged across conditions.

This section summarized various studies that examined different IRT vertical scaling procedures, such as CC, FCIP, MM, MS, Haebara, HCC and TCC scaling methods. Some of these studies simulated data based on real data and content areas such as Reading, Vocabulary and Mathematics, while other studies simulated data from simulated item parameters. Sample sizes ranged from 500 to 10,000 examinees with normal distributions represented, with the exception of the Custer etal. (2006) study that also included simulated skewed data based on the population distributions. Unfortunately, there was no consensus on which vertical scaling procedure produced the best outcome. While these simulation studies have been valuable in identifying some of the procedures to be used to evaluate results obtained in a simulation, the inconsistency in results suggests that it is unclear which procedures are best to use when creating a vertical scale.

Shortcomings of Reported Research

There are three aspects of the previous research that have not been systematically evaluated in one study. The first aspect is the properties of the data from which the real data are simulated. For example, several studies were based on data for either the Reading or Mathematics content areas. Only three studies examined more than one content area in a comparative manner with real data, and the results were mixed (Ito, et al., 2008; J. Kim, 2007; Tong & Kolen, 2007). The effect of content area on creating vertical scales has not been systematically explored. The research to date suggests that one method should be suitable for many different content areas (e.g., Mathematics, Vocabulary, and Reading). But is this a good assumption? Second, with the exception of the Custer et al.'s (2006) study in which a "slightly" skewed distribution of the real data was simulated, all of the simulation studies assumed a normal distribution. Due to current changes and expectations of 100% proficiency under the NCLB legislation the assumption of a normal distribution likely is not realistic. Negatively skewed distributions are likely to be common due to the pursuit of higher standards. What effect this change in the shape of the distribution of scores might have on vertical scaling procedures has not been systematically explored.

A third aspect not systematically evaluated is the use of evaluation methods that are useful to practitioners when creating vertical scales. For example, one study placed examinees in proficiency categories based on cut-scores (Jodoin, et al., 2003). In another study, Meng (2007) evaluated the absolute bias, SE, and RMSE between the true versus estimated proportion classification values. Other studies examined the RMSE or RMSD between pairs of item parameters, ability estimates, or equating coefficients; correlations between true versus estimated values; and convergence or model-fit criteria. While many of these methods are good measures for evaluating different scaling methods, other more practical measures used in conjunction may provide more information. Some new outcome measures in the context of vertical scaling research to evaluate state assessments include decision accuracy and consistency, and conditional standard errors of measurement at cut-scores. Decision consistency refers to the agreement between the classifications based on two nonoverlapping, equally difficult forms of a test (Livingston & Lewis, 1995). Decision accuracy refers to the extent to which actual classifications of test takers agree with those that would be made on the basis of their true scores, if their true scores would be known (Crocker & Algina, 1986; Livingston & Lewis, 1995).

The NCLB (2001) legislation requires administering tests to examinees in grades 3 through 8 in Reading and Mathematics and proficiency cut-scores are used to determine part of the measure of AYP. Another measure of AYP is change in the percentages of examinees achieving “basic proficiency” or higher between two grades and across time (i.e., different cohorts in the same grade). Measures that estimate how accurate and precise the cut-scores are important created for a vertical scale to ensure that the cut-scores and scales are adequate. Decision accuracy, decision consistency, and conditional standard errors outcome measures were not used to evaluate the outcomes in the previous research, but their use could provide more information in determining which scaling measure works best and in what context.

CHAPTER 3 METHODS

The methods used in this research study are described in the present chapter. First, the samples from which the data was simulated are described. Second, the calibration methods used to obtain the item parameters used to simulate the data conditions are provided. Third, the conditions simulated are presented. Fourth, the scaling procedures used are outlined. Fifth, cut-scores and rescaling of the cut-scores are described. Sixth, the evaluation methods for comparing the results are provided.

Research Design

Sample.

The data used in this study was from a large-scale state assessment in the United States[1]. The approximate maximum number of students that were available to be tested for this state assessment is reported in Table 1 for each of the grades in the present study. The full subset of available data was used for this study (approximately 1,500 students per grade per form).

Table 1.

Approximate Student Population per Grade

Grade	Number of students
3	81,400
4	86,000
5	82,700
6	75,800
7	81,400
8	86,000

[1] The state is not identified in compliance with the conditions set for obtaining and using the data.

The real data used included student responses to Reading and Mathematics multiple choice items for grades 3 through 8. The Reading and Mathematics assessments were administered in the same year for all grade levels. Different random samples of students in each grade were administered the Reading and Mathematics assessments. For grade 3, two different forms were administered to separate stratified random samples of students for both Reading and Mathematics. That is, grade 3 had a 3A and 3B form in both Reading and Mathematics that was on-grade. For grade levels 4 through 8, four different forms (two on grade level and two below grade level) were administered to separate stratified random samples of students for both Reading and Mathematics. There were four forms for grade 4 Reading and four forms for grade 4 Mathematics and labelled 4A and 4B for the on-grade level forms, and 3A and 3B for the below or off-level grade forms. Figure 1 presented earlier on page 19, shows the basic data collection design. Since the data collection design utilizes the randomly equivalent groups design, all of the four vertical scaling procedures were appropriate to use for the data available for this study (Kolen & Brennan, 2004). Due to the small number of items on each of the A and B forms (i.e., 25 items or less), the A and B forms for each grade level in each content area were treated as one form (i.e., the 3A and 3B forms were treated as one form). For example, the item parameters for the 20 items from the 3A form for Reading grade 3 and the item parameters for the 20 items from the 3B form for Reading grade 3 item parameters were combined for one form. The number of multiple choice items for the combined test forms is reported in Table 2 for each grade for Reading and Mathematics.

Table 2.

Number of items per form for Reading and Mathematics

	Reading		Mathematics	
Grade	on-level	off-level	on-level	off-level
3	40		42	
4	34	40	49	42
5	34	34	49	49
6	34	34	47	49
7	31	34	49	47
8	34	31	49	49

Procedure.

Calibration and Data Simulation Procedures.

The 3-PL IRT model was used to calibrate the data because this model is the calibration model currently used in most large-scale assessments and in previous research (Briggs & Weeks, 2009; Chin, et al., 2006; Cook & Douglass, 1982; Hanson & Béguin, 2002; Harris & Hooker, 1987; Ito, et al., 2008; Jodoin, et al., 2003; Karkee, et al., 2003; Keller, et al., 2004; J. Kim, 2007; S. H. Kim & Cohen, 1998; Meng, 2007; Tong & Kolen, 2007; Yen & Burket, 1997). The 3-PL item parameters were estimated for the real data using BILOG-MG (Zimowski, et al., 1996) for each of the boxes shown in Figure 1. For example, at grade 4, the combined grade 3 test forms (3A and 3B) and the combined grade 4 level test forms (4A and 4B) were calibrated separately to estimate item parameters for the four test forms. The same procedure was used for the remaining grades. Thus, the total number of calibrations for Reading and Mathematics was 22. A sample BILOG-MG syntax file for this calibration is provided in Figure A1 in Appendix A. The convergence criteria were increased from the default of 10 cycles to 500 cycles, the number of quadrature points was increased from 20 to 40, and the NOFLOAT and TPRIOR options were used for all calibrations. This was to prevent non-convergence of item parameters where possible. The number of quadrature points corresponds to points along the theta scale. That is, the theta scale is divided by the number of quadrature points on the theta scale and corresponds to each of the points along the scale. For example, if you had a theta scale from -4 to 4 and there were 9 quadrature points they would refer to the following theta points: -4, -3, -2, -1, 0, 1, 2, 3, 4. By increasing the number of quadrature points from 20 to 40 the precision of measurement of the item parameters will be increased. Also, the estimation process converges more easily.

The mean and standard deviation item parameters by grade and form for Reading and Mathematics are provided in Table 3. The *a*- and *c*- parameter estimates were relatively similar between Reading and Mathematics and across the grade levels. For Reading, with one exception (grades 4 and 5), the *b*-parameter estimates were in ascending order. In contrast, the order of the *b*-parameter estimates across the six grades was not consistent. Similar patterns for the on-level and off-level test forms were found for both Reading and Mathematics. Further, the mean *b*-parameters estimates across the grade levels were further apart for Mathematics as compared to Reading.

Table 3.

Reading and Mathematics Mean item parameters for the on-level and off-level test forms by grade

			Reading			
	a		*b*		*c*	
			On-level			
Grade	Mean	SD	Mean	SD	Mean	SD
3	0.817	0.350	-0.774	0.771	0.197	0.076
4	0.963	0.270	-0.405	0.972	0.222	0.051
5	0.874	0.401	-0.514	1.387	0.207	0.044
6	0.872	0.263	-0.083	1.470	0.199	0.056
7	0.802	0.309	-0.011	1.284	0.202	0.063
8	0.818	0.280	-0.313	1.261	0.240	0.051
			Off-level			
3						
4	0.786	0.338	-1.007	0.929	0.238	0.057
5	0.949	0.277	-0.741	0.972	0.222	0.045
6	0.962	0.425	-0.847	1.260	0.220	0.052
7	0.864	0.342	-0.318	1.397	0.202	0.072
8	0.791	0.269	-0.234	1.411	0.217	0.049
			Mathematics			
			On-level			
3	0.805	0.248	-1.287	1.017	0.207	0.048
4	0.918	0.629	-0.090	1.403	0.198	0.061
5	0.767	0.283	-0.351	1.279	0.210	0.080
6	0.947	0.382	-0.128	1.457	0.211	0.080
7	0.840	0.362	0.574	1.668	0.209	0.079
8	1.047	0.389	0.564	1.173	0.186	0.067
			Off-level			
3						
4	0.795	0.261	-1.665	1.041	0.235	0.036
5	0.850	0.374	-0.727	1.416	0.207	0.064
6	0.806	0.304	-0.912	1.111	0.211	0.063
7	0.966	0.397	-0.364	1.418	0.206	0.078
8	0.941	0.379	-0.125	1.402	0.205	0.072

The item parameter estimates were used to simulate the data. The SAS software (SAS Institute Inc., 2009) syntax was used to create the first factor - two distribution shapes (i.e., normal and negatively skewed) and the second factor - two sample sizes (i.e., 1,500 and 3,000 per-form-by-grade) for each Reading and Mathematics on-level and off-level forms at each grade level. Data were simulated for both Reading and Mathematics and each condition was replicated 100 times.

There were four steps to simulating the data. First, normal and negatively skewed *population* distributions of thetas were simulated (N = 2,000,000). The means of these population distributions increased by half a standard deviation per grade above the base grade and decreased by half a standard deviation per grade below the base grade as shown in Table 4. The base grade for both the normal and skewed distributions was grade 6. However, since the skewed distribution was not centered at zero, the grade 6 mean thetas were centered at 0.5.

Table 4.

Mean thetas for the normal and negatively skewed distributions

Grade	Normal	Skewed
3	-1.5	-1.0
4	-1.0	-0.5
5	-0.5	0.0
6	0.0	0.5
7	0.5	1.0
8	1.0	1.5

These distributions were used for this study instead of creating a vertical scale based on one of the scaling procedures to avoid bias in favour of the vertical scaling procedure used to produce the “true” vertical scale. Therefore, the samples for all conditions were selected from the constructed population of thetas. The normally distributed distributions were created using the Normal Distribution function in SAS, with the means shown in Table 4 and a standard deviation of 1. The negatively skewed distributions were created using the RAND Beta Distribution function in SAS with the two shape parameters set to 4 and 2 (see SAS website reference for full description of the RAND function; SAS Institute Inc., n.d.). The negatively skewed distribution *population* of thetas from which the conditions were drawn had skewness of -0.47 and kurtosis of -0.37. The simulated negatively skewed distributions were the pulled from population of thetas with the means shown in Table 4 and a standard deviation of 1. These distributions were created for each grade level.

The second step was to randomly select (1,500 or 3,000) examinees theta values with replacement from each population distribution and sample size condition for each grade level and test form. The samples for each subject area were selected for each test form within each grade level. For example, in the case of the normal distribution of thetas, the grade 3 level test data for Reading was drawn from the distribution with a mean theta of -1.5. For grade 4, the data for the grade 4 test form and the grade 3 form were drawn from the distribution with a mean theta of -1.0.

The third step was to calculate the probabilities for each item parameter and randomly simulate binary (0, 1) response values to create (1,500 or 3,000) item score vectors for each grade level item for both Reading and Mathematics.

A vector of item probabilities were calculated for each selected theta examinee using the 3-PL item parameters obtained from the separate calibration of the real test form data for each item. A random selection based on the Bernoulli distribution was conducted using the RAND function in SAS that uses the calculated item probability vectors for each examinee and simulated the binary (0,1) item values for each examinee (theta).

Lastly, each of the vectors was placed in the appropriate format for the data files for input into the computer programs being used for each scaling procedure. While the data formats for the FCIP and TCC scaling procedures were the same, the file formats were different for the CC, and HCC procedures.

Steps two to four were repeated 100 times, yielding 100 replicates. This procedure was repeated for each simulated condition and content area.

Scaling Procedures.

The fourth factor evaluated in this study was vertical scaling procedure. Four vertical scaling methods were considered: CC, FCIP, TCC, and HCC. The first three procedures were the most common procedures used in previous research and in practice (Briggs & Weeks, 2009; Chin, et al., 2006; Cook & Douglass, 1982; Hanson & Béguin, 2002; Harris & Hooker, 1987; Ito, et al., 2008; Jodoin, et al., 2003; Karkee, et al., 2003; Keller, et al., 2004; J. Kim, 2007; S. H. Kim & Cohen, 1998; Meng, 2007; Tong & Kolen, 2007; Yen & Burket, 1997). The fourth procedure, HCC, was used in four of the previously cited studies (Briggs & Weeks, 2009; Ito, et al., 2008; Karkee, et al., 2003; Meng, 2007) and is a promising vertical scaling procedure that should be systematically evaluated against the other three scaling procedures. The scaling designs for each of the CC, FCIP, TCC and HCC procedures are represented in Figure 3 through Figure 6 (see pages 26, 28, 34, and 36, respectively).

BILOG-MG was used to estimate the item parameters for the four vertical scaling procedures for all conditions. BILOG-MG could not be used to estimate the thetas, since BILOG-MG rescales all of the thetas with a mean of zero and the standard deviation of one. Therefore, to avoid this situation, the IRT Command Language program, called ICL, as described by Hanson in the Manual for ICL (Hanson, 2002) was used to estimate thetas for all of the conditions (i.e., the theta values from BILOG-MG were not used in this study). A description of the scoring syntax and theta score is described on page 78 in the Scoring Procedure section.

Figure 3 shows the scaling design for the CC method (see page 26). The item parameters for the items in the combined forms for Reading at all six grade levels were estimated simultaneously for this method and centered at the base grade 6 to ensure the results across scaling procedures were comparable. The same procedure was followed for Mathematics. An example BILOG-MG syntax file is presented in Figure A2 in Appendix A. The REFERENCE=4 option indicates to BILOG-MG that the center of the scale is grade 6. The other options were the same as the initial BILOG-MG run used with the real data (see Figure A1 in Appendix A).

In contrast to the CC procedure, the FCIP scaling method is a staged process (see Figure 4 on page 28). Quadrature points and weights were required for the BILOG-MG syntax to rescale the FCIP item parameters for all conditions appropriately, since BILOG-MG centers each distribution of item parameters with a mean of zero and standard deviation of one. The quadrature points were adjusted to center each grade with the half standard deviation differences shown in Table 4. The mean weights were calculated for each grade level and condition across replications from the separate concurrent calibration. These values were used for each of the FCIP replications for each set of conditions. Sample BILOG-MG syntax files for Reading and Mathematics are provided for the FCIP procedure at each grade level, except grades 3 and 6 in Figure A3 through Figure A10 in Appendix A. No rescaling was necessary at grade 3 since the item parameters were estimated at grade 4. No quadrature points were required for grade 6, since this was the base grade level. The two options of FIX on the TEST line and NOADJUST on the CALIB line were used to fix the item parameters for the appropriate form. The item parameters from the grade 6 level students' calibration were used as the starting point. The subsequent calibrations were conducted as follows:

1. Grade 6. A data set for the grade 6 level students that included the grade 6 and grade 5 items was created. CC was then used to estimate the grade 6 and grade 5 item parameters.
2. Grade 5. A data set for the grade 5 level students that included the grade 5 items and grade 4 items was created. FCIP calibration was then used where the common item parameters included in the grade 5

level test form were fixed for the values obtained in step 1 and the grade 4 item parameters were estimated.

3. Grade 4. A data set for the grade 4 level students that included the grade 4 items and grade 3 items was created. FCIP calibration was then used where the common item parameters included in the grade 4 level test form were fixed at the values obtained in Step 2 and the grade 3 item parameters were estimated.
4. Grade 3. The grade 3 item parameters were fixed for the grade 3 level students in step three and no further calibration was necessary.
5. Grade 7. A data set for the grade 7 level students that included the grade 7 common items and grade 6 items was created. FCIP calibration was then used where the common item parameters included in the grade 6 level test form were fixed at the values obtained in Step 1 and the grade 7 item parameters were estimated.
6. Grade 8. A data set for the grade 8 level students that included the grade 8 items and grade 7 items was created. FCIP calibration was then used where the common item parameters in the grade 7 test form were fixed at the values determined in Step 5 and the grade 8 item parameters were estimated.

For the TCC scaling procedure, like with FCIP, a staged process occurred (Figure 5 see page 34). Again the base grade level was 6. First, the pairs of different grade level tests administered at each grade level were concurrently calibrated. The CC procedure was used within grade level with horizontal solid lines and ellipses shown in Figure 5. For example, the grade 6 level students had the grade 6 test items and the grade 5 test items in one data set and item parameters were calibrated. A sample syntax file is shown in Figure A11 in Appendix A, and is similar to the initial run syntax file, with only the number of items in each form being different for each grade level. The data sets used for the FCIP procedure were also used for the TCC procedure. Again, six steps were required as follows:

1. Grade 6. A data set for the grade 6 level students that included the grade 6 and grade 5 items was created. CC was then used to estimate the grade 6 and grade 5 item parameters.
2. Grade 5. TCC scaling occurred by placing the grade 5 student items (both grade 5 and grade 4 items) onto a common scale using the TCC transformations with the grade 5 level items from the CC performed at step 1 as the common anchor items.
3. Grade 4. TCC scaling occurred by placing the grade 4 student items (both grade 4 and grade 3 items) onto a common scale using the TCC transformations with the grade 4 level items from the TCC scaling step 2 as the common anchor items.
4. Grade 3. TCC scaling occurred by placing the grade 3 student items (grade 3 items) onto a common scale using the TCC transformations with the grade 3 level items from the TCC scaling step 3 as the common anchor items.

5. Grade 7. TCC scaling occurred by placing the grade 7 student items (both grade 7 and grade 6 items) onto a common scale using the TCC transformations with the grade 6 level items from the CC from step 1 as the common anchor items.
6. Grade 8. The data set for the grade 8 level students created for the FCIP procedure was used. TCC scaling occurred by placing the grade 8 student items (both grade 8 and grade 7 items) onto scale using the TCC transformations with the grade 7 level items from the TCC scaling step 4 as the common anchor items.

The HCC scaling method is a modification of the CC method. As shown in Figure 6 (see page 36) the grade 6 and grade 5 students' scores were concurrently calibrated. In addition, the grades 8 and 7 students and grades 4 and 3 students were concurrently calibrated. The final step was to place the grades 8 and 7 item parameters and grades 4 and 3 item parameters onto a common scale using TCC linking. The TCC linking was adopted for this purpose since the TCC method was used for this purpose in previous research (Briggs & Weeks, 2009; Ito, et al., 2008; Karkee, et al., 2003; Meng, 2007). A sample syntax file is provided in Figure A12 in Appendix A. This syntax file is similar to the files used for the CC procedure, but instead incorporating pairs of grade levels. The HCC scaling was completed in two steps:

1. TCC scaling occurred by placing the grade 7 and 8 student item parameters (grades 8, 7, and 6 item parameters) onto the same scale using the TCC transformations with the grade 6 item parameters from the

concurrent calibration of the grade 6 and 5 level students as the anchor items.

2. TCC scaling occurred by placing the grade 4 and 3 student item parameters (grades 4 and 3 item parametrs) onto a common scale using the TCC transformations with the grade 4 item parameters from the concurrent calibration of the grade 6 and 5 level students as the anchor items.

Taken as a whole, the research design corresponds to a 2 x 2 x 2 x 4 (shape of distribution-by-sample size-by-content area-by-scaling procedure) fully crossed design, yielding 32 conditions with 100 replications per condition. Within each condition all six grade levels were simulated and evaluated.

Dependent Variables

Scoring Procedure.

Item Response Theory (IRT) pattern scoring was used for each of the simulated examinee score vectors for each of the conditions in the simulation study. Expected a posterior (EAP) scoring was used for all of the simulated data sets. While, the BILOG-MG program was used for the estimations of item parameters, it was not appropriate for estimating the EAP theta scores for the rescaled item parameters for the FCIP, TCC, and HCC procedures. To avoid possible bias that might occur by using BILOG-MG for CC and another program for the other three procedures, the IRT Command Language program (Hanson, 2002) was used to estimate the EAP pattern scores for all conditions. The theta estimates were estimated based on the final item parameters after each of the vertical scales was created. A sample syntax file for ICL is provided in Figure A13 in Appendix A. Each of the theta values were converted to a scale score by multiplying the theta value by 50 and adding 500 to facilitate the interpretation of the root-mean-squared-difference.

Cut-scores rescaling.

An initial set of cut-scores for each grade and content area was provided on the theta scale separately for each grade. These cut-scores produced three proficiency groups called *below basic, basic,* and *above basic.* An adjustment was required to place the cut-scores onto the vertical scale metric. Initially, the cut-scores were adjusted following the same procedure to determine the means of the population distributions for each grade level (see Table 4). However, in conducting the analyses for the four vertical scaling procedures, the mean scale scores did not uniformly differ by a half standard deviation as the grades increased and decreased from the base grade for the CC, TCC and HCC procedures. However, the results for the FCIP procedure did increase appropriately. Therefore to make the comparisons fairer, the quadrature points used in the FCIP procedure were adjusted so that the mean scale scores for the FCIP procedure more closely matched the mean scale scores for the other three vertical scaling procedures. The description of the adjustment for FCIP is presented first, then the description for the cut-scores rescaling is described.

The quadrature points for the FCIP procedure were adjusted until similar student scale scores were found for the CC, TCC and HCC procedures for the normal distribution and the 3,000 sample size condition. For example, if the mean scale score for CC was a half standard deviation above the scale score for FCIP, the 40 quadrature points were increased by a half standard deviation. The same quadrature points were used for all conditions for FCIP. The final quadrature points used for each grade are provided in the sample syntax file for FCIP in Figure A3 to Figure A10 in Appendix A. The mean and standard deviations of the scale scores for each condition is presented in Appendix B in Tables B1 to B4 for reference purposes.

The cut-scores were adjusted so that they were appropriate for the larger separation of scale scores found for all grade levels. Because of this, the mean scale scores for the normal distribution and 3,000 sample size condition for Reading and for Mathematics were used. First, the mean scale score for the normal distribution and 3,000 sample size was calculated for each of the four scaling methods for each grade level and content area from the 100 replications. Second, the difference in mean scale score from the expected to the actual scores was standardized by dividing by the standard deviation of 50. For example, the expected mean scale score for grade 3 Reading was 425; the mean scale score for Reading were 407.40, 399.31, 391.13, and 395.33 for CC, FCIP, TCC, and HCC, respectively. The mean of these mean scale scores was 398.29, and the difference was -26.71, which was then divided by 50, which produced a standardized difference of -0.534. This adjustment was calculated and then applied to each cut-score in each content area and grade level. Table 5 shows the final cut-scores for Reading and Mathematics on the transformed theta and score scales. These final cut-scores were used in the evaluation procedures to evaluate the four scaling procedures.

Table 5.

Final Cut-scores for Reading and Mathematics

	Reading			
	Lower cut-score		Upper cut-score	
Grade	Theta	Scale Score	Theta	Scale Score
3	-2.62	369	-0.70	465
4	-1.95	403	0.19	509
5	-1.22	439	0.76	538
6	-0.46	477	1.45	572
7	0.15	507	2.18	609
8	0.80	540	2.62	631
	Mathematics			
	Lower cut-score		Upper cut-score	
Grade	Theta	Scale Score	Theta	Scale Score
3	-3.05	347	-1.24	438
4	-2.30	385	-0.34	483
5	-1.67	416	0.27	514
6	-0.45	477	1.31	566
7	0.09	505	2.03	601
8	1.36	568	3.08	654

Evaluation Measures.

Five statistical indices were used to evaluate the results: decision accuracy, decision consistency, conditional standard error at the cut-scores, RMSD of the transformed scale scores, and correlations of the final item parameters across scaling methods.

Evaluation of Assignments to the three proficiency levels.

Three measures evaluating the standard setting outcomes for each of the different scaling procedures were employed in this study. The first two statistical measures were decision accuracy and decision consistency of the classifications of students into proficiency categories using the adjusted cut-scores. Standard setting procedures are first employed by the state agency to set the specific values for the cut-scores to divide the student population into proficiency groups, such as masters and non-masters. There are various methods to set the cut-scores but those descriptions are not provided in this text (see citation for full description of methods; Crocker & Algina, 1986). Decision accuracy indicates how accurate the decisions are in placing students in categories. The estimated probability of decision accuracy is the number of examinees accurately classified as masters using observed scores and the estimated true scores divided by the total number of examinees. Several methods have been developed to estimate decision accuracy for a single administration (Lee, Hanson, & Brennan, 2002; Livingston & Lewis, 1995; Livingston & Wingersky, 1979; Rudner, 2005; Wilcox, 1977). Decision consistency indicates how consistently the decisions are made for placing students in categories. The estimated probability of decision consistency is the number of examinees consistently classified as masters using only observed scores divided by the total number of examinees (Crocker & Algina, 1986; Livingston & Lewis, 1995). Several procedures have been developed to estimate decision consistency for a single administration (Hanson & Brennan, 1990; Huynh, 1976; Lee, et al., 2002; Livingston & Lewis, 1995; Peng & Subkoviak, 1980; Subkoviak, 1976; Wang, Kolen, & Harris, 2000; Wilcox, 1981).

The procedure used to examine the decision accuracy and decision consistency in the present study was the Livingston and Lewis (1995) procedure, and the calculations were performed using the BB-CLASS Version 1.1 computer program (Brennan, 2004). Decision accuracy and consistency outcome measures were calculated for the two adjusted cut-scores based on the final scale scores for each condition at each grade level. The mean decision accuracy and mean decision consistency for each cut-score were calculated across the 100 replications for each of the conditions and grades in this study. For the calculation of decision accuracy and consistency, the cut-score results were calculated in separate computer runs for each cut-score. The overall decision accuracy and consistency values across both cut-scores at the same time were not calculated.

The third measure to evaluate each cut-score was the conditional standard error at the cut-score. The conditional standard error at a cut-score is referred to as the standard error of the estimate of the theta value associated with the cut-score (Hambleton, et al., 1991). The standard error of estimate is defined as the square root of the reciprocal of the amount of information provided by a test at the theta value corresponding to the cut-score, $\hat{\theta}_c$:

$$SE\left(\hat{\theta}_c\right) = \frac{1}{\sqrt{I(\theta)}},$$

where $SE\left(\hat{\theta}_c\right)$ is the standard error of estimation of $\hat{\theta}_c$ and $I(\theta)$ is the information value at the cut-score (Hambleton, et al., 1991). The standard error of estimation differs for each theta point in the distribution. These calculations were based on the transformed theta scores for each of the vertical scale procedures and not on the transformed scale scores. Squared conditional standard errors at each cut-score were calculated for each replication and then averaged across the conditions and grades. Then the square root was taken of the squared conditional standard errors to calculate the conditional standard error for each set of conditions.

Evaluation of IRT outcomes.

The fourth statistical measure used to compare the four scaling methods was the root-mean-squared-difference (RMSD) between the transformed scale scores for each of the conditions averaged across replications. The RMSD for a scaling procedure for each condition is defined as:

$$RMSD_p = \sqrt{\frac{1}{r}\sum_{r=1}^{r}\frac{1}{n}\sum_{n=1}^{n}\left(SS_{p_1} - SS_{p_2}\right)^2},$$

where r is the number of replications, where n is the total number of examinees and θ_{p_1} is the scale score based on the theta for one of the two different scaling procedures and θ_{p_2} is the scale score based on the theta for another scaling procedure (e.g., CC versus FCIP). The RMSD value was calculated for each of the conditions and grades.

The fifth statistical measure used to evaluate the scaling procedures was the correlation between the item parameters obtained for each scaling method. Pearson correlations between each pair of *a*-, *b*-, and *c*- parameters were transformed to Fishers *Z* for each condition, then the mean Fishers *Z's* across conditions were calculated. The mean Fisher's *Z* values were transformed back to the correlation metric, using the anti-log of the Fisher's *Z*.

CHAPTER 4 RESULTS: READING

The results for Reading are presented in this chapter. The results for Mathematics are presented in Chapter 5. The results are presented for each evaluation measure for the four distribution shape and sample size conditions. The decision accuracy and consistency results are presented first followed by the presentation of the conditional standard errors of estimation at the cut-scores. Third, the root-mean-squared-differences (RMSDs) of the scale scores are presented. Lastly, the correlations between the item parameters across vertical scaling procedures are presented. The results presented in Chapters 4 and 5 provide a micro evaluation of the differences. Chapter 6 presents a summary of the micro evaluation, and presentation of a macro discussion and applications for practitioners.

The presentation of the results for the first condition – a normal distribution with a sample size of 1,500 – includes a summary table and a graphical representation of the results for decision accuracy, decision consistency, conditional standard error estimates, and RMSD, and a summary table for the correlations of the item parameter estimates. The presentation for each of the remaining conditions includes only a graphical representation for decision accuracy, decision consistency, conditional standard error estimates, and RMSD and a summary table for the correlations of the item parameters estimates. The tables corresponding to the graphs for the remaining conditions are provided in Appendix B.

Decision Consistency and Accuracy

Normal distribution 1,500 sample size

The decision accuracy and consistency results are presented for the normal distribution with 1,500 examinees in Table 6. The results for all four vertical scaling methods are presented for each grade and cut-score. As indicated earlier, there are three performance levels – *below basic*, *basic,* and *above basic*. For all tables the cut-score between *below basic* and *basic* is listed as the lower cut-score, and the cut-score between *basic* and *above basic* is listed as the upper cut-score. The decision accuracy and consistency values are bounded by zero and one, where, for example, a value of 0.80 is interpreted as 80 % accurate or consistent. However no specific minimum criteria have been identified for decision accuracy and consistency in the literature. Therefore, the minimum value for the present study was set at 0.80 as 0.80 represents 80% both accurate or consistent classifications, and the values below 0.80 are bolded in the tables.

Table 6.

Decision Accuracy and Consistency for Vertical Scaling for Reading, normal distribution, 1,500 sample size

Accuracy					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.825	0.817	**0.579**	**0.735**
	upper	0.890	0.910	0.930	0.932
4	lower	0.819	**0.755**	0.808	**0.766**
	upper	0.934	0.949	0.957	0.956
5	lower	**0.734**	**0.719**	**0.729**	**0.722**
	upper	0.942	0.954	0.945	0.950
6	lower	**0.676**	**0.688**	**0.688**	**0.684**
	upper	0.956	0.945	0.945	0.947
7	lower	**0.649**	**0.702**	**0.712**	**0.696**
	upper	0.988	0.978	0.980	0.981
8	lower	**0.648**	**0.723**	**0.725**	**0.714**
	upper	0.997	0.991	0.993	0.993
Consistency					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.826	**0.795**	**0.797**	0.802
	upper	0.935	0.944	0.954	0.955
4	lower	0.856	0.835	0.821	0.825
	upper	0.944	0.954	0.962	0.962
5	lower	0.856	0.848	0.848	0.849
	upper	0.937	0.948	0.938	0.943
6	lower	0.848	0.853	0.853	0.853
	upper	0.948	0.939	0.939	0.940
7	lower	0.813	0.830	0.832	0.827
	upper	0.980	0.967	0.970	0.972
8	lower	0.819	0.845	0.839	0.840
	upper	0.993	0.984	0.987	0.986

As shown in the upper panel of Table 6, the pattern for decision accuracy at the lower cut-score was not consistent across scaling procedures. For the two lower grades for the lower cut-score, the decision accuracy was below 0.80 for

TCC (0.58) and HCC (0.74) at grade 3 and for FCIP (0.76) and HCC (0.77) for grade 4. For grade 3 and 4 lower cut-score, the values of decision accuracy values were not substantially above 0.80 compared to the corresponding upper cut-score (0.81 to 0.83 vs. 0.89 to 1.00). For the four upper grades, the decision accuracy at the lower cut-score was less than 0.80 for all four scaling procedures, ranging from 0.65 (grades 7 and 8, CC) to 0.73 (grade 5, CC and TCC; grade 8, TCC).

However, the decision accuracy values for the four scaling procedures exceeded 0.80 for the upper cut-score for all grade levels. At the upper cut-score, the decision accuracy across grades and scaling methods varied from 0.89 (grade 3, CC) to essentially 1.00 (grade 8, CC). There were smaller differences across the four vertical scaling procedures at the upper cut-score, with all values of decision accuracy above 0.80.

All of the decision consistency values were equal to or greater than 0.80. However, the decision consistency values at the lower cut-score were lower than the corresponding decision consistency at the upper cut-score which ranged from 0.80 to 0.86 vs. 0.94 to 0.99 for all four scaling procedures. Given the small range among the decision consistency values across the four scaling procedures the four methods behaved similarly at each grade level and across grade levels.

The decision accuracy and consistency results presented above are graphically shown, respectively, in Figure 7 and Figure 8. The grade levels are displayed on the X-axis. Decision accuracy and consistency values are shown on the Y axis bounded by 0.40 and 1.00 to show the differences between procedures more easily. Both the lower and upper cut-scores are displayed on the same graph, where the solid line indicates the lower cut-score and the dashed line indicates the upper cut-score. Each vertical scaling procedure is shown in a different colour with a different marker, CC is blue (square marker), FCIP is red (diamond marker), TCC is green (circle marker), and HCC is yellow (triangle marker).

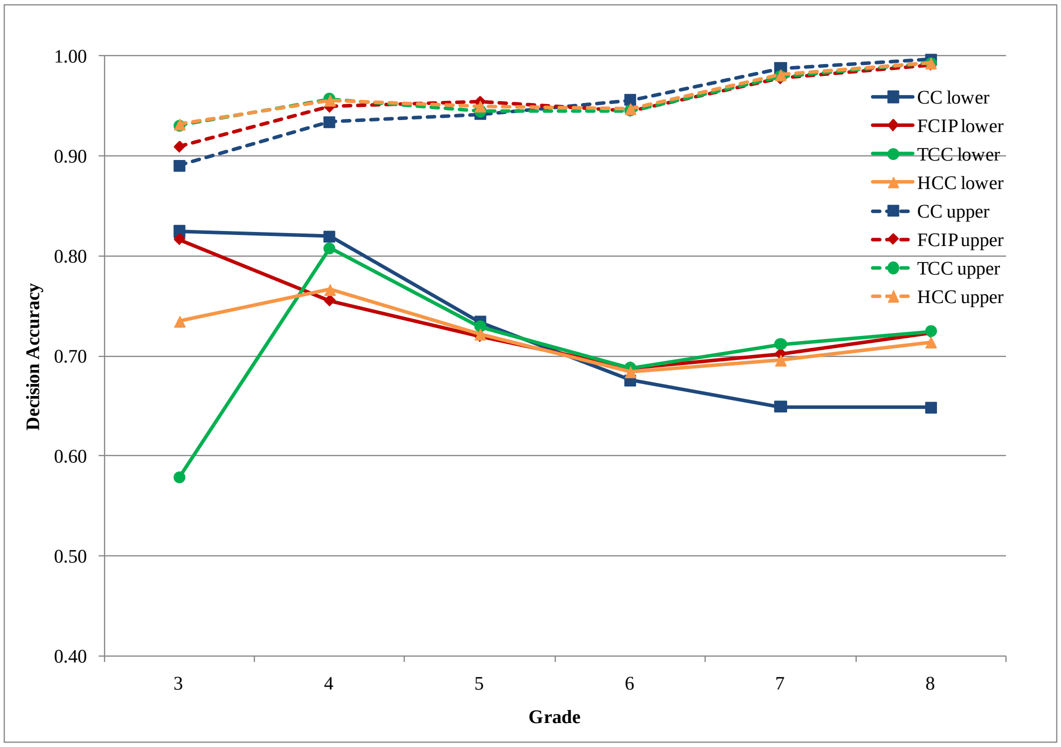

Figure 7. Decision accuracy for Reading, normal distribution, 1,500 sample size

For the lower cut-score, larger differences across the vertical scaling procedures were found, especially for the lowest grade level, with the TCC procedure having the lowest value. For example, a difference of 0.25 between the TCC (0.58) and CC (0.83) procedures were found. With these larger differences it appears the vertical scaling procedure may produce decision accuracy differentially at some grade levels. The CC procedure had the largest values for most grade levels, except for grade 7 and 8. The most consistent procedures were the FCIP and HCC procedures, with few extreme values either high or low.

For the upper cut-score, similar results across the four vertical scaling procedures were found, where all but one (CC, grade 3) of the values were greater than 0.90. The lowest values were found in the two lower grade levels for the CC procedure followed by the FCIP procedure.

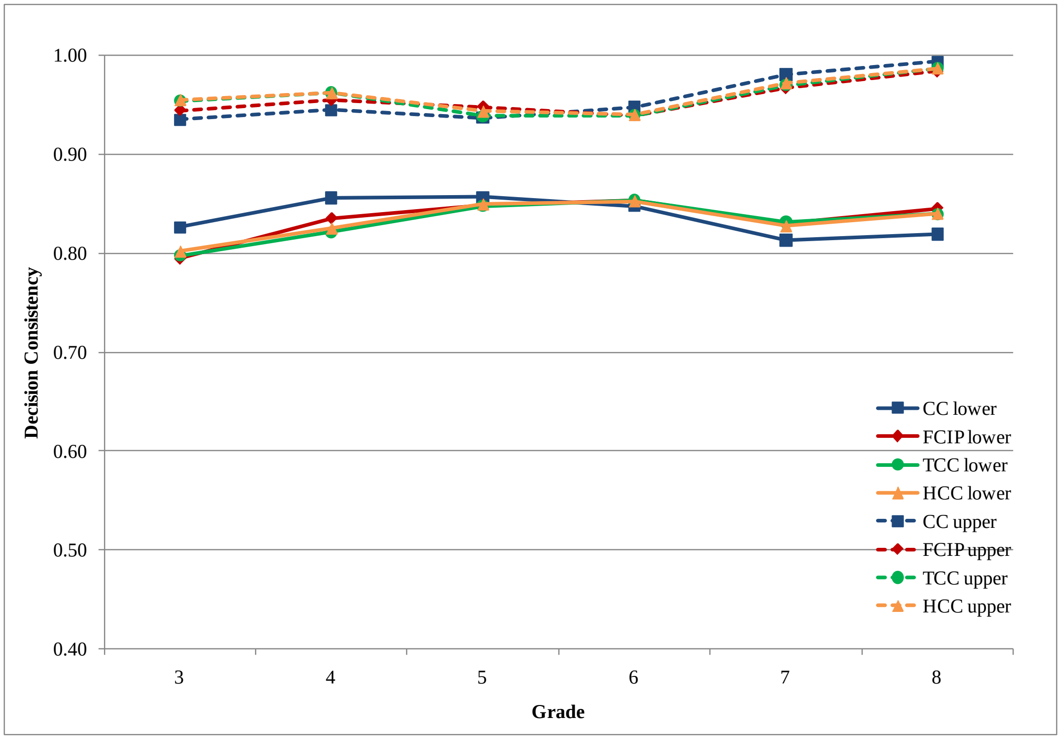

Figure 8. Decision consistency for Reading, normal distribution, 1,500 sample size

The line graphs shown in Figure 8, when compared to the line graphs in Figure 7, indicate that the decision consistency values across the four vertical scaling procedures and grades were more similar than the decision accuracy values. For the lower cut-score, the CC procedure had the largest decision consistency values for the two lower grade levels, but lower decision consistency values for the two highest grade levels, similar to those results found for decision accuracy. Generally all of the decision consistency values were above 0.80, and were consistent across the FCIP, TCC, and HCC procedures across grade levels.

Similar results were found for the upper cut-score where all the decision consistency values were above 0.90. It appears that for decision consistency the vertical scaling procedure does not impact decision consistency across grade levels.

Normal distribution 3,000 sample size

The decision accuracy and consistency results are displayed, respectively, in Figure 9 and Figure 10 for the normal distribution and a sample size of 3,000; the corresponding numerical values are provided in Table B5 in Appendix B.

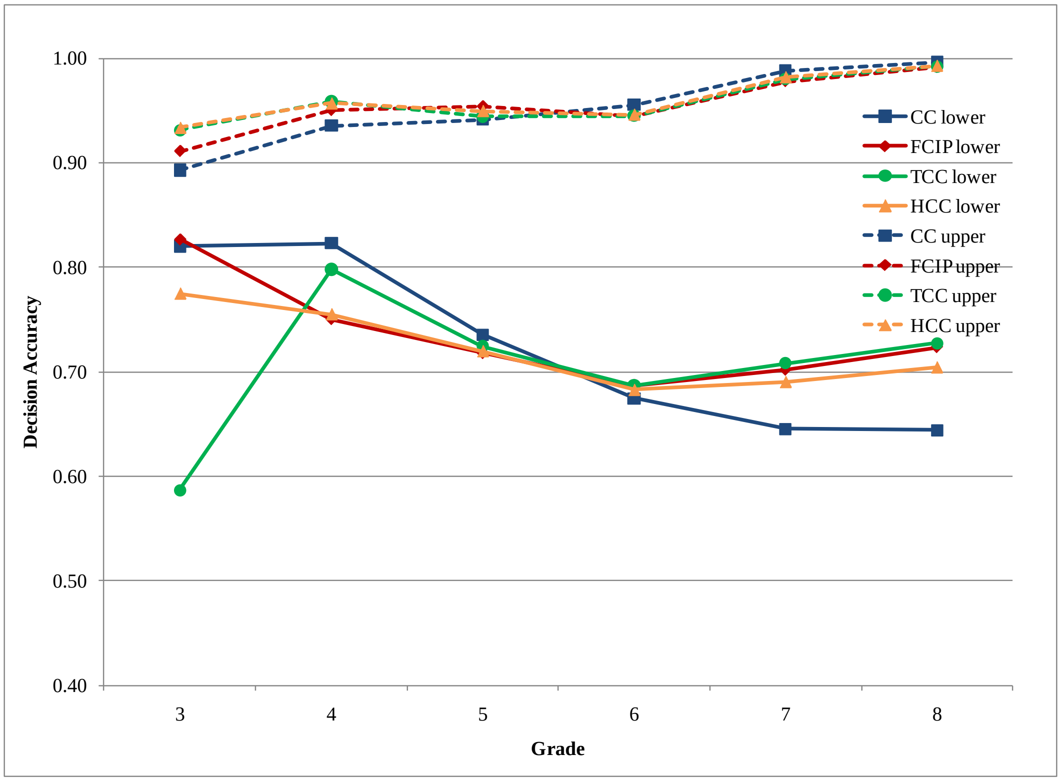

Figure 9. Decision accuracy for Reading, normal distribution, 3,000 sample size

As shown in Figure 9, the decision accuracy for Reading, normal distribution, 3,000 sample size was higher and less variable at the upper cut-score than at the lower cut-score. Larger differences across the vertical scaling procedures were found at the lower cut-score, especially for grade 3, with the TCC procedure having a markedly lower value. For example, a difference of 0.24 between TCC (0.59) and FCIP (0.83) was found. At the lower cut-score, the CC procedure generally had the largest values, except for grades 7 and 8. The most consistent procedures were the FCIP and HCC procedures.

For the upper cut-score, similar results across the four vertical scaling procedures were found, where all but one (CC, grade 3) of the decision accuracy values were greater than 0.90. The lowest values were found at grade 3. The values were somewhat consistent from grade 4 though 6, then gradually increased to 1.00 for the upper grade levels.

The results for the normal distribution and 3,000 sample size condition were fairly consistent with the results from the normal distribution 1,500 sample size condition. Therefore, it appears that sample size did not influence decision accuracy when the scale scores were normally distributed.

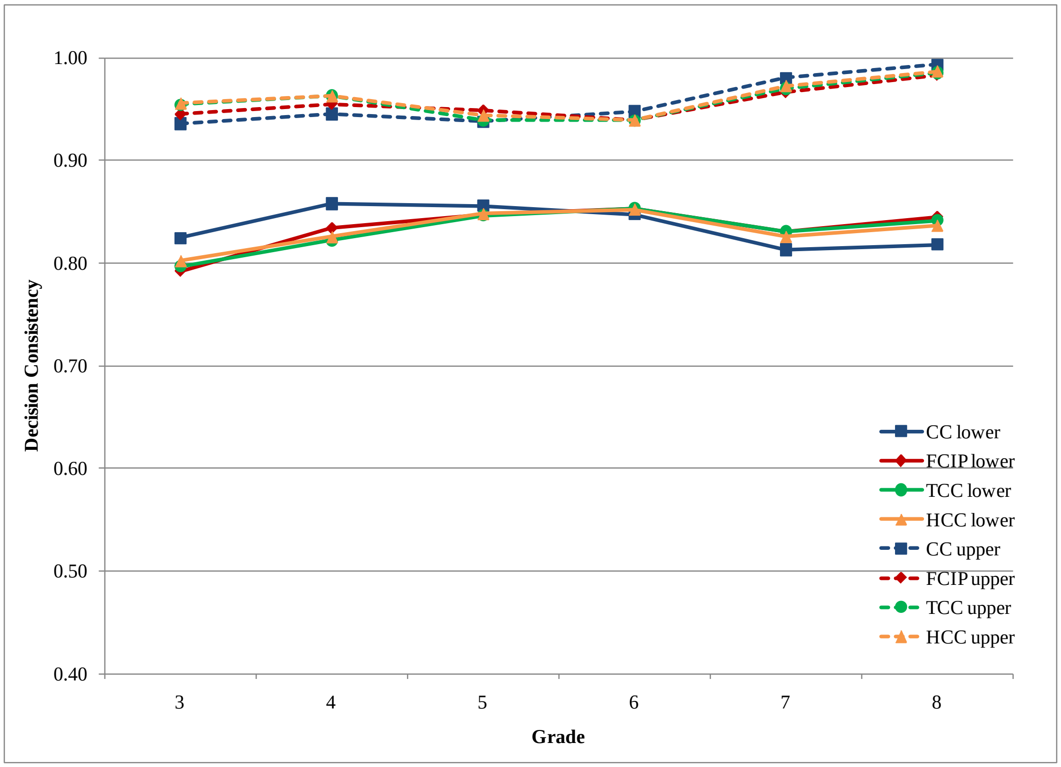

Figure 10. Decision consistency for Reading, normal distribution, 3,000 sample size

As shown in Figure 10, the decision consistency values across the four vertical scaling procedures and grades were generally quite similar. The CC procedure had the largest decision consistency values for the two lower grade levels for the lower cut-score, but lowest values for the two highest grade levels. However the values were all above 0.80.

Similar results were found for the upper cut-score, with all decision consistency values above 0.90 and close to 1.00 for the two upper grade levels. Slight differences were found at the two lower grade levels, where the CC procedure had the smallest values, and at the three upper grade levels, where the CC procedure also had the largest values.

Skewed distribution 1,500 sample size

The decision accuracy and consistency results are displayed, respectively, in Figures 11 and 12 for the skewed distribution and a sample size of 1,500; the corresponding values are provided in Table B6 in Appendix B.

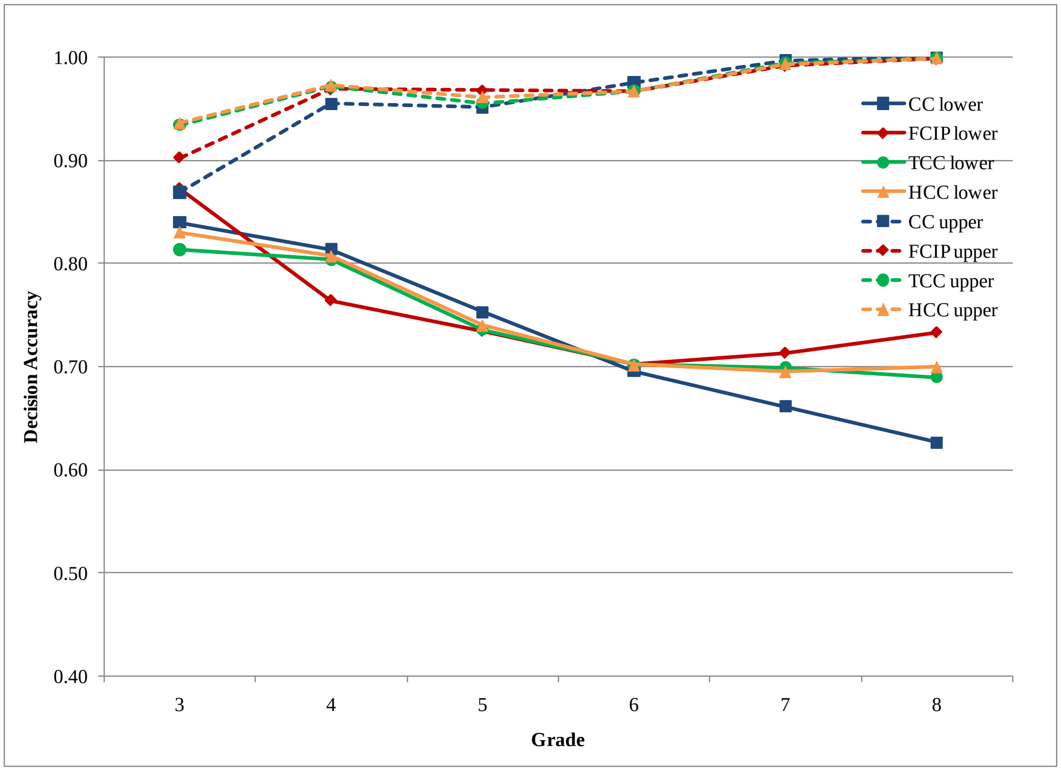

Figure 11. Decision accuracy for Reading, skewed distribution, 1,500 sample size

At the lower cut-score, there is a decrease in decision accuracy from grade 3 to grade 6 for all four scaling procedures. For grades 7 and 8 the HCC and TCC procedures remained constant at about 0.70, but the decision accuracy values for the FCIP procedure were slightly higher than the HCC and TCC procedures and the decision accuracy values for the CC procedure were slightly lower. The decision accuracy values exceeded 0.80 for all four procedures at grade 3 and for the CC, TCC, and HCC procedures at grade 4. The decision accuracy for all four procedures was less than 0.80 for the four upper grade levels.

At the upper cut-score, the decision accuracy values exceeded 0.90 with one exception (CC grade 3) for the four scaling procedures and grade levels. The greatest variability among the decision accuracy values was at grade 3, where the decision accuracy values for the CC procedure was less than 0.90, the FCIP, TCC and HCC procedures were all over 0.90.

Comparison of Figure 7 (see page 89) and Figure 11 shows that the shape of the score distribution differentially influenced the decision accuracy of the four scaling procedures, particularly at the lower cut score. In particular, whereas the decision accuracy values decreased more so for the skewed distribution condition for the upper grade levels, the decision accuracy values for the normal distribution condition were lower for the upper grade levels as grades increased.

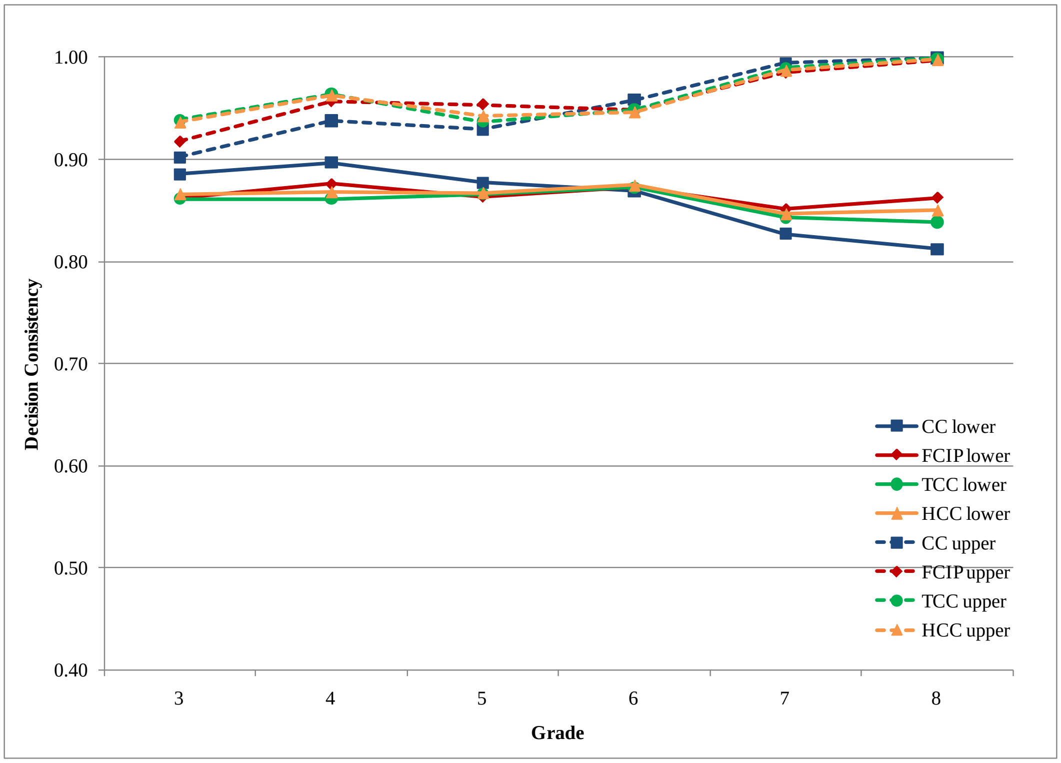

Figure 12. Decision consistency for Reading, skewed distribution, 1,500 sample size

As shown in Figure 12, the decision consistency exceeded 0.80 for all four procedures and grade levels for the lower cut-score skewed distribution, sample size of 1,500. For the lower cut-score, the CC procedure had the largest values for the three lower grade levels, and the lowest values for the two higher grade levels. The FCIP, TCC, and HCC procedures had similar values that were essentially unchanged from grade 3 to grade 6 and then decreased for grades 7 and 8.

The decision consistency values for the upper cut-score for all vertical scaling procedures were at least 0.90 across the grade levels, with lower values for grade 3 and higher values for grades 7 and 8. The decision consistency for the TCC and HCC procedures were similar across all grades, while the decision consistency of the CC procedure was the lowest at grades 3 and 4, followed by the FCIP procedure at grade 3. However, as previously indicated, the decision consistency for all four scaling procedures was at least 0.90. In contrast to decision accuracy, the decision consistency values of the four scaling procedures were not influenced as much by the change in shape for the score distribution (cf, Figure 8 page 90 and Figure 12 page 96).

Skewed distribution 3,000 sample size

The decision accuracy and consistency results are displayed, respectively, in Figures 13 and 14 for the skewed distribution and a sample size of 3,000; the numerical values are provided in Table B7 in Appendix B.

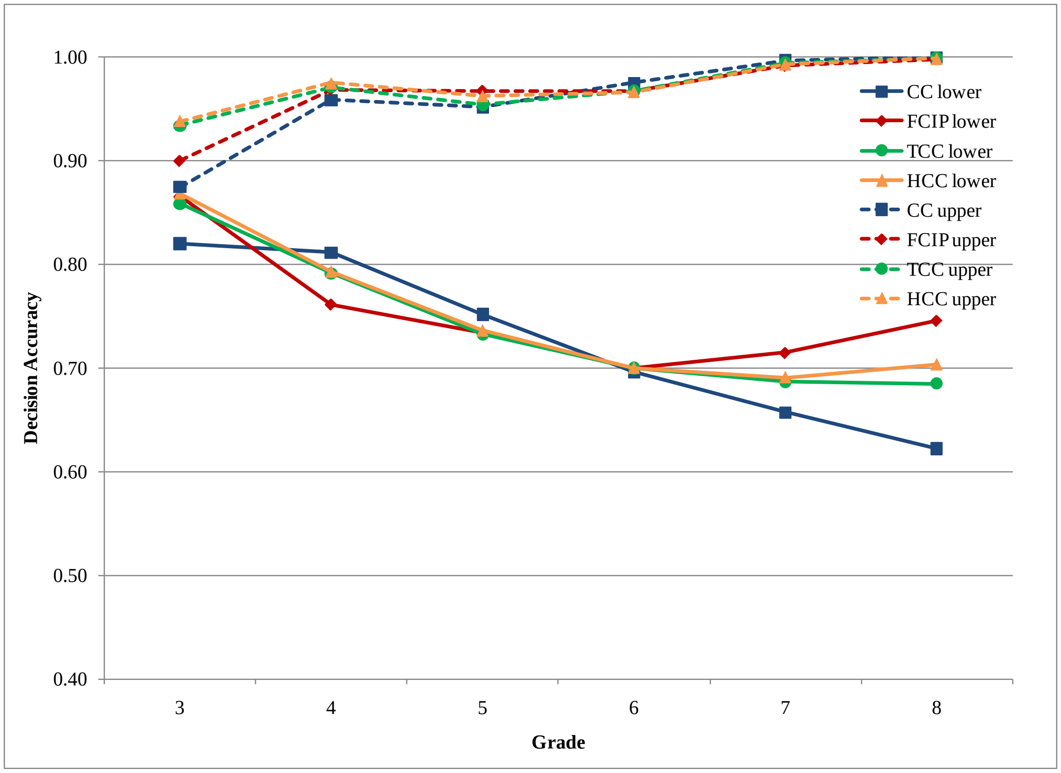

Figure 13. Decision accuracy for Reading, skewed distribution, 3,000 sample size

Similar to the skewed distribution 1,500 sample size condition, the shape of the score distribution differentially influenced the decision accuracy of the four scaling procedures, particularly at the lower cut-score, a similar way. At the lower cut-score, the decision accuracy values decreased from grade 3 to grade 6. For grades 7 and 8, the decision accuracy values for the HCC and TCC procedures remained relatively constant at about 0.70, the decision accuracy for the FCIP procedure slightly increased, and decreased for the CC procedure. There was a slight difference between the TCC and HCC procedures at grade 8 with the TCC procedure being slightly lower than 0.70. The decision accuracy values exceeded 0.80 for all four scaling procedures at grade 3 and for the CC procedure at grade 4. The decision accuracy values for all four scaling procedures were less than 0.80 for the four remaining upper grade levels.

At the upper cut-score, the decision accuracy values for the four vertical scaling procedures exceeded 0.80 for all four scaling procedures and grade levels, with the exception of the CC procedure at grade 3. The greatest variability among the decision accuracy values was at grade 3, where the values were less than 0.90 for the CC procedure and were over 0.90 for the FCIP, TCC and HCC procedures. The decision accuracy values were up to 1.00 for grades 7 and 8.

The decision accuracy for the TCC and HCC procedures at both cut-scores were similar across all grade levels. In contrast, the values for the CC and FCIP procedures varied, particularly for grades 3 and 4 and grades 7 and 8 at the lower cut-score and grade 3 at the upper cut-score.

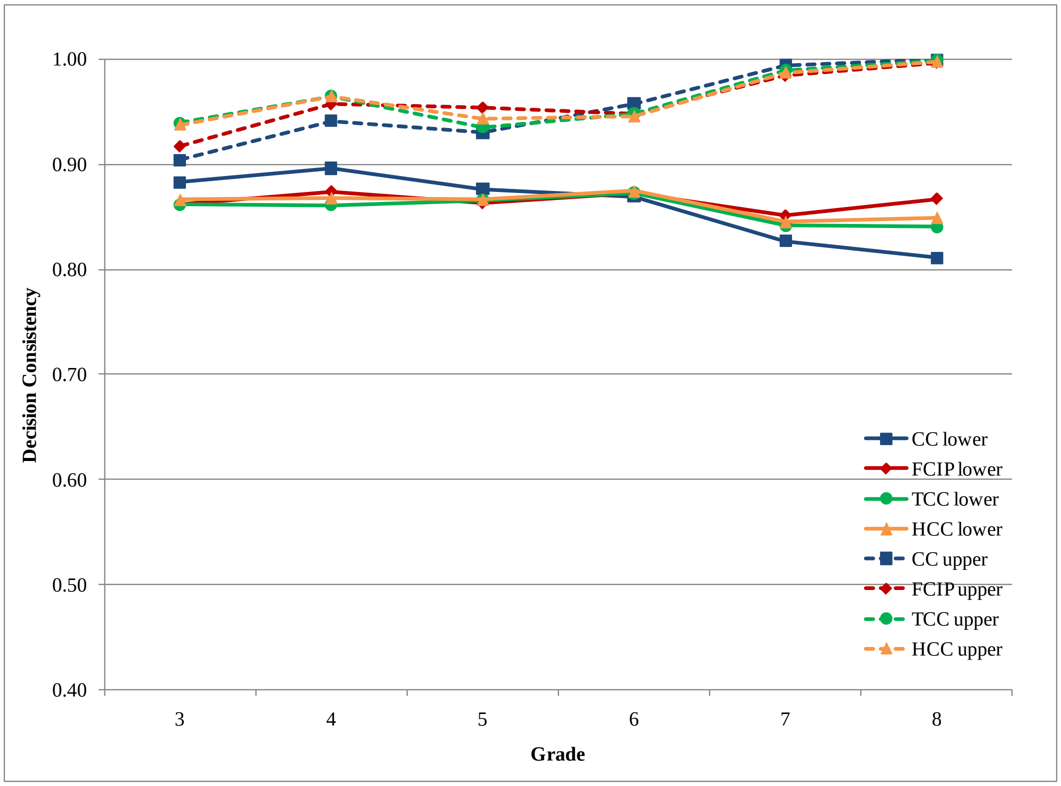

Figure 14. Decision consistency for Reading, skewed distribution, 3,000 sample size

The decision consistency values for all of the vertical scaling procedures were not influenced as much by the change in shape for the score distribution (cf., Figure 10 page 93, and Figure 14). At the lower cut-score, the decision consistency exceeded 0.80 for all four scaling procedures and grade levels. The CC procedure had the largest decision consistency values for the three lower grade levels, and the lowest values for the two higher grade levels. The FCIP, TCC, and HCC procedures had similar values that were very similar from grade 3 to grade 6 and then decreased to approximately 0.85 for grades 7 and 8. The highest decision accuracy value for grade 8 was for the FCIP procedure.

As shown in Figure 14, the decision consistency values were at least 0.90 for the four scaling procedures at the upper cut-score across the grade levels, with lower values for grade 3 and higher values for grades 7 and 8. The decision consistency for the TCC and HCC were similar across all grade levels, while the decision consistency of the CC procedure was the lowest at grades 3 and 4 followed by the FCIP procedure at grade 3. However, as previously indicated, the decision consistency for all four scaling procedures was at least 0.90.

Similar to the results found for the normal distribution and 1,500 and 3,000 sample size conditions, these results did not seem to be affected by the sample size conditions evaluated in this study. However, the skewed distribution 3,000 sample size condition compared to the normal distribution 3,000 sample size condition had similar patterns as compared to the normal and skewed 1,500 sample size conditions. That is, a similar pattern of decreasing decision accuracy as grade levels increased was found for the skewed distribution 3,000 sample size condition.

Summary for Decision Accuracy and Decision Consistency

The decision accuracy results were generally consistent with a few exceptions, more so for the upper cut-score than for the lower cut-score within distribution shape across sample size. There were some differences for decision accuracy for the lower cut-score at particular grade levels; for example, grade 3 had different values for the different vertical scaling procedures. There was a decrease in decision accuracy as grade increased for the lower cut-score for both the normal and skewed distribution conditions. However, the decision accuracy value' decrease for the skewed distribution across grade levels was more pronounced than for the normal distribution. This finding indicates that the accuracy of the decisions may not be consistent across scaling methods and grade levels. An increase in sample size from 1,500 to 3,000 did not have an impact on the results.

The decision consistency results were more consistent for both the upper and lower cut-scores than decision accuracy. While the values were higher for the upper cut-score than the lower cut-score, the values for both cut-scores were quite high. There were few differences among the four vertical scaling procedures across grade levels and the four distribution/sample size conditions. Therefore, decision consistency does not seem to be impacted for vertical scaling procedure, distribution shape, sample size, and grade level for the Reading assessment.

Conditional Standard Error

Normal distribution 1,500 sample size

The values of the conditional standard error at each cut-score are reported for the normal distribution 1,500 sample size condition in Table 7 and shown graphically in Figure 15. Since there is no clear metric to indicate if a conditional standard error is too high, no minimum value has been set.

Table 7.

Conditional Standard Error for Reading, normal distribution, 1,500 sample size condition

Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.229	0.234	0.251	0.230
	upper	0.408	0.414	0.492	0.469
4	lower	0.211	0.224	0.228	0.223
	upper	0.395	0.436	0.485	0.481
5	lower	0.290	0.289	0.315	0.305
	upper	0.681	0.701	0.685	0.704
6	lower	0.312	0.297	0.297	0.302
	upper	0.663	0.632	0.632	0.628
7	lower	0.404	0.361	0.336	0.378
	upper	0.864	0.734	0.699	0.789
8	lower	0.499	0.461	0.414	0.449
	upper	1.081	0.975	0.900	0.950

As shown in Table 7, the conditional standard errors for the lower cut-scores were consistently less than the standard errors for the upper cut-scores at each grade level. The conditional standard errors for the four scaling procedures for both the lower and upper cut-scores were similar at grades 3 and 4, and grades 5 and 6, and less so at grades 7 and 8. The conditional standard errors generally increased from grade 3 through 8, with the increase more so for the upper cut-scores than for the lower cut-scores. Further, the variability among the standard errors across grades and procedures was less for the lower cut-scores (0.21 to 0.45) than for the upper cut-scores (0.40 to 1.08). Figure 15 shows the conditional standard error results for the same information provided in Table 7.

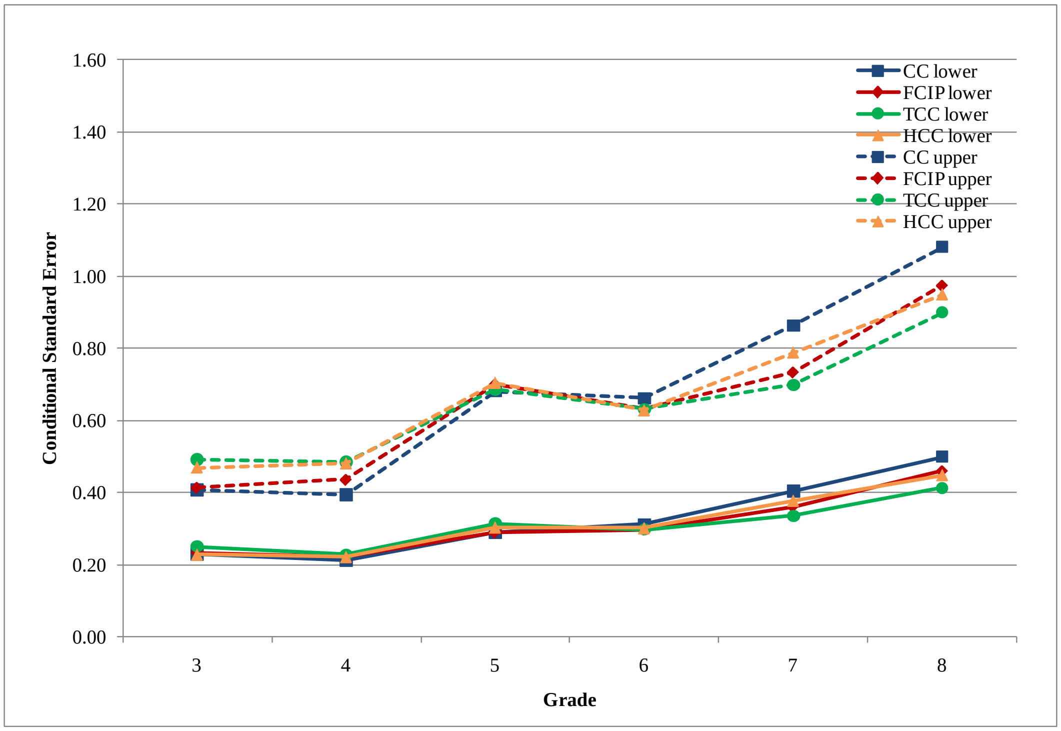

Figure 15. Conditional Standard Error for Reading, normal distribution, 1,500 sample size

As shown in Figure 15, the conditional standard errors for the four scaling procedures were flat, then increased, then flat, and then increased for pairs of grades starting with grades 3 and 4. Further, the four scaling procedures for the lower cut-score were relatively consistent with small differences across grade levels for the upper two grade levels. For the upper grade levels, lower values were found for the TCC procedure and slightly higher values were found for the CC, FCIP and HCC procedures. Although, the TCC procedure had the highest value for the lowest grade 3, these conditional standard error values for TCC were similar to the conditional standard errors for the other three scaling procedures at the upper grades.

The conditional standard error results for the upper cut-score were larger than those for the lower cut-score and larger differences between the procedures were apparent. The CC procedure had the lowest values for the three lowest grade levels but had the largest values for the three highest grade levels. Similarly, the TCC procedure had the reverse pattern, where the lowest values were for the upper grade levels and the largest values for the lowest grade levels. The most consistent procedures were HCC and FCIP, where the conditional standard error values were largely in the middle between the other scaling procedures.

Normal distribution 3,000 sample size

The mean conditional standard error results are displayed in Figure 16 for a normal distribution and a sample size of 3,000. The mean conditional standard error values are shown in Table B8 in Appendix B.

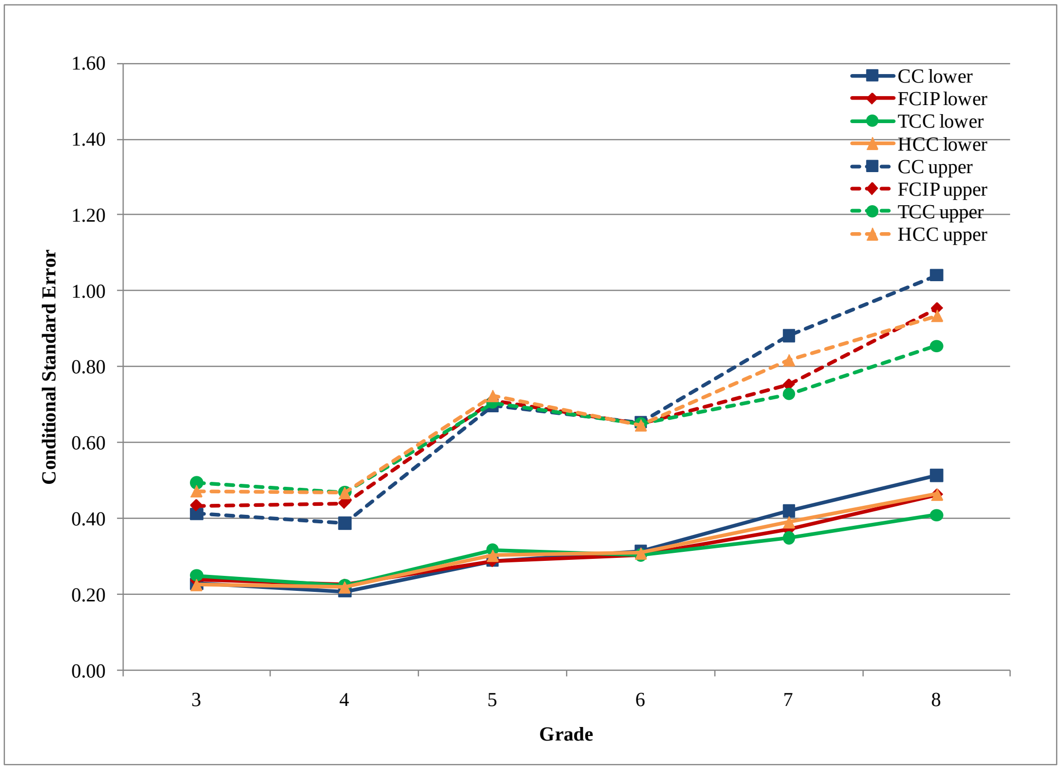

Figure 16. Conditional Standard Error for Reading, normal distribution, 3,000 sample size

As shown in Figure 16, there was close agreement among the conditional standard error values for grades 3 through 6 and generally increased across the six grade levels for the lower cut-score. At grades 7 and 8, the conditional standard error values for the four scaling procedures were slightly different, with larger differences for grade 8. Lower values were found for the TCC procedure and slightly higher values were found for the FCIP and HCC procedures.

The conditional standard error values were larger for the upper cut-score than the values for the lower cut-score. Also, larger differences between the vertical scaling procedures were found. The CC procedure had the lowest values for the two lower grade levels but had the largest values for the two highest grade levels. In contrast, the TCC procedure had the reverse pattern, where the lowest values were for the upper grade levels and the largest values for the lower grade levels. These results were consistent with those found in the normal distribution 1,500 sample size condition, indicating that sample size did not affect the results of normal distributions.

Skewed distribution 1,500 sample size

The mean conditional standard error results are displayed in Figure 17 for the skewed distribution and a sample size of 1,500. The conditional standard error values are shown in Table B9 in Appendix B.

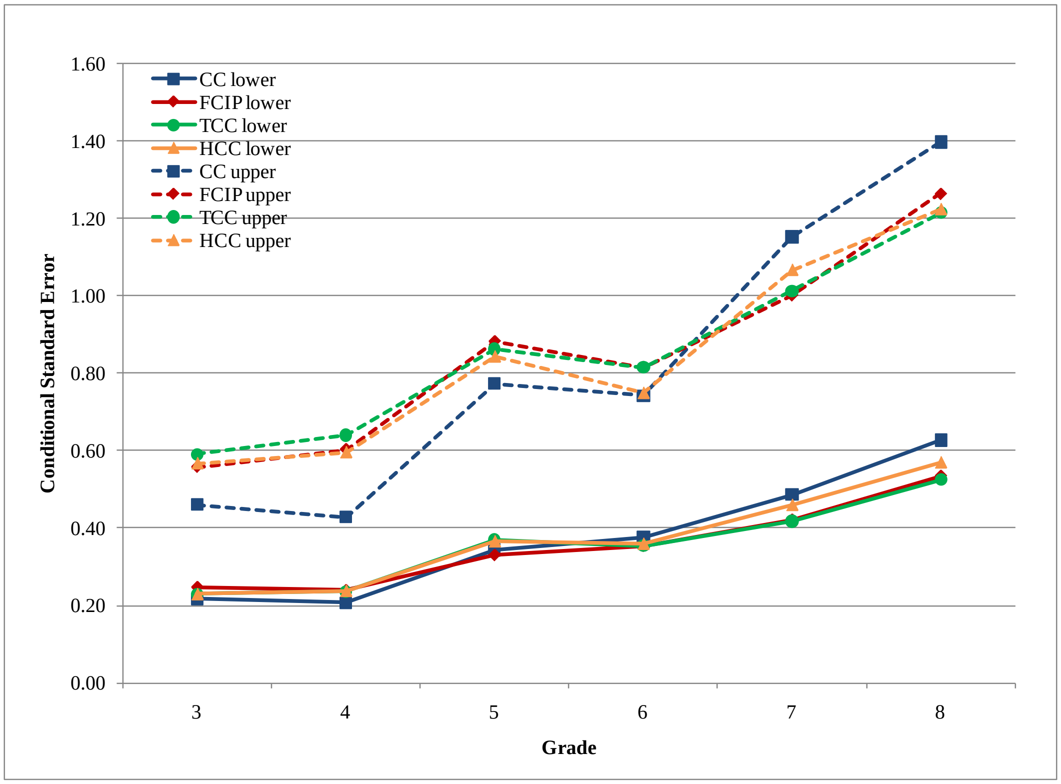

Figure 17. Conditional Standard Error for Reading, skewed distribution, 1,500 sample size

As shown in Figure 17, there was close agreement among the conditional standard errors at the lower cut-score across the four lower grade levels, with slight differences at grades 7 and 8, where the conditional standard error values for the CC procedure were the largest followed by the HCC procedure. The conditional standard errors for the TCC and FCIP procedures were similar.

The conditional standard error values for the upper cut-score were larger than those for the lower cut-score and there were larger differences among the procedures, primarily for the CC procedure. The CC procedure had the lowest values for the three lower grade levels and the largest values for the two upper grade levels. The remaining three procedures were closer in value, with a change in order across grades. For example, the conditional standard errors for the TCC procedure were slightly larger than the conditional standard errors for the HCC and FCIP procedures for grades 3 and 4, while the conditional standard errors for the HCC procedure were smaller than the other scaling procedures for grade 7. But these differences were not as large for the HCC procedure as those found for the CC procedure.

The patterns of conditional standard errors for the skewed distribution as compared to the normal distribution were slightly different. There was more variability among the conditional standard errors for the upper cut-score. Also, the values for the skewed distribution were larger than those found in the normal distribution condition. However, the values for the lower cut-score were not affected as much as the upper cut-score.

Skewed distribution 3,000 sample size

The mean conditional standard error values are displayed in Figure 18 for a skewed distribution and a sample size of 3,000. The conditional standard error values are shown in Table B10 in Appendix B.

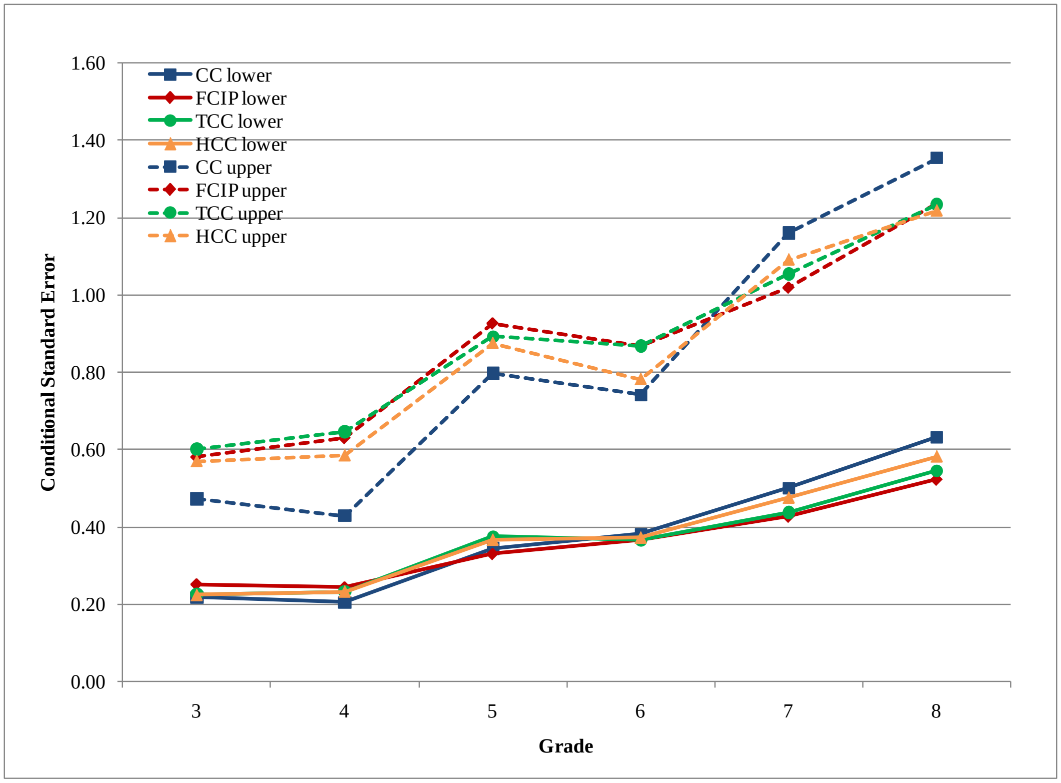

Figure 18. Conditional Standard Error for Reading, skewed distribution, 3,000 sample size

As shown in Figure 18, there was close agreement among the conditional standard errors between the four scaling procedures for the lower cut-score for grades 3 through 6, and generally increased across the six grade levels. Differences were found for the two upper grade levels. Lower values were found for the TCC and FCIP procedures, and slightly higher values were found for the FCIP and HCC procedures. Although the TCC procedure had slightly higher values for the lowest grade, these values were not too different from the values for the other scaling procedures.

The conditional standard error values for the upper cut-score were larger than those for the lower cut-score and there were larger differences among the procedures, especially for the CC procedure. The CC procedure had the lowest values for the two lower grade levels and the largest values for the two upper grade levels. In contrast, the TCC procedure had the reverse pattern, where the lowest values were found for the upper grade levels and the largest values for the lower grade levels. The most consistent procedures were HCC and FCIP, where the conditional standard error values were in between the other scaling procedures. The pattern of results was consistent with those found in the skewed distribution 1,500 sample size as well as the normal distribution conditions.

Summary for Conditional Standard Error

The conditional standard errors were relatively consistent across vertical scaling procedures for the normal distribution condition. This consistency was found more so for the lower cut-score than the upper cut-score values. There was an increase in the conditional standard errors as grade increased. When the skewed distribution condition was introduced, the conditional standard errors for the upper cut-score differed across vertical scaling procedures more so than for the normal distribution conditions, the CC procedure differed the most from the other procedures. The values of the conditional standard error values decreased slightly when the sample size increased.

Root-Mean-Squared-Difference

Normal distribution 1,500 sample size

The root-mean-squared-differences for the pairs of the four vertical scaling procedures for the normal distribution, 1,500 sample size condition are presented in Table 8. As shown, there are six RMSDs among the four scaling procedures for each grade. Similar to the conditional standard error results there is no metric to indicate which values are too high.

Table 8.

Root-Mean-Squared-Difference for Vertical Scaling procedures for Reading, normal distribution, 1,500 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	9.121	17.249	13.182	10.316	6.615	5.268
4	8.537	14.170	11.933	7.569	5.830	2.626
5	4.997	5.563	4.048	4.478	2.346	3.061
6	4.657	4.657	4.019	0.000	1.088	1.088
7	8.939	9.300	7.146	2.536	3.525	3.161
8	9.529	11.405	7.892	9.749	3.067	9.093

The results reported in Table 8 indicate that the agreement between pairs of scale scores can be clustered into two groups. The first cluster, which had lower RMSDs across grade levels, includes the following pairs: FCIP/TCC, FCIP/HCC, and TCC/HCC. The second cluster which had larger RMSDs across grade levels includes the following pairs: CC/FCIP, CC/TCC, and CC/HCC. In the first cluster, the RMSDs ranged between 0.00 and 7.57 with three exceptions – grade 3, FCIP/TCC (10.32), grade 8, FCIP/TCC (9.75), and TCC/HCC (9.09); the greatest agreements were for the FCIP and HCC procedure pairs. For the second cluster eight of the mean RMSDs exceeded 9.0 and the remainder varied between 4.02 and 8.94. The CC was one of the members of each pair in the second cluster. Larger RMSD values were found for the two lowest and two upper grade levels.

Figure 19 shows the RMSD results for the same information provided in Table 8. The grade levels are shown on the X-axis. The value of the RMSD is shown on the Y-axis. The differences between the CC/FCIP pair is shown in blue (square markers), CC/TCC is red (diamond markers), CC/HCC is green (circle markers), FCIP/TCC is purple (triangle markers), FCIP/HCC is yellow (asterisk markers), and TCC/HCC is orange (cross markers).

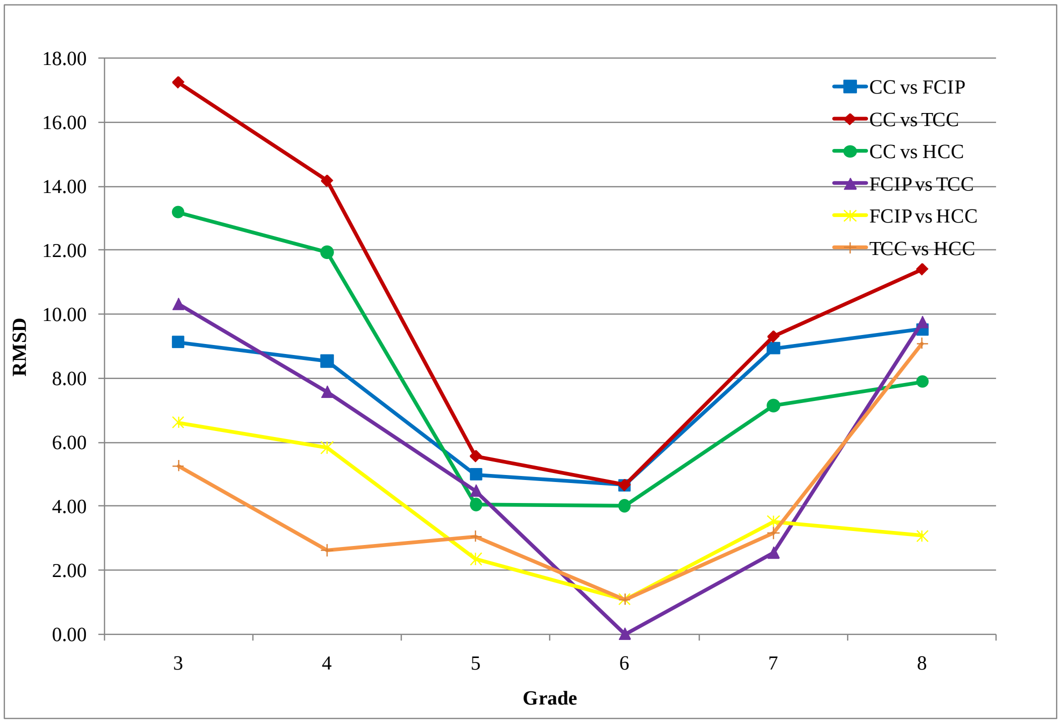

Figure 19. Root-Mean-Squared-Difference for Reading, normal distribution, 1,500 sample size

As shown in Figure 19, the graphs are somewhat parabolic in shape, with low values at grades 5 and 6, and increasing values on both sides. The difference between pairs of vertical scaling procedures was greater at grades 3 and 4 than at grades 7 and 8. The three pairs with the CC procedure tended to have larger RMSDs than the pairs in which the CC procedure was not a member. The largest differences were for the CC/TCC pair. Similar but smaller differences were found for the CC/HCC pair. Smaller differences were found for the FCIP/HCC and CC/FCIP pairs. Smaller differences were found for the TCC/HCC pair for most grade levels, but a higher difference was found for grade 8. Less extreme differences were found when the FCIP and HCC procedures were one of the pairs with other procedures.

Normal distribution 3,000 sample size

Figure 20 shows the RMSD results for the normal distribution 3,000 sample size. The values of the RMSD are provided in Table B11 in Appendix B.

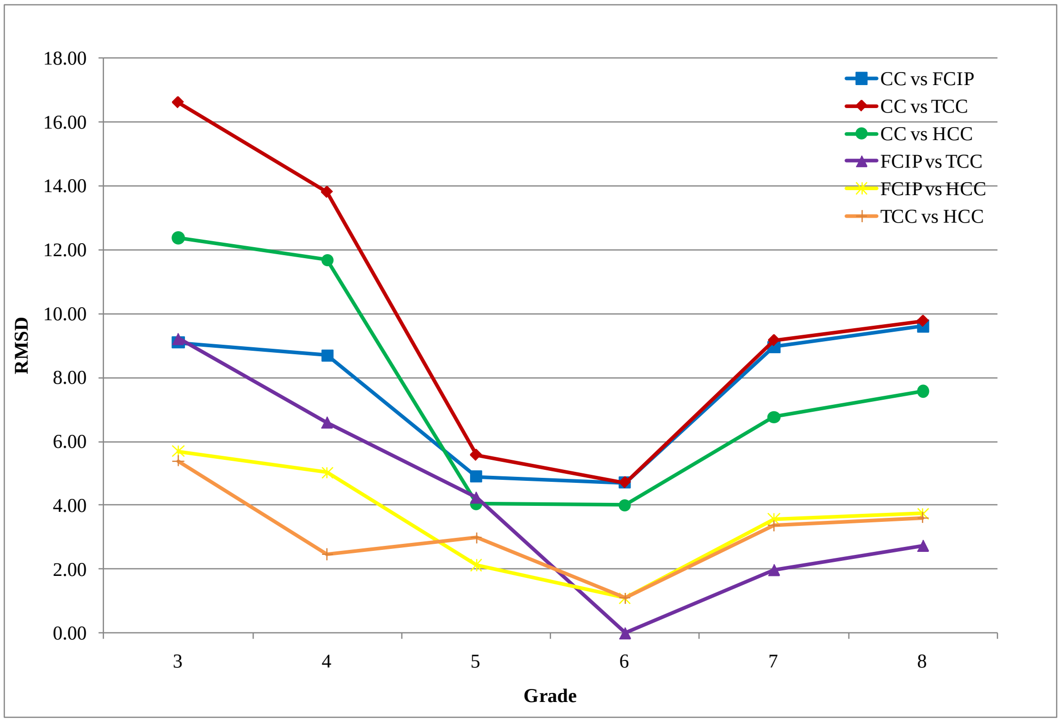

Figure 20. Root-Mean-Squared-Difference for Reading, normal distribution, 3,000 sample size

The RMSDs of the normal distribution 3,000 sample size condition patterns were very similar to those found for the normal distribution 1,500 sample size condition patterns. However, the RMSD values were smaller for the larger sample size condition in some instances. As shown in Figure 20, the graphs are somewhat parabolic in shape, with low values at grades 5 and 6, and increasing values for both the two upper and lower grades. The largest differences among the four scaling procedures were found at the two lowest grade levels particularly for grade 3. Pairs that included the CC procedure tended to have larger RMSDs than the pairs in which the CC procedure was not a member. The largest differences between procedures occurred for the CC/TCC pair followed by the differences for the CC/HCC pair. Smaller differences were found for the FCIP/HCC and the CC/FCIP pairs. However, while smaller differences were found for the TCC/HCC pair for most grade levels, slightly different values were found for grade 8. The least extreme differences were found for the FCIP/HCC pair.

Skewed distribution 1,500 sample size

Figure 21 shows the RMSD results for skewed distribution 1,500 sample size. The values of the RMSD are provided Table B12 in Appendix B.

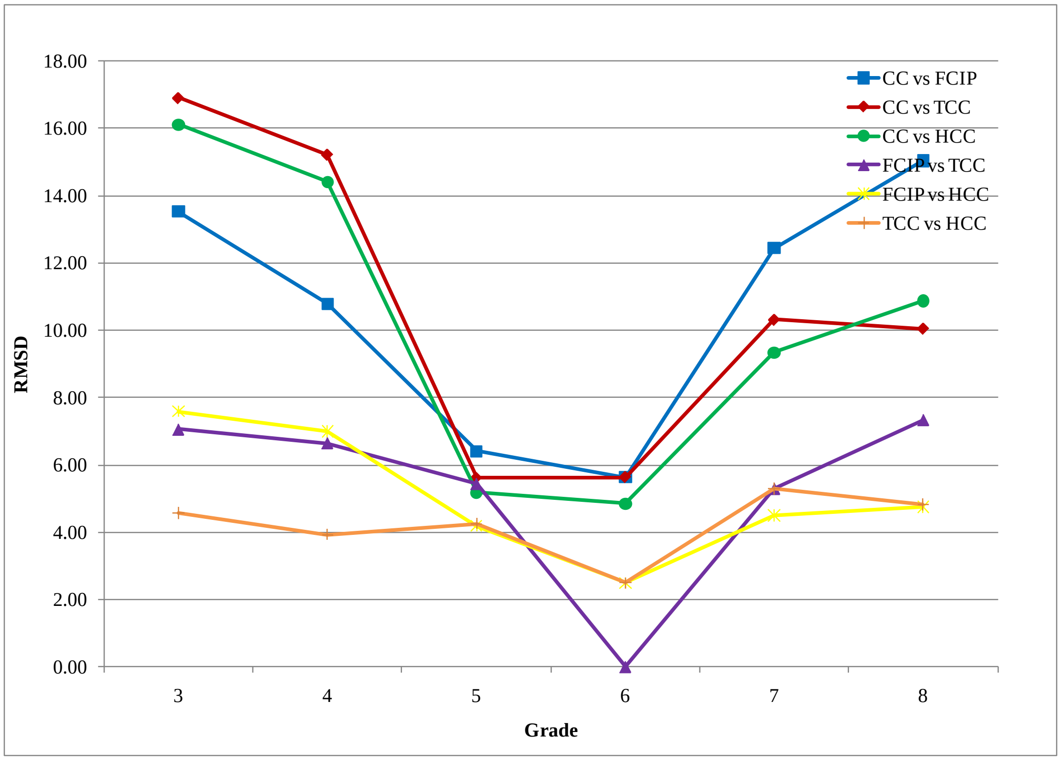

Figure 21. Root-Mean-Squared-Difference for Reading, skewed distribution, 1,500 sample size

As shown in *Figure 21*, the graphs are somewhat parabolic in shape, with low values at grades 5 and 6, and increasing values on both sides. The largest differences between procedures occurred when the CC procedure was paired with each of the FCIP, TCC and HCC procedures, particularly at grades 3 and 4 and grades 7 and 8. The differences for the CC/TCC pair increased as the grade levels increased or decreased from the base grade of 6 with a similar pattern to the CC/HCC pair. Smaller differences were found for the FCIP/HCC and TCC/HCC pairs. While smaller differences were found for the FCIP/HCC pair for most grade levels, a larger difference was found for grade 7. The less extreme differences were found for the TCC/HCC pair, and there was no difference for the FCIP/TCC pair at grade 6.

The results of the skewed distribution 1,500 sample size condition patterns were different than those found for the normal distribution sample size condition with a sample size of 1,500 (cf. Figure 19 see page 113 and Figure 21). The pairs that had the largest RMSDs were the same as those found in the normal distribution conditions. But the CC/FCIP, FCIP/TCC, FCIP/HCC RMSDs pairs were either higher or lower depending on the grade level. However, the CC procedure still had larger differences when paired with the other vertical scaling procedures.

Skewed distribution 3,000 sample size

Figure 22 shows the RMSD results for skewed distribution 3,000 sample size. The RMSD values are shown in Table B13 in Appendix B.

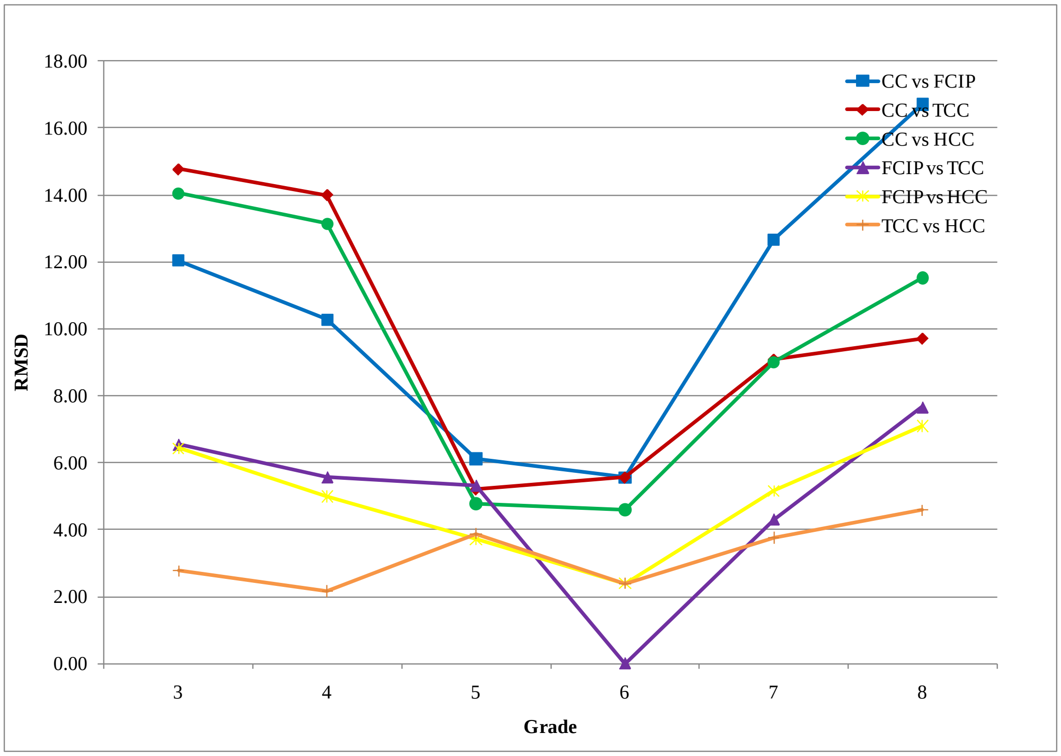

Figure 22. Root-Mean-Squared-Difference for Reading, skewed distribution, 3,000 sample size

As shown in Figure 22, the graphs are somewhat parabolic in shape. The largest differences between procedures occurred for the CC/FCIP pair for grades 5 through 8. However, the CC/TCC pair had larger differences for grades 3 and 4. Similar but smaller differences were found for the CC/HCC pair. The differences for the CC/HCC procedures increased as the grades increased or decreased from the base grade of 6 with a similar pattern to the CC/TCC pair. Smaller differences were found for the FCIP/HCC and TCC/HCC pairs. However, while smaller differences were found for the FCIP/HCC pair for most grade levels, a larger difference was found for grade 5. Less extreme differences were found for the TCC/HCC pair.

Comparison of Figure 21 and Figure 22 indicates that sample size had little effect for the skewed distributions. The RMSD values for the lower grades were smaller for this larger sample size condition, and higher for the upper grade levels. These results were similar to those found in the skewed distribution 1,500 sample size condition.

Summary for Root-Mean-Squared-Difference

The RMSDs for the six pairs of vertical scaling procedures showed a lot of variability especially at the lower and upper grades. The variability among RMSDs was somewhat smaller for the larger sample size conditions. This indicates that the scale scores were sometimes quite different between vertical scaling procedures. The patterns were similar for the normal distribution for both sample size conditions. While there were differences between the normal and skewed distributions, the RMSD values were different depending on the pair members. The RMSDs were larger when the CC procedure was paired with the other three scaling procedures. However, since these results are aggregated over all scale scores it is difficult to determine how large the differences were for the scale scores at the cut-scores. That is, the largest differences may not have occurred near either cut-score, so it may be there was no impact on decision accuracy and consistency.

Correlations between Item Parameters

Normal distribution 1,500 sample size

The mean correlations between the pairs of *a*-, *b*-, and *c*-parameter estimates obtained from the four scaling procedures are presented for the normal distribution 1,500 sample size condition in Table 9. The results are presented for all six sets of correlations of the item parameters for each grade. While correlations of 0.50 are considered moderate in value (Glass & Hopkins, 1996), values less than 0.50 are bolded since the item parameter correlations were expected to be higher.

Table 9.

Correlations of item parameters for Vertical Scaling for Reading, normal distribution, 1,500 sample size

			a			
Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	0.935	0.918	0.999	0.839	0.935	0.920
4	0.912	0.813	0.811	0.624	0.625	0.996
5	0.948	0.935	1.000	0.831	0.950	0.932
6	0.884	0.884	0.877	1.000	0.995	0.995
7	0.896	0.888	0.999	0.998	0.896	0.890
8	0.997	0.981	0.998	0.978	0.992	0.991
			b			
3	0.963	0.915	0.996	0.866	0.961	0.935
4	0.982	0.978	0.977	0.944	0.941	0.999
5	0.989	0.987	1.000	0.970	0.989	0.987
6	0.991	0.991	0.993	1.000	0.999	0.999
7	0.989	0.988	1.000	1.000	0.990	0.989
8	0.999	0.997	1.000	0.998	0.999	0.998
			c			
3	0.705	0.934	0.994	0.613	0.688	0.931
4	0.553	0.847	0.842	**0.451**	**0.369**	0.964
5	0.511	0.797	0.997	**0.269**	**0.496**	0.795
6	0.891	0.891	0.898	1.000	0.970	0.970
7	0.864	0.871	0.998	0.998	0.868	0.876
8	0.918	0.919	0.994	0.962	0.890	0.936

As shown in Table 9, the pattern of mean correlations for the *b*-parameter estimates is less complex than the pattern for the *a*-parameter estimates which is less complex than the pattern for the *c*-parameter estimates. For the *a*-parameter mean correlations, with the exception of FCIP/TCC and FCIP/HCC at grade 4, 11 were between 0.80 and 0.90, while the rest were at least 0.90. Seven of the 11 mean correlations between 0.80 and 0.90 involved CC (two at grade 4, three at grade 6, and two at grade 7).

The mean correlations for the *b*-parameter estimates were above 0.90, with one exception, the mean correlation for the FCIP/TCC pair for grade 3 which was 0.87. All of the correlations were high for all scaling pairs.

In the case of the *c*-parameter the mean correlations can be divided into two sets. The first set includes the pairs CC/TCC, CC/HCC, and TCC/HCC, for which the correlations were all at least 0.80 across grade levels. The second set includes CC/FCIP, FCIP/TCC and FCIP/HCC; the mean correlations for these pairs of procedures were less than 0.72 for the three lower grades and above 0.86 for the three upper grades. Further, the mean correlations for the *c*-parameter estimates for the scaling procedures were lower than the mean correlations for the corresponding *a*- and *b*-parameter estimates. The lower mean correlations for the *c*-parameters were likely due to fixing the parameters for the on-level test forms for the lower grades. The fixing of the *c*-parameters did not affect the upper grade levels since the fixed parameters were for the off-level test forms for those grade levels. Since the FCIP procedure results in fixed parameters from the upper grade level, these results may be expected as they are not estimated independently of the grade level.

Normal distribution 3,000 sample size

The mean correlations between pairs of *a*-, *b*-, and *c*-parameter estimates obtained from the different scaling procedures are presented for the normal distribution 3,000 sample size condition in Table 10.

Table 10.

Correlations of item parameters for Vertical Scaling for Reading, normal distribution, 3,000 sample size

				a		
Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	0.948	0.935	0.999	0.885	0.949	0.936
4	0.929	0.845	0.835	0.690	0.688	0.997
5	0.957	0.940	1.000	0.857	0.959	0.938
6	0.887	0.887	0.882	1.000	0.996	0.996
7	0.900	0.895	0.999	0.999	0.899	0.896
8	0.997	0.986	0.998	0.982	0.992	0.993
				b		
3	0.965	0.918	0.996	0.866	0.959	0.938
4	0.984	0.979	0.975	0.950	0.944	0.999
5	0.990	0.987	1.000	0.974	0.990	0.988
6	0.991	0.991	0.992	1.000	1.000	1.000
7	0.989	0.988	1.000	1.000	0.989	0.988
8	0.999	0.997	1.000	0.998	0.998	0.998
				c		
3	0.707	0.935	0.990	0.625	0.687	0.930
4	0.507	0.834	0.822	**0.478**	**0.390**	0.960
5	0.518	0.769	0.996	**0.305**	**0.497**	0.770
6	0.888	0.888	0.887	1.000	0.972	0.972
7	0.855	0.860	0.998	0.998	0.862	0.867
8	0.932	0.918	0.996	0.972	0.913	0.937

Similar to the normal distribution 1,500 sample size condition, the pattern of mean correlations for the *b*-parameter estimates is less complex than the pattern for the *a*-parameter estimates which is less complex than the pattern for the *c*-parameter estimates. For the *a*-parameter mean correlations, with the exception of grade 4, FCIP/TCC, and FCIP/HCC, 10 were between 0.80 and 0.90, and the rest were at least 0.90. Six of the 10 mean correlations between 0.80 and 0.90 involved CC (two at grade 4, three at grade 6, and one at grade 7).

The mean correlations for the *b*-parameter estimates were above 0.90, with one exception (FCIP/TCC grade 3 0.87). All of the correlations were high for all scaling pairs.

For the *c*-parameter the mean correlations again can be divided into two sets. The first set included the pairs involved CC/TCC, CC/HCC, and TCC/HCC; the mean correlations for these pairs of scaling procedures were all at least 0.80 across the grades. The second set includes CC/FCIP, FCIP/TCC and FCIP/HCC; the mean correlations involving these procedures were less than 0.71 for the three lower grades, and above 0.86 for the three upper grades. Further, the mean correlations among the *c*-parameter estimates for the scaling procedures were lower than the mean correlations for the corresponding *a*- and *b*-parameter estimates, likely due to the same reasons provided for the normal distribution 1,500 sample size condition. The lower mean correlations for the *c*-parameters were likely due to fixing the parameters for the on-level test forms for the lower grades. The fixing of the *c*-parameters did not affect the upper grade levels since the fixed parameters were for the off-level test forms for those grade levels.

Skewed distribution 1,500 sample size

The mean correlations between pairs of *a*-, *b*-, and *c*-parameter estimates obtained from the different scaling procedures are presented for the skewed distribution 1,500 sample size condition in Table 11.

Table 11.

Correlations of item parameters for Vertical Scaling for Reading, skewed distribution, 1,500 sample size

			a			
Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	0.936	0.917	0.999	0.898	0.943	0.927
4	0.834	0.653	0.756	0.684	0.694	0.962
5	0.928	0.915	0.999	0.834	0.935	0.924
6	0.892	0.892	0.890	1.000	0.967	0.967
7	0.900	0.901	0.998	0.999	0.903	0.905
8	0.985	0.979	0.999	0.986	0.987	0.987
			b			
3	0.945	0.948	0.998	0.860	0.941	0.959
4	0.980	0.966	0.978	0.946	0.950	0.996
5	0.985	0.990	1.000	0.969	0.986	0.990
6	0.987	0.987	0.992	1.000	0.996	0.996
7	0.991	0.991	1.000	1.000	0.990	0.990
8	0.998	0.996	1.000	0.996	0.998	0.997
			c			
3	0.656	0.937	0.996	0.603	0.641	0.932
4	0.516	0.821	0.841	**0.385**	**0.285**	0.953
5	**0.457**	0.723	0.993	**0.248**	**0.454**	0.716
6	0.879	0.879	0.913	1.000	0.937	0.937
7	0.767	0.764	0.996	0.998	0.783	0.782
8	0.820	0.856	0.997	0.959	0.810	0.886

Similar to the normal distribution conditions, the pattern of mean correlations for the *b*-parameter estimates is less complex than the pattern for the *a*-parameter estimates which is less complex than the pattern for the *c*-parameter estimates. For the *a*-parameter mean correlations, with the exception of grade 4, CC/TCC (0.65), CC/HCC (0.76), FCIP/TCC (0.68), and FCIP/HCC (0.69) procedures at grade 4, six were between 0.80 and 0.90, and the rest were at least 0.90. The four correlations below 0.80 were close to 0.70 or above. Four of the six mean correlations between 0.80 and 0.90 involved CC (one at grade 4, and three at grade 6).

The mean correlations for the *b*-parameter estimates were above 0.90, with the same exception, the mean correlation between the FCIP/TCC procedures for grade 3. All of the correlations were high for all scaling pairs.

For the *c*-parameter the mean correlation can be divided into two sets. The first set included the pairs CC/TCC, CC/HCC, and TCC/HCC; the mean correlations for the pairs of scaling procedures were all at least 0.80 across the grades. The second set includes CC/FCIP, FCIP/TCC and FCIP/HCC; the mean correlations for the pairs of scaling procedures were less than 0.66 for the three lower grades and above 0.82 for the three upper grades. Further, the mean correlations among the *c*-parameter estimates for the scaling procedures were lower than the mean correlations for the corresponding *a*- and *b*-parameter estimates. Like the normal distributions, the lower mean correlations for the *c*-parameters were likely due to fixing the parameters for the on-level test forms for the lower grades.

Skewed distribution 3,000 sample size

The mean correlations between pairs of *a*-, *b*-, and *c*-parameter estimates obtained from the different scaling procedures are presented for the skewed distribution 3,000 sample size condition in Table 12.

Table 12.

Correlations of item parameters for Vertical Scaling for Reading, skewed distribution, 3,000 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
			a			
3	0.945	0.925	0.998	0.928	0.954	0.937
4	0.827	0.647	0.774	0.737	0.742	0.952
5	0.942	0.922	0.999	0.867	0.950	0.933
6	0.895	0.895	0.908	1.000	0.976	0.976
7	0.907	0.907	0.997	0.999	0.905	0.908
8	0.976	0.963	0.995	0.980	0.979	0.973
			b			
3	0.945	0.945	0.996	0.865	0.938	0.961
4	0.979	0.961	0.974	0.947	0.950	0.994
5	0.984	0.990	0.999	0.973	0.987	0.991
6	0.983	0.983	0.992	1.000	0.996	0.996
7	0.992	0.992	0.999	1.000	0.988	0.988
8	0.996	0.997	1.000	0.998	0.995	0.997
			c			
3	0.655	0.932	0.989	0.612	0.630	0.926
4	**0.473**	0.779	0.801	**0.409**	**0.288**	0.944
5	**0.407**	0.670	0.990	**0.272**	**0.409**	0.668
6	0.847	0.847	0.893	1.000	0.919	0.919
7	0.716	0.705	0.992	0.998	0.742	0.733
8	0.792	0.788	0.993	0.940	0.794	0.814

Similar to the previous conditions, the pattern of mean correlations for the *b*-parameter estimates is less complex than the pattern for the *a*-parameter estimates which is less complex than the pattern for the *c*-parameter estimates. For the *a*-parameter mean correlations, with the exception of grade 4, CC/TCC (0.65), CC/HCC (0.77), FCIP/TCC (0.74), and FCIP/HCC (0.74), four were between 0.80 and 0.90, and the rest were at least 0.90. The four correlations below 0.80 were close to 0.70 or above. Three of the four mean correlations between 0.80 and 0.90 involved CC (one at grade 4, and two at grade 6).

The mean correlations for the *b*-parameter estimates were above 0.94, with the exception, the mean correlation for the FCIP/TCC pair for grade 3 (0.86). All of the correlations were high for all scaling pairs.

For the *c*-parameter the mean correlation can again be divided into two sets. The first set included the pairs involved CC/TCC, CC/HCC, and TCC/HCC; the mean correlations for these pairs of scaling procedures were all at least 0.80 across the grades. The second set includes CC/FCIP, FCIP/TCC and FCIP/HCC; the mean correlations for these pairs of scaling procedures were less than 0.66 for the three lower grades, and above 0.72 for the three upper grades. Further, the mean correlations among the *c*-parameter estimates for the scaling procedures were lower than the mean correlations for the corresponding *a*- and *b*-parameter estimates. The lower mean correlations for the *c*-parameters were likely due to fixing the parameters for the on-level test forms for the lower grades.

Summary for Correlations

The correlations between the six pairs of scaling procedures were not affected by the distribution shape or sample size. The highest correlations were found for the *b*-parameter, where all were greater than 0.90 except for FCIP/TCC. For the correlations for the *a*-parameter the majority of the pairs were above 0.80, with the exception of four pairs at grade 4. Lastly, the correlations for the *c*-parameter were generally quite high, with some exceptions when FCIP was a member. The smallest *c*-parameter correlations occurred for pairs that included FCIP with the lower grade levels (3, 4, and 5). This was most likely due to the lower grades having the off-level item parameters fixed. The item parameters for the *a*-, *b*-, and *c*- parameters did not differ based on the distribution type or sample size. There was no consistent pattern of which pair had the highest values across the *a*-, *b*-, and *c*-parameters, but generally the values were high with the few exceptions presented above. These results indicate similar ranking for the vertical scaling procedure item parameters.

CHAPTER 5 RESULTS: MATHEMATICS

The results for Mathematics are presented in this chapter. The results are presented for each evaluation measure for the four distribution shape and sample size conditions. The decision accuracy and consistency results are presented first followed by presentation of the conditional standard errors of estimation at the cut-scores. Third, the root-mean-squared-differences of the scale scores are presented. Lastly, the correlations between the item parameters across vertical scaling procedures are presented. The results presented in Chapters 4 and 5 provide a micro evaluation of the differences. Chapter 6 presents a summary of the micro evaluation, and a macro discussion and applications for practitioners.

The presentation of the results for the first condition – normal distribution with a sample size of 1,500 – includes a summary table and a graphical representation of the results for decision accuracy, decision consistency, conditional standard error estimates, and RMSD and a summary table for the correlations of the item parameter estimates. The remaining conditions only include a graphical representation for decision accuracy, decision consistency, conditional standard error estimates and RMSD and a summary table for the correlations of the item parameters estimates. The remaining tables appear in Appendix B.

Decision Consistency and Accuracy

Normal distribution 1,500 sample size

The decision accuracy and consistency results are presented for the normal distribution with 1,500 examinees in Table 13. The results for all four vertical scaling methods are presented for each grade and cut-score. As indicated earlier, there are three performance levels – *below basic*, *basic,* and *above basic*. For all tables the cut-score between *below basic* and *basic* is listed as the lower cut-score, and the cut-score between *basic* and *above basic* is listed as the upper cut-score. The decision accuracy and consistency values are bounded by zero and one, where, for example, a value of 0.80 is interpreted as 80 % accurate or consistent. However no specific minimum criteria have been identified for decision accuracy and consistency in the literature. Therefore, the minimum value for the present study was set at 0.80 as 0.80 represents 80% both accurate or consistent classification and the values below 0.80 are bolded in the tables.

Table 13.

Decision Accuracy and Consistency for Vertical Scaling for Mathematics, normal distribution, 1,500 sample size

Accuracy					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	**0.559**	**0.467**	**0.484**	**0.519**
	upper	0.916	0.934	0.930	0.935
4	lower	0.807	0.829	0.805	0.815
	upper	0.949	0.961	0.960	0.962
5	lower	0.819	0.815	**0.794**	0.802
	upper	0.922	0.931	0.919	0.930
6	lower	**0.672**	**0.684**	**0.684**	**0.680**
	upper	0.927	0.915	0.915	0.919
7	lower	**0.681**	**0.707**	**0.715**	**0.699**
	upper	0.930	0.929	0.917	0.927
8	lower	**0.661**	**0.690**	**0.690**	**0.674**
	upper	0.874	0.907	0.870	0.877
Consistency					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.818	0.815	0.814	0.818
	upper	0.945	0.955	0.953	0.956
4	lower	0.833	0.817	0.818	0.820
	upper	0.954	0.963	0.962	0.964
5	lower	0.884	0.880	0.872	0.878
	upper	0.936	0.944	0.938	0.942
6	lower	0.841	0.845	0.845	0.844
	upper	0.942	0.935	0.935	0.937
7	lower	0.842	0.848	0.850	0.846
	upper	0.941	0.939	0.932	0.938
8	lower	0.886	0.888	0.890	0.887
	upper	0.915	0.930	0.912	0.916

As shown in Table 13, the decision accuracy for the four scaling procedures exceeded 0.80 for the upper cut-score for all grade levels. However, the pattern for decision accuracy at the lower cut-score was not consistent across grade levels. For example, the decision accuracy was less than 0.80 for all four vertical scaling procedures for grade 3 (from 0.47 to 0.56) and grades 6, 7, and 8 (from 0.66, CC, grade 8 to 0.72, TCC, grade 7). In contrast, the decision accuracy for grade 4 ranged from 0.80 to 0.83.

At the upper cut-score, the decision accuracy across grades and scaling methods varied from 0.87 (grade 8, CC) to 0.96 (grade 4, FCIP, TCC, HCC). There were smaller differences among the four vertical scaling procedures, with all decision accuracy values above 0.80.

The decision consistency values for both cut-scores were all greater than 0.80 for all four scaling procedures and grade levels. However, the decision consistency values at the lower cut-score were lower than the corresponding decision consistency at the upper cut-score (0.81 to 0.89 vs. 0.92 to 0.96) for all four scaling procedures. Given the small range, the decision consistency values across the four scaling procedures behaved similarly across grade levels.

The decision accuracy and consistency results presented in Table 13 are graphically shown, respectively, in Figure 23 and Figure 24. The grade levels are displayed on the X-axis. Decision accuracy and consistency values are shown on the Y axis bounded by 0.40 and 1.00 to show the differences between procedures more easily. Both the lower and upper cut-scores are displayed on the same graph, where the solid line indicates the lower cut-score and the dashed line indicates the upper cut-score. Each vertical scaling procedure is shown in a different colour with a different marker, CC is blue (square marker), FCIP is red (diamond marker), TCC is green (circle marker), and HCC is yellow (triangle marker).

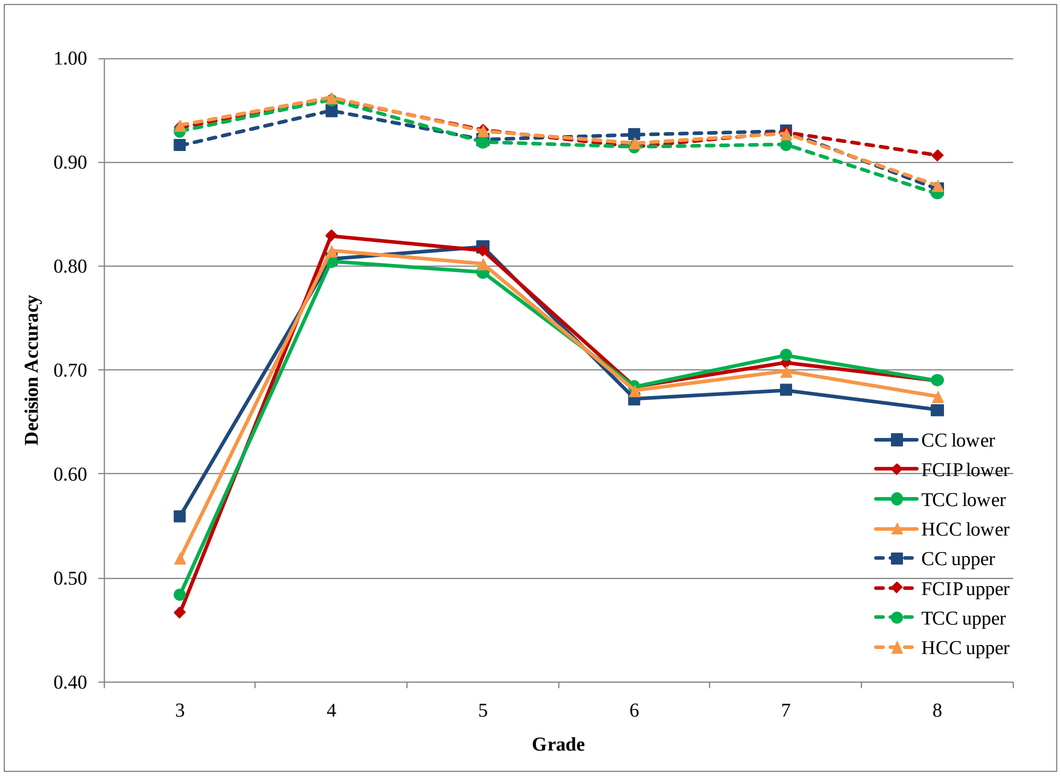

Figure 23. Decision accuracy for Mathematics, normal distribution, 1,500 sample size

As shown in Figure 23, for the lower cut-score larger differences across the grade levels occurred rather than differences across vertical scaling procedures. At the lower cut-score, only at grades 4 and 5 were the values above 0.80, and the smallest values (< 0.56) were found at grade 3. Decision accuracy values, less than 0.80, were found at grade 6, 7 and 8 (approximately 0.70) with smaller differences among the four vertical scaling procedures at the two upper grade levels. The decision accuracy values for the lower cut-score showed some inconsistencies across grade levels. For example, the difference between CC (0.56) and FCIP (0.47) at grade 3 was 0.09. The largest difference of 0.36

occurred across grade levels, which ranged from 0.47 (FCIP grade 3) to 0.83 (FCIP grade 4). Generally, the CC procedure had the largest decision accuracy values, except for the two upper grade levels. For the upper cut-score, similar results across the four vertical scaling procedures were found, where most of the values were greater than 0.90, with the exception of the CC, TCC and HCC procedures for grade 8. Most of the values were quite high and the differences were small.

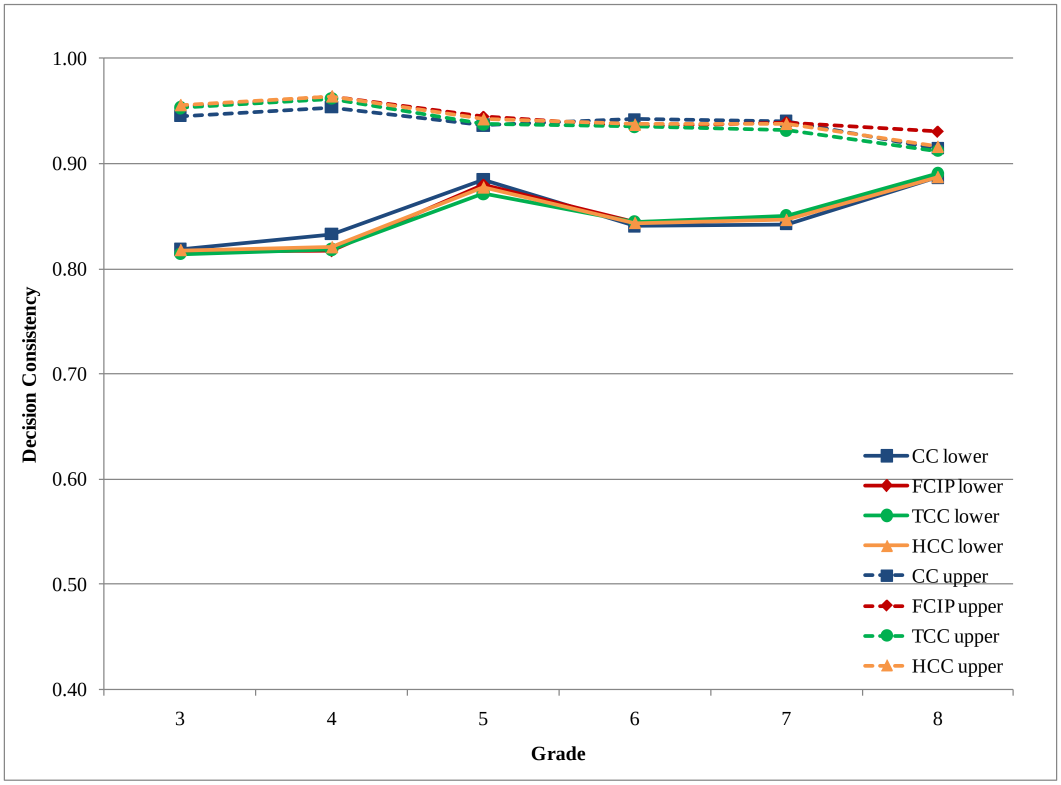

Figure 24. Decision consistency for Mathematics, normal distribution, 1,500 sample size

The decision consistency values shown in Figure 24 were more similar across the four scaling procedures and grades than the values of decision accuracy shown in Figure 23. At the lower cut-score, the CC procedure had the largest values for grades 4 and 5 and the CC procedure had similar decision consistency values with the other procedures for the other grade levels. However, there were few differences across the scaling procedures, and all of the results were above the 0.80 level.

Similar decision consistency values were found for the upper cut-score, where slight differences were found at the two lower grade levels, where the CC, TCC, and HCC procedures had the smallest values, though these differences were really small. However, most of the results showed fairly consistent values all above 0.90 and approached 0.95 at the two lower grades. It appears that for decision consistency the vertical scaling procedure does not impact the values even across grade levels.

Normal distribution 3,000 sample size

The decision accuracy and consistency results are displayed, respectively, in Figures 25 and 26 for the normal distribution and a sample size of 3,000 and the values are provided in Table B14 in Appendix B.

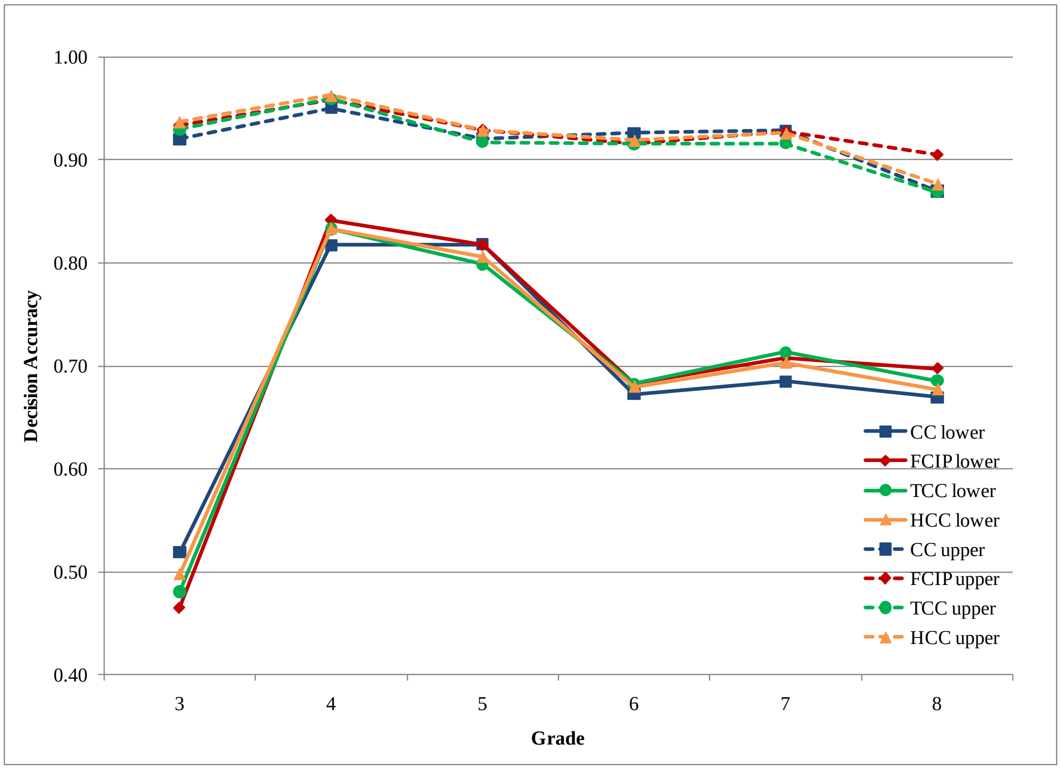

Figure 25. Decision accuracy for Mathematics, normal distribution, 3,000 sample size

The decision accuracy values varied more across the grade levels and occurred more for the lower cut-score than for the upper cut-score. At the lower cut-score, only at grades 4 and 5 were the decision accuracy values above 0.80. The smallest values, which were approximately 0.50 were found at grade 3, while the somewhat larger values, approximately 0.70, were found at grade 6, 7 and 8. Further, there was more variability among the values between the four scaling procedures at grade 3, followed by grades 7 and 8. The larger differences occurred across grade levels ranging from 0.47 (FCIP grade 3) to 0.84 (FCIP grade 4) with a difference of 0.37. Generally, the CC procedure had the largest values, except for the two upper grade levels.

For the upper cut-score, similar results across the four vertical scaling procedures were found, where most of the values were greater than 0.90, with the exception of the CC, TCC, and HCC procedures for grade 8, for which the values were slightly less than 0.90. The differences between the four scaling procedures were small. The decision accuracy and decision consistency results are very similar to the 1,500 sample size condition. Sample size does not seem to have an effect on both decision accuracy and consistency.

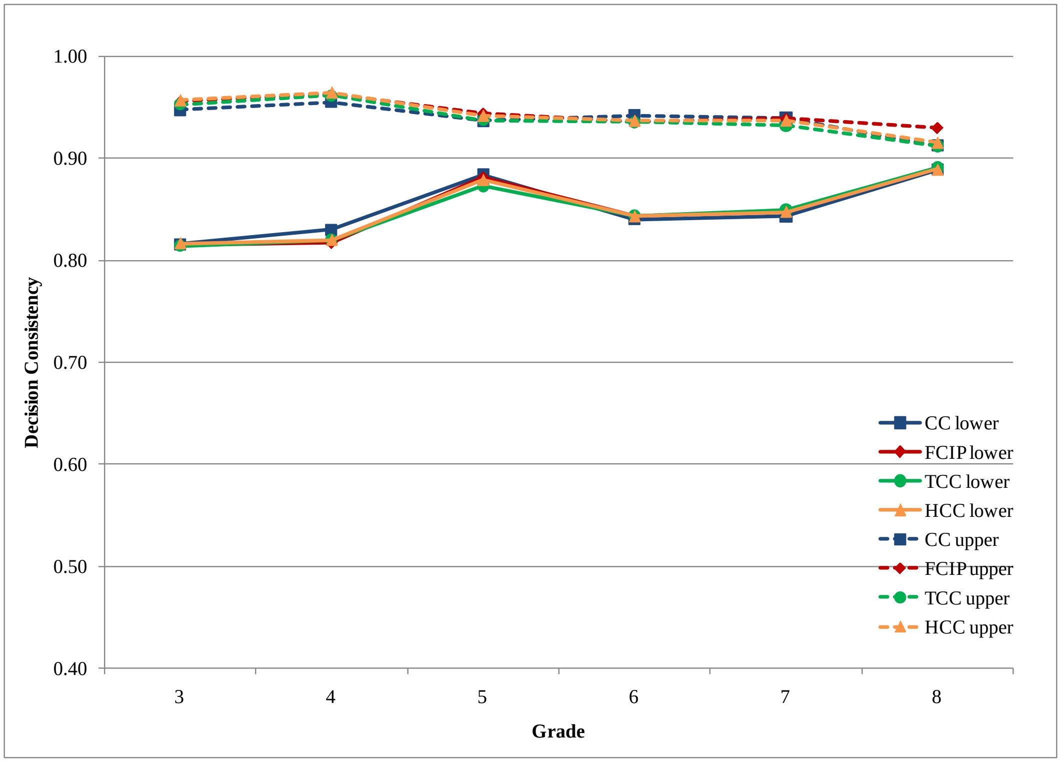

Figure 26. Decision consistency for Mathematics, normal distribution, 3,000 sample size

As shown in Figure 26, the decision consistency values showed similar values across the four vertical scaling procedures for both cut-scores and across grade levels. At the lower cut-score, the decision consistency was above 0.80 at all grade levels for each scaling procedure. Further, differences for the four scaling procedures were with small, except for grade 4 where the CC procedure had the largest values. However, there were few differences across the scaling procedures, and all of the results were above the 0.80 level.

Similar results were found for the upper cut-score where slight differences were found at the two lower grade levels, where the CC, TCC, and HCC procedures had the smallest values, though these differences were really small. However, most of the results showed fairly consistent values all above 0.90 and approached 0.95 at the two lower grades. It appears that for decision consistency the vertical scaling procedure does not impact the values even across grade levels.

Skewed distribution 1,500 sample size

The decision accuracy and consistency results are displayed, respectively, in Figures 27 and 28 for the skewed distribution and a sample size of 1,500. The numerical values are provided in Table B15 in Appendix B.

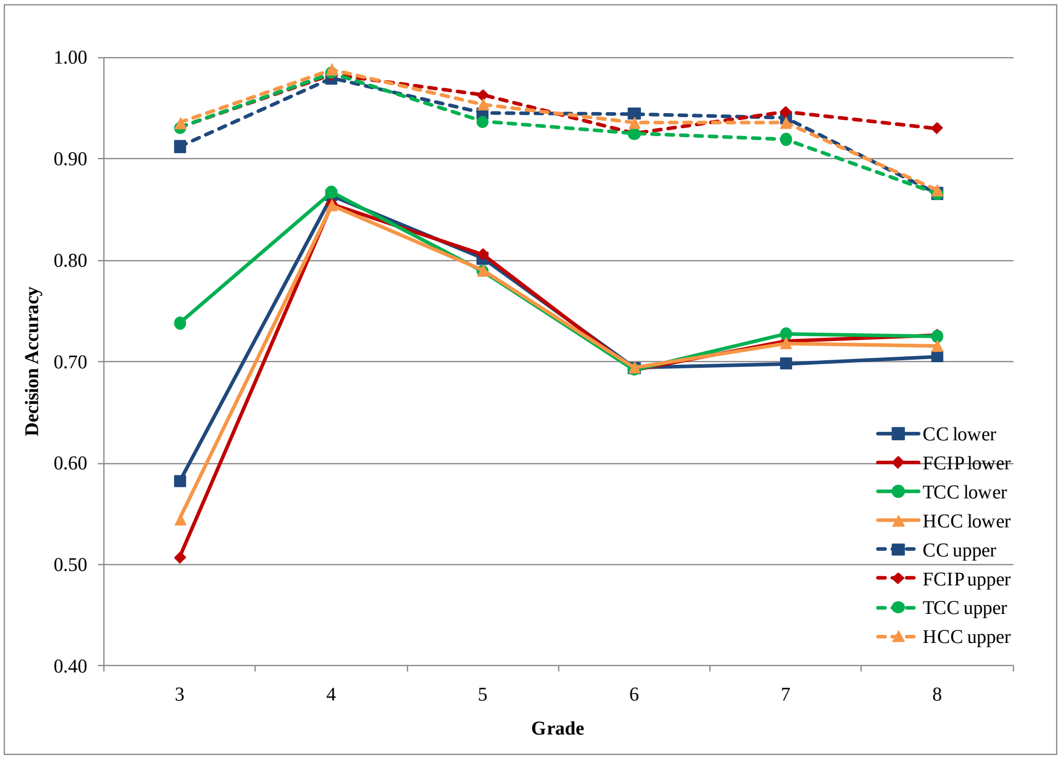

Figure 27. Decision accuracy for Mathematics, skewed distribution, 1,500 sample size

At the lower cut-score, the decision accuracy values at grade 4 (CC, FCIP, TCC, and HCC) and grade 5 (CC and FCIP) were above 0.80. The lowest decision accuracy occurred at grade 3, which ranged from 0.50 (FCIP) to 0.74 (TCC). Decision accuracy values less than 0.80 were found at grades 6, 7 and 8, which ranged from 0.70 to 0.72. The decision accuracy values at the lower cut-score showed some inconsistencies across the four scaling procedures across grade levels, particularly at grade 3. The largest difference between scaling procedures, 0.23, occurred between TCC (0.74) and FCIP (0.51) at grade 3. The larger difference between scaling procedures of 0.36 occurred across grade levels and ranged from 0.51 (FCIP grade 3) to 0.87 (TCC grade 4). However, the decision accuracy values differed in pattern between the skewed distribution as compared to the normal distribution, especially for grade 3.

For the upper cut-score, similar results for the decision accuracy values for the four vertical scaling procedures were found, where most of the values were greater than 0.90, with the exception of the CC, TCC and HCC procedures for grade 8. However, most of the values were quite high and the differences were small for most grade levels, with a larger difference between the FCIP procedure and the CC, TCC and HCC procedures only for grade 8.

Comparison of Figure 23 (see page 136) and Figure 27 indicates that the shape of the score distribution differentially influenced the decision accuracy of the four scaling procedures, more so at the lower cut-score than at the upper cut-score. For example, the decision accuracy values were more variable at grade 3 for the lower cut-score and grade 8 at the upper cut-score.

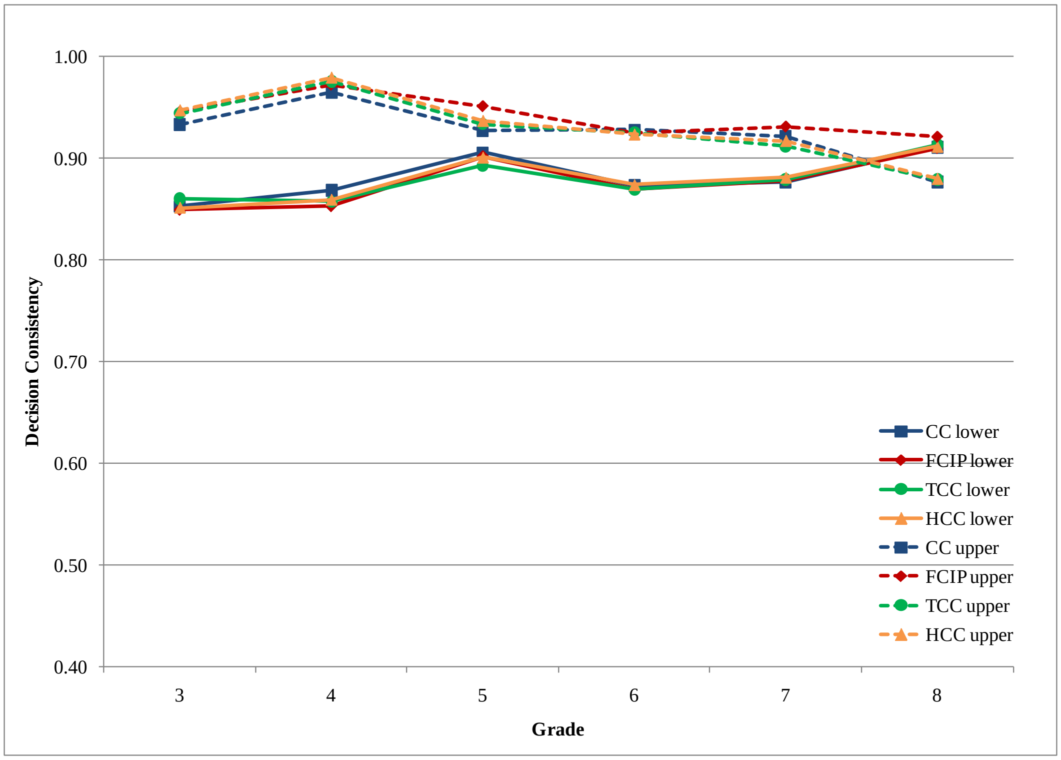

Figure 28. Decision consistency for Mathematics, skewed distribution, 1,500 sample size

The patterns of the decision consistency values shown in Figure 28 are similar to the patterns of the decision consistency values in Figure 27. At grade 4 the CC procedure had the largest values; otherwise the values of decision consistency were generally the same. Similar decision consistency values across vertical scaling procedures were found for the upper cut-score for grades 5 and 8, where the CC, TCC, and HCC procedures had the smallest values. The largest difference occurred at grade 8 between the FCIP and the CC, TCC and HCC procedures. It appears that for decision consistency the vertical scaling procedure does not impact the values even across grade levels. These results were similar to those found for the normal distribution 1,500 sample size condition.

Skewed distribution 3,000 sample size

The decision accuracy and consistency results are displayed, respectively, in Figures 29 and 30 for the skewed distribution and a sample size of 3,000. The numerical values are provided in Table B16 in Appendix B.

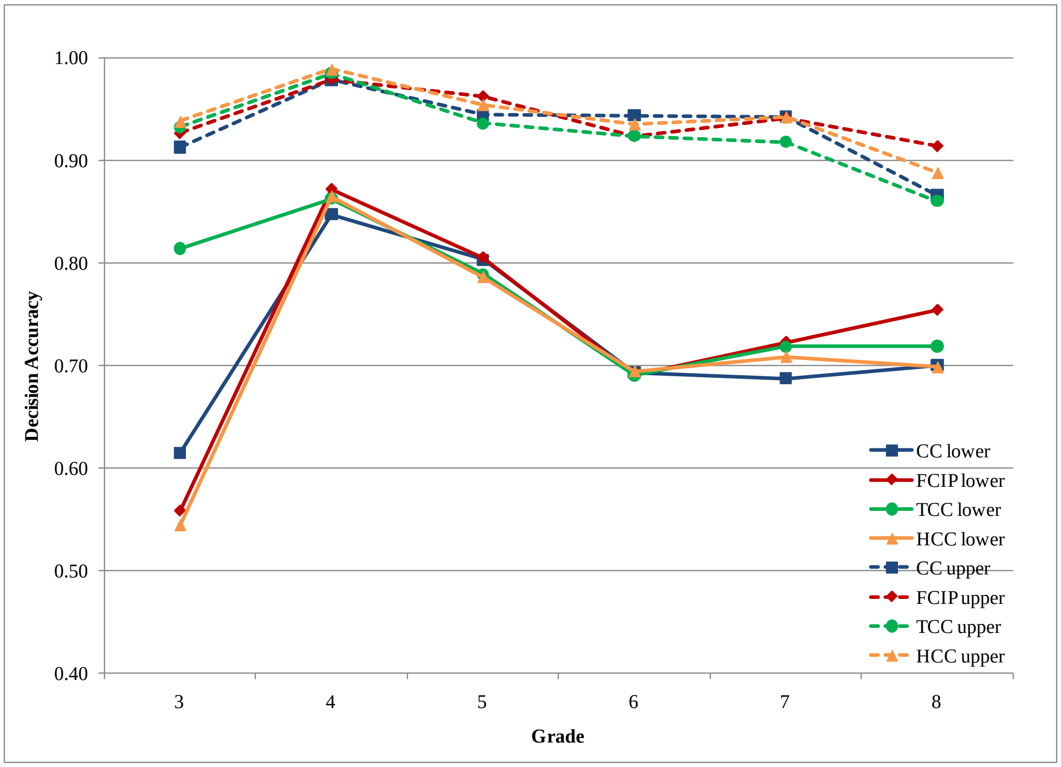

Figure 29. Decision accuracy for Mathematics, skewed distribution, 3,000 sample size

As shown in Figure 29, the lowest decision accuracy values for the lower cut-score occurred at grade 3, which ranged from 0.54 (HCC) to 0.82 (TCC). Lower decision accuracy values were found at grades 6, 7, and 8, which varied between 0.68 and 0.76. The decision accuracy values for the lower cut-score showed some inconsistencies across vertical scaling procedures, particularly with grade 3 with a difference of 0.27 between TCC (0.81) and HCC (0.55). The larger difference of 0.32 occurred across grade levels, which ranged from 0.55 (HCC grade 3) to 0.87 (FCIP and HCC grade 4).

For the upper cut-score, the decision accuracy values were greater than 0.90 for most grade levels, except for grade for the CC, TCC, and HCC procedures. The differences between the four vertical scaling procedures were generally small, with the largest differences between the FCIP procedure and the CC, TCC and HCC procedures at grade 8.

Comparison of Figure 25 (see page 139) and Figure 29 show that the shape of the score distribution did not influence the decision accuracy of the four scaling procedures. For example, the decision accuracy values for the lower cut-score had larger differences at grade 3 and smaller differences at grades 7 and 8, similar to those found in the normal distribution. However, the differences at grade 3 were larger for the skewed distribution than for the normal distribution, in part because the TCC procedure had a higher value for the skewed distribution as compared to the normal distribution.

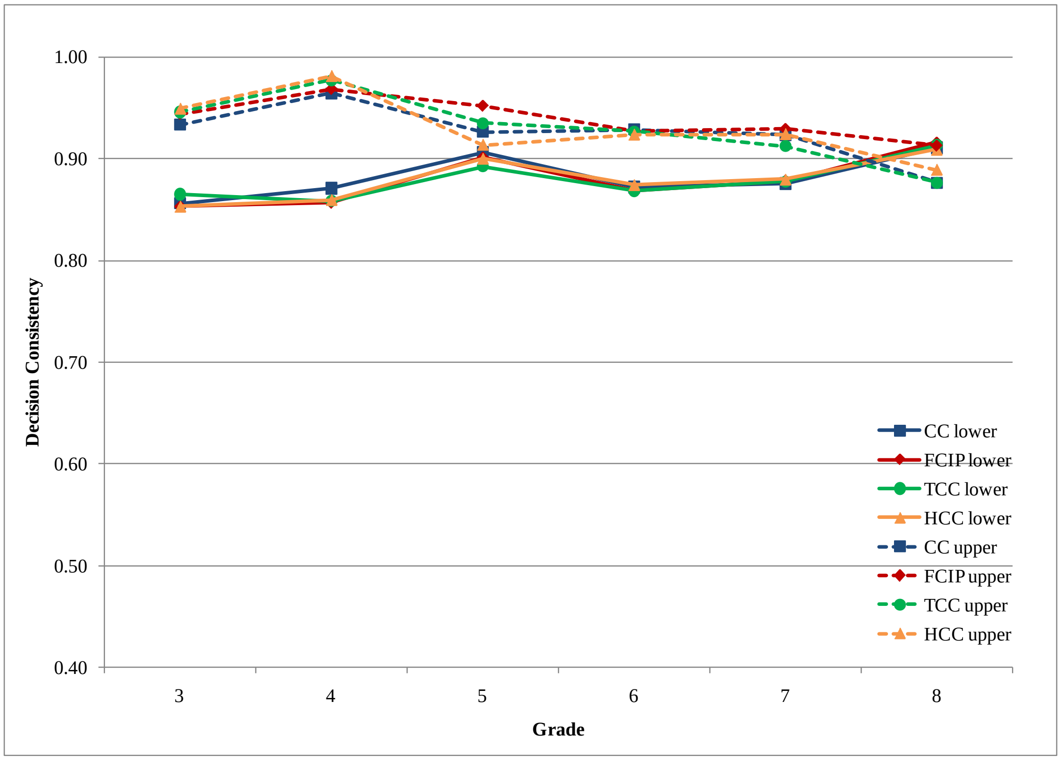

Figure 30. Decision consistency for Mathematics, skewed distribution, 3,000 sample size

The patterns of decision consistency values shown in Figure 30, when compared to the patterns of values in Figure 29, showed that there was greater agreement among vertical scaling procedures and grade levels than for decision accuracy. At the lower cut-score, the CC procedure had the largest values for grade 4 and agreed with the other procedures for the other grade levels. However, there were few differences across the scaling procedures, and all of the results were above the 0.80 level.

Most of the decision consistency values were consistent with values all above 0.90 that approached 0.95 at the two lower grades. The largest difference occurred at grade 8 between the FCIP and the CC, TCC and HCC procedures. It appears that for decision consistency the vertical scaling procedure does not impact the values even across grade levels.

Summary for Decision Accuracy and Decision Consistency

The decision accuracy results indicated a great deal of agreement across the four scaling procedures for both sets of cut-scores for the normal and skewed distributions and sample sizes. For Mathematics, the differences for the normal distributions occurred more across grades than across vertical scaling procedures. While there were some differences across vertical scaling procedures, these differences primarily occurred for grade 3. There was a decrease in decision accuracy as grade increased for the lower cut-score for grade 4 through 8. There were virtually no differences across vertical scaling procedures and grade levels for the upper cut-score. This indicates that the accuracy of the decisions is not consistent for all grade levels. An increase of sample size from 1,500 to 3,000 did not have an impact on the results.

The decision consistency results were more similar for both the upper and lower cut-scores. The values were larger for the upper cut-score, but all of the values exceeded 0.80. There were few differences found across vertical scaling procedure and grade levels. In fact, there were few differences found between the distribution type or sample size as well. Therefore, decision consistency does not seem to be impacted for the different distributions and sample sizes for this Mathematics assessment.

Conditional Standard Error

Normal distribution 1,500 sample size

The values for conditional standard error at the cut-score are reported for the normal distribution 1,500 sample size condition in Table 14 and shown graphically in Figure 31. Since, there is no clear metric to indicate if a conditional standard error is too high no maximum value has been set.

Table 14.

Conditional Standard Error for Mathematics, normal distribution, 1,500 sample size

		1500			
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.285	0.270	0.283	0.258
	upper	0.484	0.495	0.499	0.487
4	lower	0.180	0.188	0.180	0.178
	upper	0.236	0.256	0.260	0.265
5	lower	0.209	0.213	0.205	0.212
	upper	0.329	0.354	0.324	0.343
6	lower	0.197	0.199	0.199	0.201
	upper	0.300	0.297	0.297	0.305
7	lower	0.230	0.200	0.204	0.220
	upper	0.375	0.347	0.335	0.353
8	lower	0.228	0.180	0.199	0.212
	upper	0.390	0.350	0.338	0.367

The conditional standard errors for the lower cut-scores were consistently lower than the conditional standard errors for the upper cut-scores. The conditional standard errors for both the lower and upper cut-scores were similar at grades 5 and 6 across the four scaling procedures, and systematically increased from grade 3 through 8, with the increase more so for the upper cut-scores than for the lower cut-scores. The variability among the conditional standard errors across grades and procedures was less for the lower cut-scores (0.18 to 0.29) than for the upper cut-scores (0.24 to 0.50). The results were similar across the four scaling methods with slight differences for the upper cut-score for the two upper grade levels. Figure 31 shows the conditional standard error results for the same information provided in Table 14.

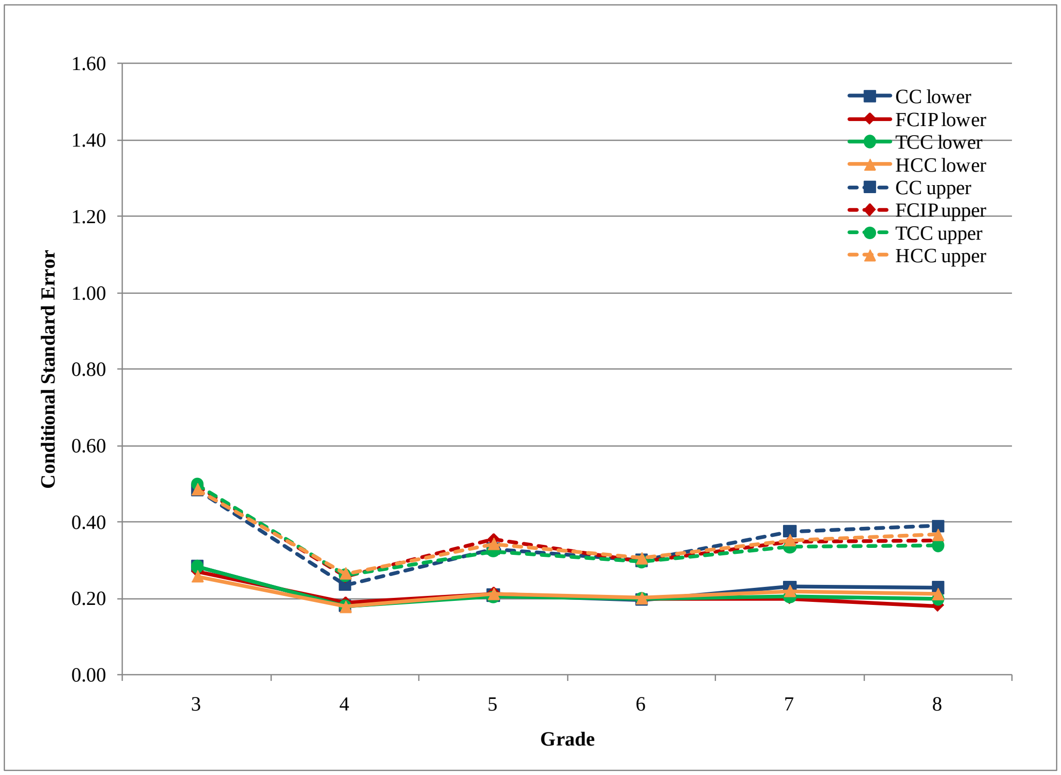

Figure 31. Conditional Standard Error for Mathematics, normal distribution, 1,500 sample size

As shown in Figure 31, there was strong agreement between the conditional standard errors for the lower cut-score, where small differences were found for the upper two grade levels. The lowest value was found for the FCIP procedure and slightly higher values were found for the CC, TCC and HCC procedures, particularly at grade 8. Although, the TCC procedure had slightly higher values for the lowest grade, these differences were minimal.

The conditional standard error values for the upper cut-score were larger than those for the lower cut-score and slight differences between the scaling procedures occurred at the two upper grade levels. Further, the values were very similar for the four scaling procedures across the grade levels, varying by less than 0.04. The largest conditional standard error values were for grade 3 (0.50), the smallest was for grade 4 (0.24).

Normal distribution 3,000 sample size

The mean conditional standard error results are displayed in Figure 32 for a normal distribution and a sample size of 3,000. The conditional standard error values are shown in Table B17 in Appendix B.

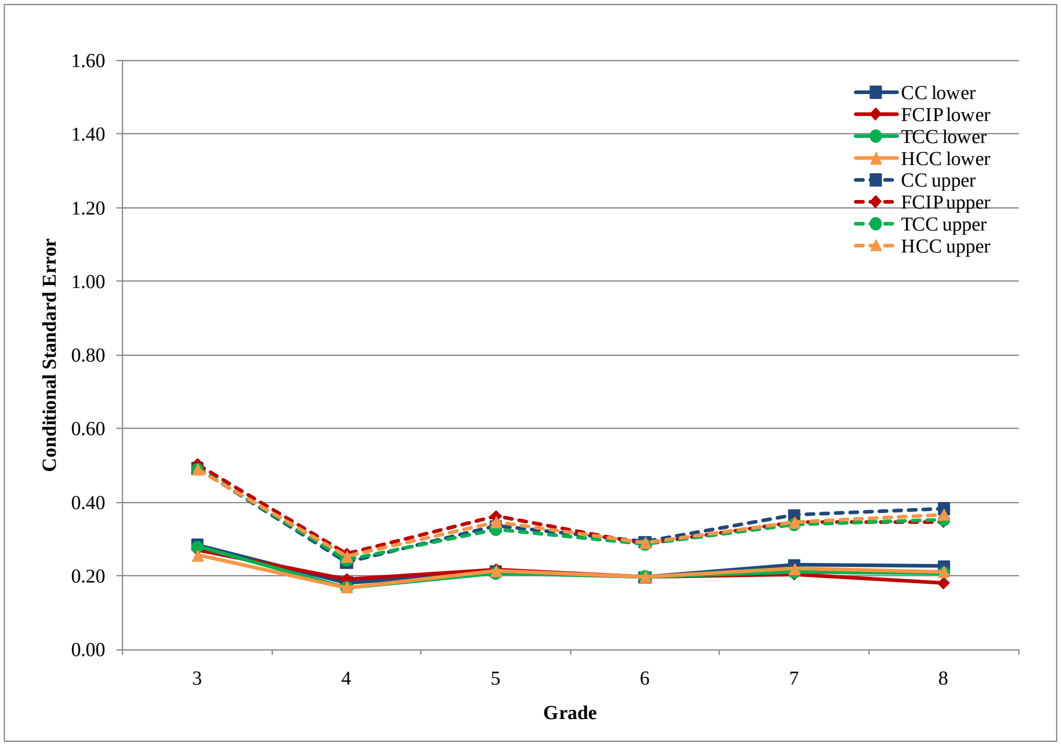

Figure 32. Conditional Standard Error for Mathematics, normal distribution, 3,000 sample size

As shown in Figure 32, the conditional standard errors for the four scaling procedures at the lower cut-score were relatively constant from grades 4 through 6. The largest conditional standard errors were for grade 3, approximately 0.30, which decreased to approximately 0.20 for the remaining grade levels. For grade 7 and, particularly at grade 8, the conditional standard errors slightly diverge. For grade 8, the CC procedure had the largest value, approximately 0.22, and the FCIP procedure had the lowest value, approximately 0.18.

The conditional standard error values at the upper cut-score were larger than the conditional standard errors at the lower cut-score. Further, the values were very similar for the four scaling procedures across the grade levels, varying by less than 0.04. The largest conditional standard error values were for grade 3 (0.50), the smallest was for grade 4 (0.30). These results were very similar to those found in the normal distribution 1,500 sample size condition.

Skewed distribution 1,500 sample size

The mean conditional standard error results are displayed in Figure 33 for a skewed distribution and a sample size of 1,500. The conditional standard error values are shown in Table B18 in Appendix B.

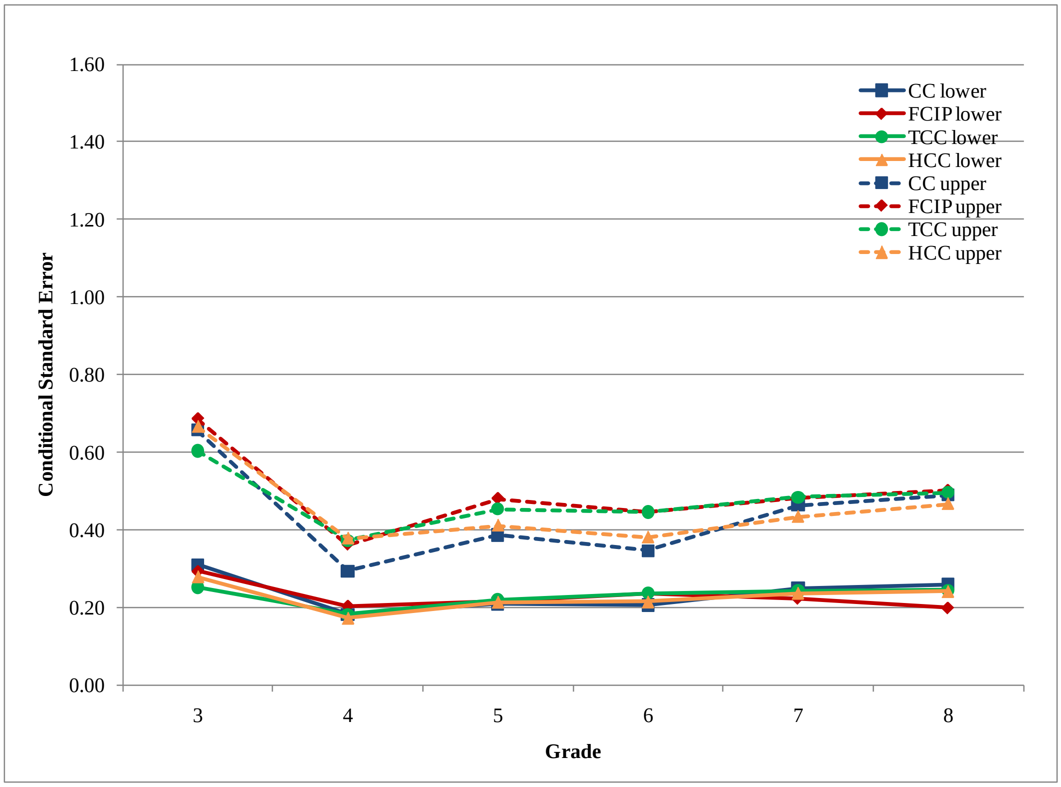

Figure 33. Conditional Standard Error for Mathematics, skewed distribution, 1,500 sample size

As shown in Figure 33, the conditional standard errors at the lower cut-score across the four vertical scaling procedures were quite close for grades 4 through 7. At grade 3, the conditional standard error values for the CC procedure were the largest and the conditional standard errors for the TCC procedure were the lowest, with the difference of 0.07. At grade 8, the lowest conditional standard error was for the FCIP procedure, while the conditional standard errors for the CC, TCC and HCC procedures were essentially the same, with a difference of 0.07.

The conditional standard error results for the upper cut-score were larger than those for the lower cut-score and larger differences between the procedures occurred at grades 4, 5, and 6. The CC procedure had the smallest values compared to the other three procedures for grade 4, 5, and 6, and the FCIP and TCC procedures had the largest values. The conditional standard errors for the lower cut-scores across all grade levels were lower and did not differ across grade levels. The conditional standard errors were largest for the four vertical scaling procedures at grade 3 and were smaller for the remaining grade levels.

Skewed distribution 3,000 sample size

The mean conditional standard error values are displayed in Figure 34 for a skewed distribution and a sample size of 3,000. The conditional standard error values are shown in Table B19 in Appendix B.

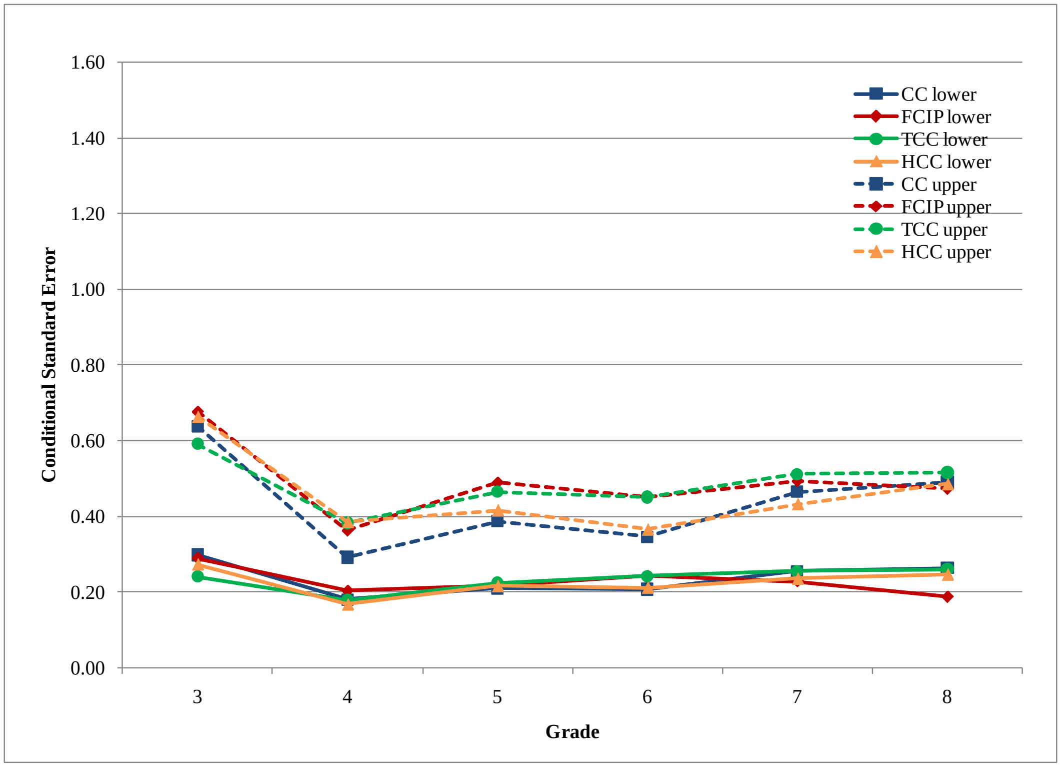

Figure 34. Conditional Standard Error for Mathematics, skewed distribution, 3,000 sample size

As shown in Figure 34, the conditional standard errors for the lower cut-score were relatively consistent, where small differences were found for grade 8. For grade 8, the lowest value was found for the FCIP procedure with slightly higher values for the CC, TCC and HCC procedures. The conditional standard errors were again higher for grade 3.

The conditional standard error values for the upper cut-score were larger than those for the lower cut-score and larger differences between the scaling procedures occurred at grades 4, 5, 6 and 7. For grade 8, the conditional standard errors were more similar. The CC procedure had smaller values compared to the other three procedures for grade 4, 5, and 6, and the FCIP and TCC procedures had the largest values. The conditional standard errors for the lower cut-score across all grade levels were lower and did not differ much across grade levels. The conditional standard errors were largest for the four vertical scaling procedures at grade 3 and were smaller for the remaining grade levels.

Summary of Conditional Standard Error

The conditional standard errors were very consistent across vertical scaling procedures for the normal and skewed distribution conditions for both sets of cut-scores. However, the conditional standard errors were larger in value for the upper cut-score than for the lower cut-score for all four scaling procedures. The change in distribution slightly changed the conditional standard error patterns among the vertical scaling procedures for the skewed distribution conditions as compared to the normal distribution conditions. The sample size slightly changed the conditional standard error values as they decreased with the larger sample size.

Root-Mean-Squared-Difference

Normal distribution 1,500 sample size

The root-mean-squared-differences between the six pairs of the four vertical scaling procedures for the normal distribution, 1,500 sample size condition are presented in Table 15. Similar to the conditional standard error results there is no metric to indicate which values are too high.

Table 15.

Root-Mean-Squared-Difference for Vertical Scaling procedures for Mathematics, normal distribution, 1,500 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	10.653	5.653	4.468	10.160	9.980	3.680
4	7.397	6.960	6.368	5.263	5.319	1.673
5	3.464	4.077	2.718	4.901	2.991	3.415
6	4.051	4.051	3.392	0.000	1.132	1.132
7	5.253	6.075	2.892	3.126	3.965	4.238
8	7.396	4.904	3.046	5.827	5.718	2.927

The results reported in Table 15 indicate that the agreement between pairs of scale scores was affected by grade level more so than vertical scaling procedure. Most of the RMSDs were below 5.0. The CC/FCIP procedure had RMSDs greater than 5.0 for grades 3, 4, 7, and 8. The FCIP/TCC and FCIP/HCC pairs had RMSDs above 5.0 for grades 3, 4, and 7. The CC/HCC pair was above 5.0 only for grade 4, and the TCC/HCC pair had lower values for all grade levels. The largest RMSDs were found at grade 3 which is the farthest grade level from the base grade.

Figure 35 shows the RMSD results for the same information provided in Table 15. The grade levels are shown on the X-axis. The value of the RMSD is shown on the Y-axis. The RMSDs for the CC/FCIP pair is shown in blue (square markers), CC/TCC pair is red (diamond markers), CC/HCC pair is green (circle markers), FCIP/TCC pair is purple (triangle markers), FCIP/HCC pair is yellow (asterisk markers), and TCC/HCC pair is orange (cross markers).

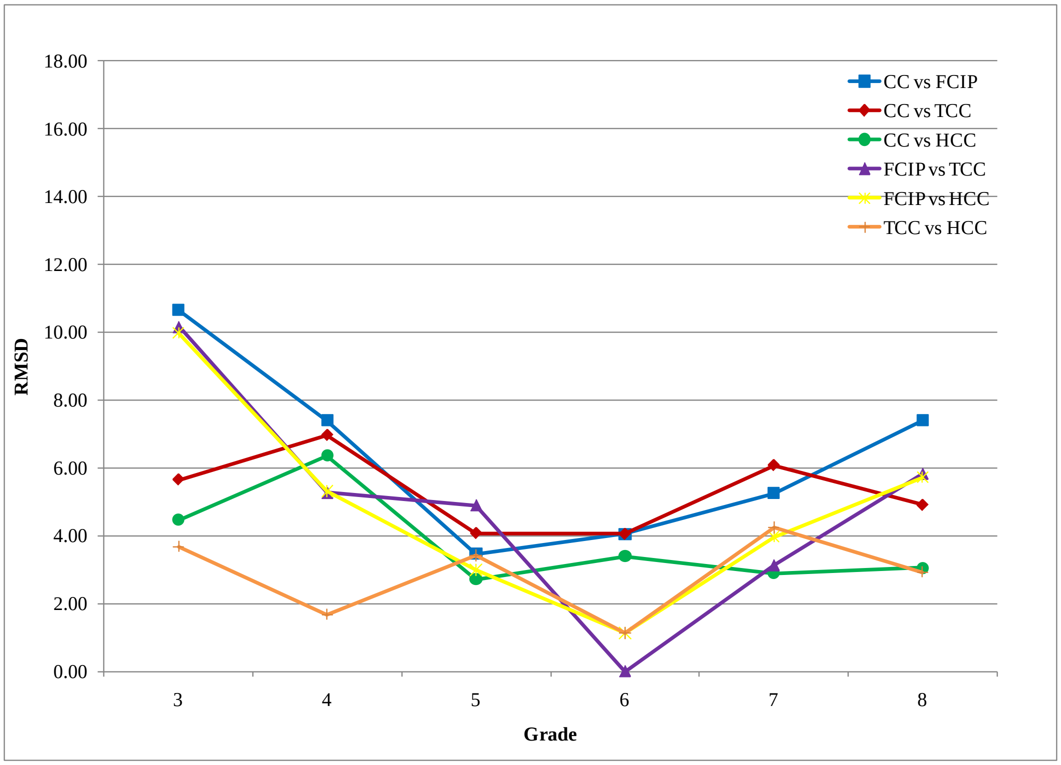

Figure 35. Root-Mean-Squared-Difference for Mathematics, normal distribution, 1,500 sample size

As shown in Figure 35, the patterns are parabolic in shape with the lowest RMSDs occurred for grades 5 and 6 and larger RMSDs found for the two upper grade levels and the two lowest grade levels. The largest RMSDs occurred for the CC/FCIP pair for grades 3, 4 and 8. For grade 5 the largest differences between procedures occurred for the FCIP/TCC pair, and for grade 7 the CC/TCC pair. Similar but smaller differences were found for the CC/HCC pair. The differences for the CC/FCIP pair were larger as the grades increased or decreased from the base grade of 6. Smaller differences were found for the CC/HCC pair for grades 5 through 7, and the TCC/HCC pair for nearly all grade levels.

Normal distribution 3,000 sample size

Figure 36 shows the RMSD results for the normal distribution 3,000 sample size. The values of the RMSD are provided in Table B20 in Appendix B.

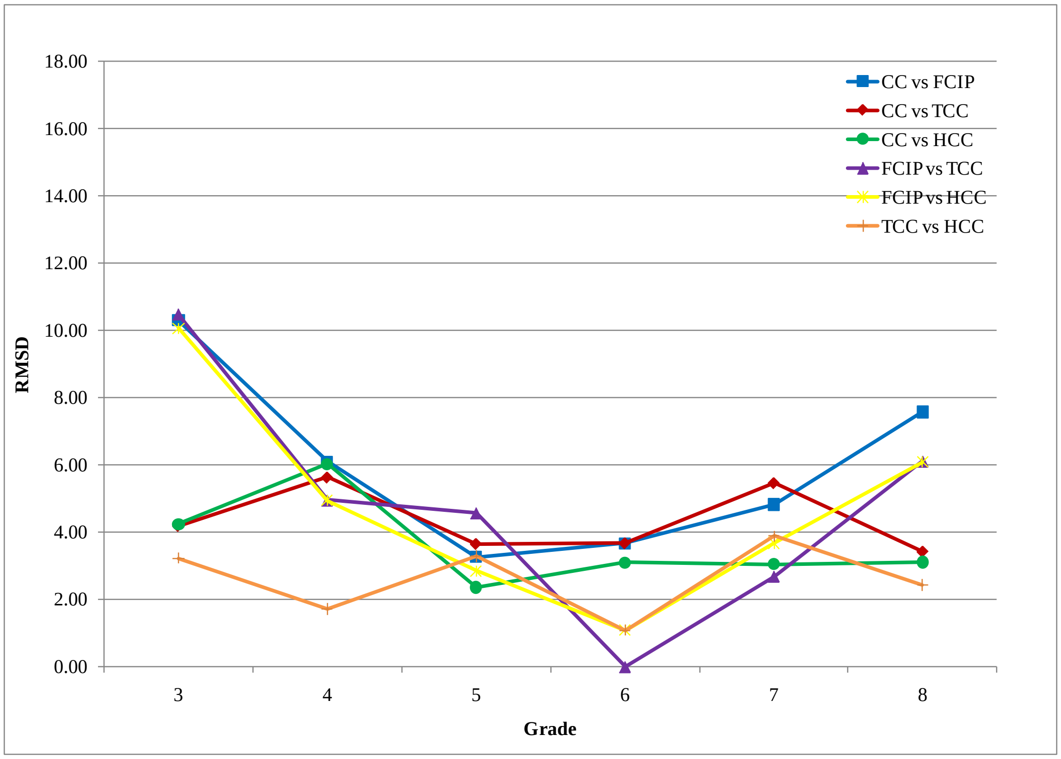

Figure 36. Root-Mean-Squared-Difference for Mathematics, normal distribution, 3,000 sample size

As shown in Figure 36, the largest RMSDs for the CC/FCIP, FCIP/TCC, FCIP/HCC pairs were found for grades 3 and CC/FCIP for grade 8. For grade 7 the largest differences between procedures occurred for the CC/TCC pair. Similar but smaller differences were found for the CC/HCC pair. The differences for the CC/FCIP pair were larger as the grades increased or decreased from the base grade of 6. Smaller differences were found for the CC/HCC pair for grades 5 through 7, and the TCC/HCC pair for nearly all grade levels. These results were similar to those found in the normal distribution 1,500 sample size condition.

Skewed distribution 1,500 sample size

Figure 37 shows the RMSD results for skewed distribution 1,500 sample size. The values of the RMSD are provided Table B21 in Appendix B.

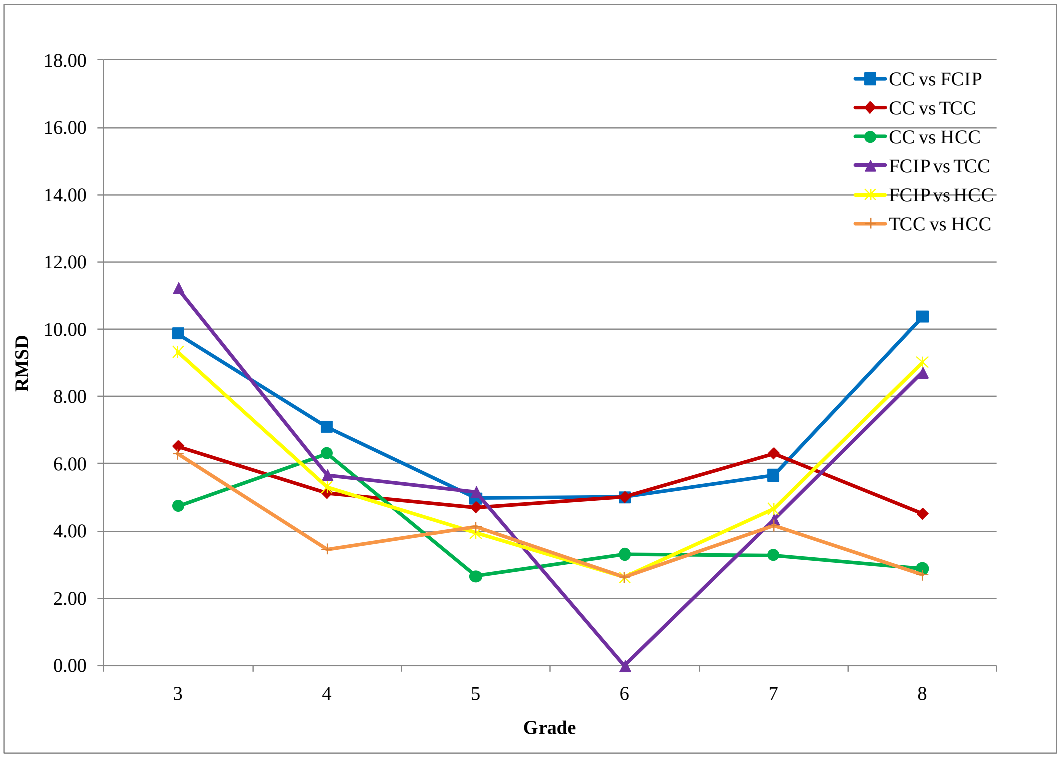

Figure 37. Root-Mean-Squared-Difference for Mathematics, skewed distribution, 1,500 sample size

The RMSDs of the skewed distribution 1,500 sample size condition patterns were slightly different to those found for the normal distribution sample size condition patterns (cf. Figure 35 see page 160 and Figure 37). As shown in Figure 37, the largest differences between procedures occurred for FCIP/TCC and CC/FCIP, particularly at grades 3, 4, 7, and 8. There were also larger differences in scale scores for the FCIP/HCC pair. But the differences among CC/TCC, CC/HCC, and TCC/HCC procedures were smaller across all grade levels as compared to the normal conditions. The most consistent differences across the grades involved HCC compared to CC and TCC. There was no difference for the FCIP/TCC pair at grade 6.

Skewed distribution 3,000 sample size

Figure 38 shows the RMSD results for skewed distribution 3,000 sample size. The RMSD values are shown in Table B22 in Appendix B.

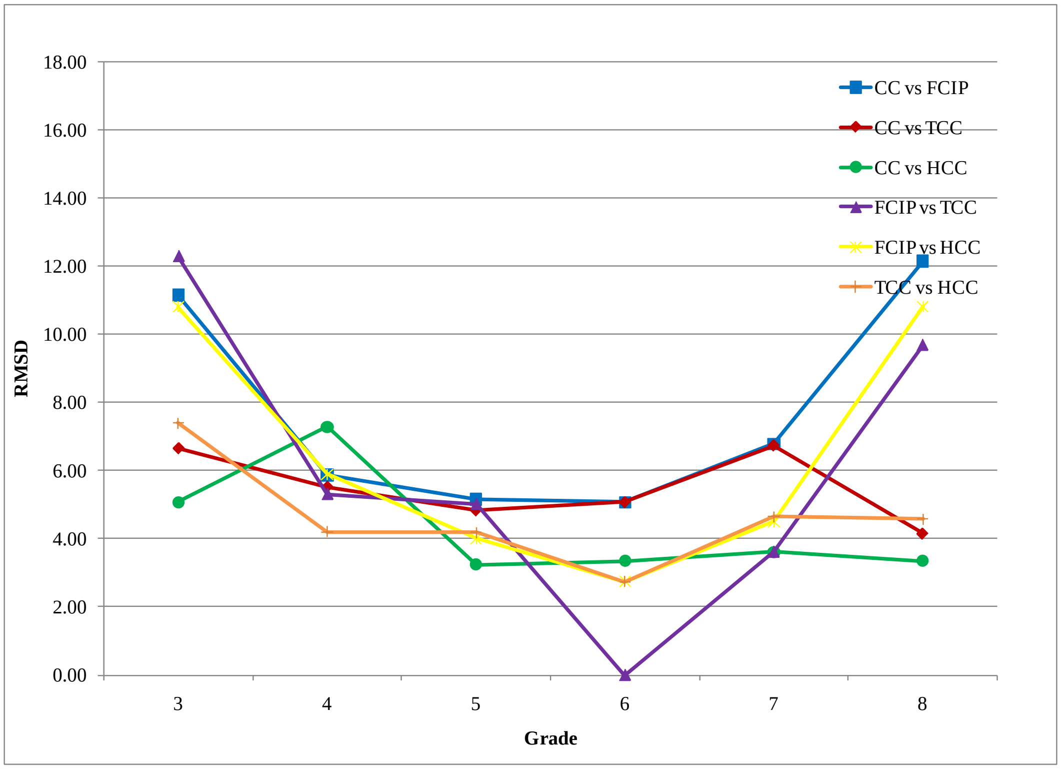

Figure 38. Root-Mean-Squared-Difference for Mathematics, skewed distribution, 3,000 sample size

This figure shows that the largest differences between procedures occurred for the FCIP/TCC pair for grades 3 and CC/FCIP for grade 8. For grade 7 the largest differences between procedures occurred for the CC/TCC pair. Similar but smaller differences were found for the CC/FCIP and FCIP/HCC pairs. The differences for the CC/FCIP pair were larger as the grades increased or decreased from the base grade of 6. Smaller differences were found for the CC/HCC procedures for grades 5 through 7, and TCC/HCC pair for nearly all grade levels.

Summary of Root-Mean-Squared-Difference

The RMSDs showed variability especially at the tail ends of the grade span. This indicates that the resulting scale scores were sometimes different between vertical scaling procedures. The patterns were similar for the normal distribution for both sample size conditions. But some of the values were slightly smaller for the larger sample size condition. There were differences between the normal and skewed distributions, where more variability was found for the skewed distribution conditions. However, since these results are aggregated over all scale scores it is difficult to determine what impact scale score differences at the cut-scores could have had on classification decisions. That is, the largest differences may be at the extreme ends of the score scale.

Correlations between Item Parameters

Normal distribution 1,500 sample size

The mean correlations between the pairs of *a*-, *b*-, and *c*-parameter estimates obtained from the four scaling procedures are presented for the normal distribution 1,500 sample size condition in Table 16. The results are presented for all six sets of correlations of the item parameters for each grade. While correlations of 0.5 are considered moderate in value (Glass & Hopkins, 1996), values less than 0.5 are bolded since the correlations were expected to be higher.

Table 16.

Correlations of item parameters fors Vertical Scaling for Mathematics, normal distribution, 1,500 sample size

			a			
Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	0.931	0.858	0.998	0.728	0.930	0.857
4	0.934	0.857	0.851	0.760	0.755	0.998
5	0.925	0.801	1.000	0.712	0.928	0.799
6	0.902	0.902	0.901	1.000	0.996	0.996
7	0.719	0.738	0.997	0.998	0.694	0.716
8	0.998	0.992	0.998	0.997	0.999	0.996
			b			
3	0.969	0.985	1.000	0.938	0.968	0.985
4	0.977	0.972	0.972	0.932	0.931	1.000
5	0.957	0.979	1.000	0.915	0.957	0.980
6	0.987	0.987	0.988	1.000	0.999	0.999
7	0.976	0.974	1.000	1.000	0.976	0.974
8	0.999	0.999	1.000	1.000	1.000	0.999
			c			
3	**0.208**	0.856	0.997	**0.215**	**0.195**	0.850
4	0.633	0.874	0.871	0.650	0.609	0.972
5	0.671	0.899	0.999	0.649	0.667	0.899
6	0.919	0.919	0.935	1.000	0.968	0.968
7	0.882	0.905	0.999	0.996	0.880	0.903
8	0.970	0.951	0.995	0.990	0.970	0.966

As shown in Table 16, the pattern of mean correlations for the *b*-parameter estimates is less complex than the pattern for the *a*-parameter estimates which is less complex than the pattern for the *c*-parameter estimates. For the *a*-parameter mean correlations, nine of the mean correlations were less than 0.80 (FCIP/TCC grade 3; FCIP/TCC and FCIP/HCC grade 4; FCIP/TCC and TCC/HCC grade 5; and CC/FCIP, CC/TCC, FCIP/HCC, and TCC/HCC grade 7), five were between 0.80 and 0.90, while the rest were greater than 0.90. Six of the nine mean correlations less than 0.80 involved FCIP.

The mean correlations for the *b*-parameter estimates were above 0.93. All of the correlations were high for all scaling pairs.

In the case of the *c*-parameter the mean correlation can be divided into two sets. The first set includes the pairs CC/TCC, CC/HCC, and TCC/HCC; the mean correlations for these pairs of scaling procedures were all at least 0.80 across the grade levels. The second set includes CC/FCIP, FCIP/TCC and FCIP/HCC; the mean correlations for these pairs of procedures were less than 0.67 for the three lower grades, and above 0.88 for the three upper grades. Further, the mean correlations for the *c*-parameter estimates for the scaling procedures were lower than the mean correlations for the corresponding *a*- and *b*-parameter estimates. The lower mean correlations for the *c*-parameter were likely due to fixing the parameters for the on-level test forms for the lower grades. The fixing of the *c*-parameters did not affect the upper grade levels since the fixed parameters were for the off-level test forms for those grade levels. Since the FCIP procedure results in fixed parameters from the upper grade level, these results may be expected as they are not estimated independently of the grade level.

Normal distribution 3,000 sample size

The mean correlations between pairs of *a*-, *b*-, and *c*-parameter estimates obtained from the different scaling procedures are presented for the normal distribution 3,000 sample size condition in Table 17.

Table 17.

Correlations of item parameters for Vertical Scaling for Mathematics, normal distribution, 3,000 sample size

			a			
Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	0.949	0.887	0.998	0.793	0.947	0.885
4	0.940	0.880	0.875	0.797	0.795	0.999
5	0.922	0.803	0.999	0.734	0.932	0.801
6	0.922	0.922	0.921	1.000	0.998	0.998
7	0.650	0.670	0.996	0.999	0.622	0.644
8	0.999	0.997	0.999	0.999	0.999	0.998
			b			
3	0.971	0.984	1.000	0.941	0.970	0.984
4	0.977	0.970	0.967	0.932	0.928	1.000
5	0.941	0.974	0.999	0.913	0.949	0.977
6	0.989	0.989	0.989	1.000	0.999	0.999
7	0.972	0.970	1.000	1.000	0.971	0.969
8	0.999	0.999	1.000	1.000	1.000	0.999
			c			
3	**0.194**	0.845	0.997	**0.214**	**0.171**	0.833
4	0.594	0.833	0.824	0.641	0.601	0.972
5	0.630	0.863	0.992	0.662	0.641	0.878
6	0.923	0.923	0.936	1.000	0.971	0.971
7	0.852	0.880	0.995	0.995	0.834	0.861
8	0.970	0.960	0.996	0.992	0.973	0.972

Similar to the normal distribution 1,500 sample size condition, the pattern of mean correlations for the *b*-parameter estimates is less complex than the pattern for the *a*-parameter estimates which is less complex than the pattern for the *c*-parameter estimates. For the *a*-parameter mean correlations, eight of the mean correlations were less than 0.80 (FCIP/TCC grade 3; FCIP/TCC and FCIP/HCC grade 4; FCIP/TCC grade 5; and CC/FCIP, CC/TCC, FCIP/HCC, and TCC/HCC grade 7), six were between 0.80 and 0.90, while the rest were greater than 0.90. Six of the eight mean correlations less than 0.80 involved FCIP.

The mean correlations for the *b*-parameter estimates were above 0.91. All of the correlations were high in value for all scaling pairs.

For the *c*-parameter the mean correlation can be divided into two sets. The first set included the pairs involved CC/TCC, CC/HCC, and TCC/HCC; the mean correlations involved these scaling procedures were all at least 0.80 across the grades. The second set includes CC/FCIP, and FCIP/TCC and FCIP/HCC; the mean correlations involved these procedures were less than 0.66 for the three lower grades, and above 0.83 for the three upper grades. Further, the mean correlations for the *c*-parameter estimates for the scaling procedures were lower than the mean correlations for the corresponding *a*- and *b*-parameter estimates. The lower mean correlations for the *c*-parameters were likely due to fixing the parameters for the on-level test forms for the lower grades. The fixing of the *c*-parameters did not affect the upper grade levels since the fixed parameters were for the off-level test forms for those grade levels.

Skewed distribution 1,500 sample size

The mean correlations between pairs of *a*-, *b*-, and *c*-parameter estimates obtained from the different scaling procedures are presented for the skewed distribution 1,500 sample size condition in Table 18.

Table 18.

Correlations of item parameters for Vertical Scaling for Mathematics, skewed distribution, 1,500 sample size

			a			
Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	0.930	0.809	0.997	0.736	0.938	0.833
4	0.891	0.764	0.807	0.775	0.772	0.968
5	0.881	0.685	0.999	0.738	0.892	0.696
6	0.777	0.777	0.870	1.000	0.932	0.932
7	0.700	0.717	0.995	0.997	0.680	0.699
8	0.994	0.989	0.999	0.998	0.990	0.985
			b			
3	0.967	0.988	1.000	0.937	0.967	0.989
4	0.973	0.963	0.971	0.927	0.932	0.997
5	0.935	0.972	0.999	0.902	0.939	0.977
6	0.981	0.981	0.990	1.000	0.995	0.995
7	0.975	0.975	1.000	1.000	0.975	0.974
8	0.999	0.998	1.000	1.000	0.999	0.999
			c			
3	**0.262**	0.835	0.997	**0.269**	**0.250**	0.831
4	0.654	0.851	0.870	0.627	0.591	0.961
5	0.637	0.828	0.995	0.643	0.632	0.825
6	0.882	0.882	0.919	1.000	0.934	0.934
7	0.849	0.861	0.998	0.996	0.848	0.859
8	0.955	0.938	0.994	0.988	0.949	0.949

Similar to the normal distribution conditions, the pattern of mean correlations for the *b*-parameter estimates is less complex than the pattern for the *a*-parameter estimates which is less complex than the pattern for the *c*-parameter estimates. For the *a*-parameter mean correlations, 13 of the mean correlations were less than 0.80 (FCIP/TCC grade 3; CC/TCC, FCIP/TCC, FCIP/HCC grade 4; CC/TCC and FCIP/TCC grade 5; CC/TCC, FCIP/TCC, and TCC/HCC grade 5, CC/FCIP, and CC/TCC grade 6; CC/FCIP, CC/TCC, FCIP/HCC, and TCC/HCC grade 7), seven were between 0.80 and 0.90, and the rest were at least 0.90. Nine of the 13 mean correlations less than 0.80 involved TCC.

The mean correlations for the *b*-parameters estimates were above 0.90. All of the correlations were high in value for all scaling pairs.

For the *c*-parameter the mean correlation can be divided into two sets. The first set included the pairs involved CC/TCC, CC/HCC, and TCC/HCC; the mean correlations involved these scaling procedures were all at least 0.83 across the grades. The second set includes CC/FCIP, and FCIP/TCC and FCIP/HCC; the mean correlations involved these procedures were less than 0.65 for the three lower grades, and above 0.85 for the three upper grades. Further, the mean correlations for the *c*-parameter estimates for the scaling procedures were lower than the mean correlations for the corresponding *a*- and *b*-parameter estimates. The lower mean correlations for the *c*-parameters was likely due to the fixing the parameters for the on-level test forms for the lower grades. The fixing of the *c*-parameters did not affect the upper grade levels since the fixed parameters were for the off-level test forms for those grade levels.

Skewed distribution 3,000 sample size

The mean correlations between pairs of *a*-, *b*-, and *c*-parameter estimates obtained from the different scaling procedures are presented for the skewed distribution 3,000 sample size condition in Table 19.

Table 19.

Correlations of item parameters for Vertical Scaling for Mathematics, skewed distribution, 3,000 sample size

			a			
Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	0.935	0.822	0.998	0.786	0.944	0.847
4	0.888	0.772	0.817	0.809	0.795	0.958
5	0.867	0.668	0.997	0.755	0.889	0.683
6	0.790	0.790	0.906	1.000	0.905	0.905
7	0.645	0.657	0.983	0.996	0.625	0.642
8	0.989	0.983	1.000	0.996	0.986	0.979
			b			
3	0.963	0.986	1.000	0.932	0.962	0.988
4	0.970	0.955	0.967	0.922	0.929	0.995
5	0.927	0.969	0.999	0.900	0.935	0.976
6	0.981	0.981	0.992	1.000	0.993	0.993
7	0.971	0.969	0.999	1.000	0.969	0.967
8	0.998	0.998	1.000	0.999	0.999	0.998
			c			
3	**0.249**	0.840	0.996	**0.240**	**0.222**	0.830
4	0.621	0.802	0.821	0.626	0.595	0.959
5	0.586	0.789	0.987	0.640	0.586	0.796
6	0.882	0.882	0.923	1.000	0.929	0.929
7	0.791	0.808	0.996	0.993	0.790	0.807
8	0.940	0.942	0.996	0.982	0.935	0.948

Similar to the previous conditions, the pattern of mean correlations for the *b*-parameter estimates is less complex than the pattern for the *a*-parameter estimates which is less complex than the pattern for the *c*-parameter estimates. For the *a*-parameter mean correlations, 12 of the mean correlations were less than 0.80 (FCIP/TCC grade 3; CC/TCC and FCIP/HCC grade 4; CC/TCC, FCIP/TCC and TCC/HCC grade 5; CC/FCIP and CC/TCC grade 6; CC/FCIP, CC/TCC, FCIP/HCC, and TCC/HCC grade 8), seven were between 0.80 and 0.90, and the rest were greater than 0.90. Nine of the 12 mean correlations involved TCC.

The mean correlations for the *b*-parameter estimates were above 0.90. All of the correlations are high for all vertical scaling pairs.

For the *c*-parameter the mean correlation can be divided into two sets. The first set included the pairs involved CC/TCC, CC/HCC, and TCC/HCC; the mean correlations involved these scaling procedures were all at least 0.83 across the grades. The second set includes CC/FCIP, and FCIP/TCC and FCIP/HCC; the mean correlations involved these procedures were less than 0.65 for the three lower grades, and above 0.85 for the three upper grades. Further, the mean correlations for the *c*-parameter estimates for the scaling procedures were lower than the mean correlations for the corresponding *a*- and *b*-parameter estimates. The lower mean correlations for the *c*-parameters were likely due to fixing the parameters for the on-level test forms for the lower grades.

Summary of Correlations

The correlations between the six pairs of scaling procedures were not affected by the distribution shape or sample size. The highest correlations were found for the *b*-parameter, where all were greater than 0.90. For the correlations for the *a*-parameter the majority of the pairs were above 0.80, with a few exceptions depending on vertical scaling pair and grade level. Lastly, the correlations for the *c*-parameter were generally quite high, with some exceptions when FCIP was a member. The smallest *c*-parameter correlations occurred for pairs that included FCIP with the lower grade levels (3, 4, and 5). This was most likely due to the lower grades having the off-level item parameters fixed. The item parameters for the *a*-, *b*-, and *c*- parameters did not differ based on the distribution type or sample size. There was no consistent pattern of which pair had the highest values across the *a*-, *b*-, and *c*-parameters, but generally the values found were high with the few exceptions presented above. These results indicate similar ranking for the vertical scaling procedure item parameters.

CHAPTER 6 DISCUSSION AND CONCLUSIONS

A brief description of the vertical scaling procedures, evaluation measures, factors in simulation study and four research questions used in this study are presented first. The findings are summarized and highlighted next at the micro level, followed by the limitations of the study. The results at a macro level are then discussed in the context of current research to date. Finally, recommendations for practice and future research are presented.

Summary of Study and Research Questions

Four vertical scaling procedures were evaluated in the present study. The first vertical scaling procedure was CC, which simultaneously estimates the item parameters across the test forms for the six grade levels included. FCIP was the second procedure, which took the item parameters at the base form for grade 6 and fixed the parameters for the adjacent grade level(s). For example, the grade 6 level form from the grade 6 students was fixed for the grade 5 students' calibration of the grade 5 form. TCC was the third procedure, which estimated the item parameters separately for each grade and then linked the item parameters using the TCC procedure (Stocking & Lord, 1983). The HCC procedure was the fourth procedure, which concurrently calibrated adjacent pairs of consecutive grades instead of all grades, and linked the pairs using the TCC procedure.

Four factors were evaluated in this simulation study. The first factor was vertical scaling procedure. The second factor was content area, which included a Reading and Mathematics assessment in grades 3 through 8. The third factor was distribution type, which included a normal distribution and negatively skewed distribution. The fourth factor was sample size, which included 1,500 and 3,000 examinees per form. There were 100 replications for each simulated condition.

Five evaluation measures were used in this study. The first two were decision accuracy and decision consistency, which were calculated for each cut-score at each grade level. Third, conditional standard errors were calculated for each cut-score at each grade level. Fourth, RMSD values were calculated between the final scale scores for each of the six pairs of vertical scaling procedures for each grade. Fifth, correlations of the final item parameters were calculated between each of the vertical scaling procedures for each grade.

The following four research questions were addressed using simulated data based on real data:

1. Do vertical scaling methods perform the same for the five evaluation criteria?
2. Does distribution shape have an effect on the five evaluation criteria?
3. Does content area have an effect on the five evaluation criteria?
4. Does sample size have an effect on the five evaluation criteria?

Summary of Findings

The following summary highlights the micro level results for the four factors in this study. Each of the four questions is summarized separately: vertical scaling procedure, distribution shape, content area, and sample size.

Vertical Scaling Procedures

Each of the four vertical scaling procedure’s micro level results are presented separately: concurrent calibration, fixed common item parameters, test characteristic curve, and hybrid characteristic curve.

Concurrent Calibration.

1. The decision accuracy values ranged from 0.52 to 0.86 for the lower cut-score and 0.87 to 1.00 for the upper cut-score. For the lower cut-score, the CC procedure had the lowest values for grades 7 and 8 for most conditions. The CC procedure had similar or larger values at each of the other grade levels as compared to the other vertical scaling procedures. For the upper cut-score, the CC procedure had the lowest values for grades 3 and 4 (0.87 to 0.89), and similar or larger values at each of the other grade levels as compared to the other vertical scaling procedures.
2. The decision consistency values ranged from 0.81 to 0.91 for the lower cut-score and 0.88 to 1.00 for the upper cut-score. While the CC procedure had a similar pattern of lower decision accuracy values for the grades described above, the decision consistency values were all much higher and the differences were much smaller.
3. The conditional standard errors ranged from 0.18 to 0.63 for the lower cut-score and 0.24 to 1.40 for the upper cut-score. Small differences between scaling procedures were found for the lower cut-score standard error. The CC procedure had the lowest conditional standard errors for grades 3 and

4 (0.24 to 0.66) and the highest values for grades 7 and 8 (0.37 to 1.40) for the upper cut-score.

4. The RMSDs ranged from 2.35 to 17.25 when CC was a member. The values were the smallest for the two middle grades (2.35 to 6.41), larger for grades 7 and 8 (2.88 to 16.70), and the largest for grades 3 and 4 (4.18 to 17.25).

5. The item parameter correlations were generally quite high (>0.60) for most of the conditions. Lower *c*-parameter correlations were found for the three lowest grade levels for the CC procedure when paired with the FCIP (0.19 to 0.70).

Fixed Common Item Parameters.

1. The decision accuracy values ranged from 0.47 to 0.87 for the lower cut-score and 0.90 to 1.00 for the upper cut-score. The FCIP procedure had similar or larger values than the other scaling procedures, with a few exceptions. The lowest values for the lower cut-score were found for grade 3 (0.56 - Mathematics normal, 1,500 and 0.52 - 3,000; 0.58 - Mathematics skewed, 1,500), and grade 4 (0.81 - Reading, skewed, 1,500 and 0.81 - 3,000). For the upper cut-score, the FCIP procedure had no noticeable differences with the other vertical scaling procedures.

2. The decision consistency values ranged from 0.79 to 0.92 for the lower cut-score and 0.91 to 1.00 for the upper cut-score. While the pattern for the FCIP procedure was similar to the pattern of decision accuracy values,

the decision consistency values were all much higher and the differences were much smaller.

3. The conditional standard errors ranged from 0.18 to 0.53 for the lower cut-score and 0.26 to 1.26 for the upper cut-score. Small differences between scaling procedures were found for both cut-scores.

4. The RMSDs ranged from 0.00 to 16.70 when FCIP was a member. The values were the smallest for the two middle grade levels (0.00 to 6.41), larger for grades 7 and 8 (1.98 to 16.70), and the largest in value for grades 3 and 4 (4.93 to 13.52).

5. The item parameter correlations were generally quite high (<0.60) for most of the conditions. Lower *c*-parameter correlations were found for the three lowest grade levels for the FCIP procedure when paired with each of the other scaling procedures (0.17 to 0.71).

Test Characteristic Curve.

1. The decision accuracy values ranged from 0.48 to 0.87 for the lower cut-score and 0.86 to 1.00 for the upper cut-score. The TCC procedure had similar or larger values than the other scaling procedures, with a few exceptions. The lowest values for the lower cut-score were found for grade 3 (0.58 - Reading normal, 1,500 and 0.59 - 3,000; 0.81 - Reading skewed, 1,500). For the upper cut-score, the TCC procedure had no noticeable differences with the other vertical scaling procedures.

2. The decision consistency values ranged from 0.80 to 0.91 for the lower cut-score and 0.88 to 1.00 for the upper cut-score. While the pattern for

decision consistency for the TCC procedure was similar to the pattern of decision accuracy values, the decision consistency values were all much higher and the differences were much smaller.

3. The conditional standard errors ranged from 0.17 to 0.55 for the lower cut-score and 0.24 to 1.23 for the upper cut-score. Small differences between scaling procedures were found for both cut-scores.

4. The RMSDs ranged from 0.00 to 17.25 when TCC was a member. The values were the smallest for the two middle grade levels (0.00 to 5.64), larger for grades 7 and 8 (1.98 to 11.40), and the largest in value for grades 3 and 4 (1.67 to 17.25).

5. The item parameter correlations were generally quite high (>0.60) for most of the conditions. Lower *c*-parameter correlations were found for the three lowest grade levels for the TCC procedure when paired with the FCIP procedure (0.21 to 0.66).

Hybrid Characteristic Curve.

1. The decision accuracy values ranged from 0.50 to 0.87 for the lower cut-score and 0.87 to 1.00 for the upper cut-score. The HCC procedure had similar or larger values than the other scaling procedures, with one exception, grade 3 (0.54 - Mathematics skewed 3,000). For the upper cut-score, the HCC procedure had no noticeable differences with the other vertical scaling procedures.

2. The decision consistency values ranged from 0.80 to 0.91 for the lower cut-score and 0.88 to 1.00 for the upper cut-score. While the patterns of

decision consistency for the HCC procedure was similar to the pattern of decision accuracy values, the decision consistency values were all much higher and the differences were much smaller.

3. The conditional standard errors ranged from 0.17 to 0.58 for the lower cut-score and 0.25 to 1.22 for the upper cut-score. Small differences between scaling procedures were found for both cut-scores.
4. The RMSDs ranged from 1.08 to 16.11 when HCC was a member. The values were the smallest for the two middle grade levels (1.08 to 5.18), larger for grades 7 and 8 (2.43 to 11.52), and the largest in value for grades 3 and 4 (1.67 to 16.11).
5. The item parameter correlations were generally quite high (<0.60) for most of the conditions. Lower *c*-parameter correlations were found for the three lowest grade levels for the HCC procedure when paired with the FCIP procedure (0.17 to 0.69).

Distribution Type

In contrast to decision accuracy and consistency, the micro level results are presented comparatively for the normal distribution and skewed distributions.

1. Of the five outcome measures, decision accuracy was most affected by the shape of the distribution at the lower cut-score. Further, the differences found for decision accuracy were content area dependent. For Reading, the decision accuracy values decreased as grades increased for the skewed conditions, for the lower cut-score. In contrast for Mathematics, the decision accuracy values for the vertical scaling procedures for the skewed

conditions larger differences were found primarily at grade 3 and smaller differences at grades 7 and 8, which were not found in the normal conditions for the lower cut-score. For the upper cut-score, neither the Reading nor Mathematics decision accuracy values changes from the normal as compared to the skewed distributions.

2. The decision consistency values were similar across scaling procedures and grade levels with small differences between the normal and skewed distributions.
3. The patterns of conditional standard errors for the vertical scaling procedures across grade levels were relatively similar for the normal and skewed distributions. But for the normal distribution there was less variability across scaling procedures, primarily for the upper cut-score.
4. There were no noticeable differences between the normal and skewed distributions for the RMSDs.
5. There were no noticeable differences between the normal and skewed distributions for the item parameter correlations.

Content Area

Similar to the summary for distribution type, the content area micro level results are presented comparatively for Reading and Mathematics.

1. The decision accuracy values ranged from 0.58 to 1.00 for Reading, and from 0.47 to 0.99 for Mathematics. Lower values were found for the lower cut-score as compared to the upper cut-score for both Reading and Mathematics. For the lower cut-score, Reading generally had higher

accuracy values compared to Mathematics for grade 3 (0.58 to 0.87 vs. 0.47 to 0.81), lower values for grades 4 and 5 (0.72 to 0.82 vs. 0.79 to 0.87), and similar values for grade 6 through 8. Whereas Reading had more variability across scaling procedures than Mathematics, Mathematics had more variability across grade levels. For the upper cut-score, Reading had no noticeable differences as compared to Mathematics.

2. The decision consistency values ranged from 0.79 to 1.00 for Reading, and from 0.81 to 0.98 for Mathematics. Reading had decision consistency values that were similar across scaling procedures and grade levels while there were small differences with Mathematics.

3. The conditional standard errors ranged from 0.21 to 1.40 for Reading, and from 0.17 to 0.69 for Mathematics. The conditional standard errors for Reading were larger and more variable than the conditional standard errors for Mathematics. Further, the conditional standard errors for Reading increased as grade level increased, while the conditional standard errors for Mathematics were slightly more consistent across grade levels and vertical scaling procedures.

4. The RMSDs were generally larger in value for Reading (0.00 to 17.25) as compared to Mathematics (0.00 to 12.30).

5. There were no noticeable differences between the item parameter correlations for Reading and Mathematics.

Sample Size

1. The sample size factor had little to no effect on the values and patterns for the outcome measures, with the exception of the conditional standard errors and RMSDs. The conditional standard errors and RMSDs were smaller for the larger sample size condition, but the patterns for the outcome measures were nearly identical.

Limitations of the Study

There were four limitations to this study. First, the decision accuracy and consistency rates for false positive and negative could have been examined to determine whether incorrect decisions whether the inaccurate results occurred by incorrectly placing students in the lower or upper proficiency categories. The scope of this study did not include identification of false positive or false negative classification rates. Rather, the intent of the present study was to provide a general indication of how well the different vertical scaling procedures impacted the overall decision accuracy and consistency results.

Second, the results of this study are specific to the particular set of real data used to simulate the conditions conducted in this study, namely examinee responses to the Reading and Mathematics assessment administered in a particular U.S. state assessment. Other data sets obtained from different state assessments in which vertical scaling was implemented were not used.

Third, tests with mixed item formats were not addressed in the present study. Consequently, the findings of this study can only extend to assessments containing only multiple choice items.

Fourth, the dimensionality of the vertical scaling data cannot easily be determined. Because the item information occurs by grade level and only adjacent grade levels have items in common, a principal component analysis cannot be conducted for the full grade span.

Discussion

In this section, the results of this study are integrated and discussed for three aspects of the previous research, which were previously identified in the section Short Comings of Reported Research on vertical scaling (see page 62). These results are discussed at the macro level. First, negatively skewed distributions and normal distributions were included in the present study, which had not been compared in previous vertical scaling research. Second, Reading and Mathematics content area data files were used to simulate and compare different vertical scaling procedures. Third, three new outcome measures for standard setting outcomes were compared for the vertical scaling procedures, namely, decision accuracy and decision consistency, and conditional standard errors of estimation.

Distribution type

First, the research to date has not compared a skewed distribution to a normal distribution in a simulation study. Only one previous study evaluated a non-normal distribution (Custer, et al., 2006), where only a slightly skewed distribution that matched the real data was simulated and no comparison to a normal distribution was made. In the present study, the type of distribution, normal or moderately skewed, had an impact primarily on the decision accuracy results, more so for Reading than Mathematics. The decision accuracy values for the four scaling procedures for Reading were affected primarily at the upper and lower grade levels and for Mathematics were affected at grade 3. Distribution type had little to no impact on the decision consistency, conditional standard error, RMSDs, and item parameter correlations outcome measures. This result indicates that as negatively skewed distributions are found, the distribution type will not necessarily impact the vertical scaling results for most of the outcome measures. However, since the decision accuracy results were impacted, caution should be taken when implementing a vertical scale for distributions that have a negatively skewed distribution. Practitioners can use this information to focus evaluation of vertical scaling on the decision accuracy results since these were the most impacted by the skewed distribution conditions.

Content area

Second, this study was the first to evaluate more than one content area using simulation techniques. It is important to ensure that the results of a vertical scaling procedure are appropriate as the results indicate content area dependent results, specifically for classification decisions. Testing agencies that implement vertical scaling procedures typically want to use the same scaling procedure across content areas, and it is important to ensure that this process is appropriate. Three previous studies, as outlined in the shortcomings of reported research, compared content areas, although not in a simulation study (Ito, et al., 2008; J. Kim, 2007; Tong & Kolen, 2007). Larger RMSD results between HCC and CC were found as grades were farther from the base grade level in the Ito et al. (2008)

study. However, while Ito et al. found that the RMSDs were larger for Mathematics than for Reading, larger RMSDs were found in Reading than Mathematics in the present study.

Kim (2007) evaluated CC, FCIP, and TCC for different content areas in normed assessment batteries in terms of effect sizes and grade-to-grade growth. However, since only effect sizes and grade-to-grade growth were the outcome measures, this study is difficult to compare to the current study. The Kim study found that there were differences between the content areas, similar to the findings from this study. Even though the outcome measures were different in the two studies both studies support that content areas are differentially affected by the outcome measures used in each study.

Tong and Kolen (2007) compared Thurstone scaling design to the 3-PL test and common item designs for four content areas with similar outcome measures to the Kim study. Differences in content areas were also found, but the results are difficult to directly compare to the current study given the outcome measures were different. However, these three studies from the literature indicate that vertical scaling procedures do not always produce the same outcomes for different content areas.

In the current study, although decision consistency had similar results for Reading and Mathematics, the results for decision accuracy were different depending on the grade level, scaling procedure, content area, and distribution type, and the conditional standard errors and RMSDs were larger for the Reading content area as compared to Mathematics. It is unclear if there is something inherently different between the constructs of Reading and Mathematics that will influence vertical scaling, or that the different number of items on the test forms had an effect on the results.

Standard setting outcome measures

This study evaluated the four vertical scaling procedures using different outcome measures, namely decision accuracy, decision consistency and conditional standard errors at the lower and upper cut-score(s). Two studies evaluated vertical scaling with standard setting outcome measures. Meng (2007) evaluated CC, TCC, and two versions of HCC for the absolute bias, SE, and RMSE between the true and estimated proportion classification values. However, the results from Meng's study were difficult to interpret as the differences were summed across grade levels and then averaged across conditions to compare the vertical scaling procedures. In the Meng study comparisons to truth were made, and in the current study, measures of classification decisions were compared. In contrast to Meng's study, the current study did not find inconsistent patterns across the scaling procedures and grade levels for the standard setting outcomes used. In fact, fairly consistent patterns were seen for all the outcome measures and conditions.

In another study, Jodoin et al. (2003) placed examinees in proficiency categories based on cut-scores and compared CC, FCIP, and MS for the percentages of students within the categories, which is similar to a classification measurement. Jodoin et al. evaluated these results with real data and not replicated samples, and evaluated year-to-year growth and not a single year vertical scaling design. Jodoin et al. found that the MS and FCIP procedures produced a slightly higher degree of agreement for proficiency categories than the CC procedure, and the FCIP procedure placed examinees in the next higher proficiency category more often than the MS and CC procedures for EAP and inconsistently classified students for MLE. The results of the current study consistently indicate differences between vertical scaling procedures similar to the Jodoin et al. study, but since the Jodoin et al. study was a year-to-year evaluation with only one grade level, a direct comparison is more difficult. Neither Meng (2007) nor Jodoin et al. (2003) evaluated decision accuracy, decision consistency or conditional standard errors.

The decision accuracy results seemed most affected by the different vertical scaling procedures. Comparison of vertical scaling procedures on the impact of decision accuracy is important to ensure that the procedures being used are adequate for all grade levels. For Reading, differences were found for the decision accuracy results primarily for grade levels farther from the base grade level, where the HCC procedure had no really low values for any of the conditions and had similar or larger values than the values for CC, FCIP, and TCC procedures for most conditions. For a Mathematics assessment, all procedures performed similarly, except for the skewed distributions, where the TCC showed higher values. These results indicate that practitioners need to treat with caution the decision accuracy results when implementing a vertical scale.

Conclusions

Vertical scaling research has developed and expanded. There are no "true" values for a vertical scale, because a particular vertical scaling procedure needs to be implemented to create the scale. This creates problems in evaluating "truth" even when real data is being used. The approach in this research study was to evaluate each procedure as a test developer would in practice. Therefore, the outcome measures were evaluated using standard setting outcomes that are commonly used when developing vertical scales. Practitioners should take care to ensure that the reliability of the cut-score decisions are not affected at all grade levels, because measures of growth in Annual Yearly Progress rely on good measurement at all grade levels.

Taken together, a pattern of which vertical scaling procedure produced the highest decision accuracy results somewhat emerged. Since decision accuracy was the most affected outcome measure, and it measures how accurately students are placed in proficiency categories, the decision accuracy results should be evaluated and considered when deciding which scaling procedure to implement. The CC procedure produced the lowest decision accuracy values for the grade levels farthest from the base grade, but in opposite directions for both the lower and upper cut-scores (higher accuracy results were found for the lower grades for the lower cut-score, and higher accuracy results were found for the upper grades for the upper cut-score). Unfortunately the vertical scales developed for state assessments do not usually include two or three grade levels under NCLB, and therefore the use of the CC procedure may not be appropriate.

The FCIP, TCC and HCC procedures generally produced either the highest or similar results for the five outcome measures, although the TCC and FCIP procedure did have a few more low values depending on the content area and conditions of the data used to create the vertical scale. Unfortunately, the item parameter correlations for the FCIP procedure had some of the lowest values, specifically for the *c*-parameter for some grade levels. The correlations were generally quite high for all other pairings, but fixing the item parameters seemed to have affected the *c*-parameter estimation for the FCIP procedure. Further, the FCIP procedure is difficult to implement as the specification of quadrature points is required in BILOG-MG, when creating a vertical scale. Whereas LOGIST 5.0 was used in the study conducted by Becker and Forsyth (1992), PARSCALE 3.5 was used by Jodoin et al. (2003), only BILOG-MG and MULTILOG are specifically capable of handling multiple grades (Kolen & Brennan, 2004). Use of MULTILOG, would also require quadrature points if the FCIP procedure is used. The different *c*-parameters and the requirement of the quadrature points may indicate that using the FCIP procedure to create a vertical scale is not recommended.

The HCC procedure seemed to have more stable results than the TCC and FCIP procedures and did not have the issue of low *c*-parameter correlations like the FCIP procedure. Therefore, the HCC procedure could be used for both Reading and Mathematics as the results were good or the best for most conditions. However, if the state department of education allows for different procedures for content areas then the HCC procedure would be most appropriate for Reading while the TCC procedure would be most appropriate for Mathematics.

The results of this study indicate how difficult it is to create a vertical scale that is accurate and can be used to measure growth over time. Given the poor decision accuracy results for some grade levels in this study, can the measures of growth from a vertical scale be trusted? Given present accountability practices, can we rely on IRT vertical scaling methods to measure such growth? Unfortunately for now, the methods considered in the present study are the methods most prominently used. Therefore procedures that determine the decision accuracy and consistency of the decisions made about students' placement in proficiency categories for each of the grade levels should be of interest. There is already difficulty in creating test forms that increase in difficulty and complexity and shows growth grade-by-grade. In addition, accurate cut-scores in the score distribution of each grade test form must be established to reflect year-to-year growth. In addition to these two aspects, the results of this research study indicate that classification decisions should also not be neglected. To ensure fairness, it is necessary that educational practitioners working with federal and state governments responsible for education closely attend to these issues to ensure that the scores used to place each student in a proficiency category is valid and that the placements are accurate and consistent.

Recommendations for Practice

This study has a few implications for practitioners for implementing a vertical scale using one of the four scaling procedures. First, choice of vertical scaling procedure requires three considerations: content area, distribution type, and range of grade levels. For Reading, differences were found for the decision accuracy results, primarily for the grade levels furthest from the base grade level for the lower cut-score. For the Mathematics, there were fewer differences between the vertical scaling procedures, but where there were differences in some cases they were rather large. Practitioners should carefully examine the decision accuracy results to ensure that high values are found for all grade levels at both cut-scores for each content area assessed, especially when negatively skewed distributions are found. Negatively skewed distributions are commonly seen in testing programs developed to satisfy the requirements of NCLB. The information at the lower cut-score is typically used to determine AYP.

For the most part the FCIP and HCC procedures were the most consistent in not having very low values for any grade level for Reading and the TCC procedure for Mathematics. But the implementation of the FCIP using BILOG-MG is problematic as the appropriate quadrature points were required to create the grade separation for a vertical scale. The “correct” method of creating the appropriate scale is not well documented for the FCIP procedure. Overall, the HCC procedure appears to be the best for Reading, and the TCC procedure for Mathematics. But if a common procedure across content areas is implemented, then the HCC procedure should be used.

Third, a sample size of 1,500 was adequate for the outcome measures evaluated in this study. Therefore, collection of a sample size of 1,500 per form for the data collection design used in this study is adequate.

Recommendations for Future Research

There are several areas that should be examined in future research studies. First, MLE or MAP estimates can be used instead of the EAP estimate to see if similar patterns of results are found, particularly with decision accuracy, decision consistency, and conditional standard errors. The primary focus of this study was to evaluate the four vertical scaling procedures in terms of decision accuracy, decision consistency, and conditional standard errors. Evaluating MLE scoring method could be helpful to determine if the results from the current study are generalizable to MLE.

Second, the Reading and Mathematics tests in the present study only included multiple choice items. The study should be replicated with a combination of multiple choice and constructed response items given educational assessments often include both item types.

Third, false positive and false negative rates should be examined. The values could determine if the decisions being made for the different vertical scaling procedures place students in the lower or upper proficiency category and thereby indicate the type of error by the conditions evaluated in this study.

Fourth, smaller sample sizes could be implemented in a similar study. Ensuring that similar results are found with a smaller sample of 1,000 or even 500 examinees may be useful to practitioners. A smaller sample size would enable the practitioner to reduce resource load and sampling issues.

REFERENCES

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*(2), 147-162. doi: 10.1111/j.1745-3984.1991.tb00350.x

Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement, 29*(4), 341-354. doi: 10.1111/j.1745-3984.1992.tb00382.x

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444. doi: 10.1177/014662168200600405

Brennan, R. L. (2004). *Manual for BB-Class: A computer program that uses the Beta-binomial model for classification consistency and accuracy*. [Computer software] Version 1.1.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3-14. doi: 10.1111/j.1745-3992.2009.00158.x

Burket, G. R. (1991). *PARDUX*. [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Camilli, G., Yamamoto, K., & Wang, M. M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379-388. doi: 10.1177/014662169301700407

Chin, T. Y., Kim, W., & Nering, M. L. (2006). *Five statistical factors that influence IRT vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment, 1*(4), 329-347.

Cook, L. L., & Douglass, J. B. (1982). *Analysis of fit and vertical equating with the three-parameter model*. Paper presented at the annual meeting of the American Education Research Association, New York, NY.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

CTB/McGraw-Hill. (2001). *TerraNova, The Second Edition (CAT/6)*. Monterey, CA: Author.

Custer, M., Omar, M. H., & Pomplun, M. (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education, 19*(2), 133-149. doi: 10.1207/s15324818ame1902_4

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in eduation and psychology* (3rd ed.). Boston, MA: Allyn & Bacon.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement, 5*(2), 187-201. doi: 10.1177/014662168100500204

Gustafsson, J. E. (1979). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement, 16*(3), 153-158. doi: 10.1111/j.1745-3984.1979.tb00096.x

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

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

Hanson, B. A. (2002). IRT Command Language. Retrieved from http://www.b-a-h.com/software/irt/icl/

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3-24. doi: 10.1177/0146621602026001001

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*, 345-359. doi: 10.1111/j.1745-3984.1990.tb00753.x

Harris, D. J. (1991). A comparison of Angoff's design I and design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement*, *28*(3), 221-235. doi: 10.1111/j.1745-3984.1991.tb00355.x

Harris, D. J., & Hooker, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*(2), 151-159. doi: 10.1177/014662168701100203

Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27*, 3-17.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-30). New York, NY: Springer Science + Business Media, LLC.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: Praeger Publishers.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement*, *19*(2), 139-147. doi: 10.1111/j.1745-3984.1982.tb00123.x

Hoover, H. D., Dunbar, S. D., & Frisbie, D. A. (2003). *The Iowa tests. Guide to development and research*. Itasca, IL: Riverside Publishing.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*, 253-264. doi: 10.1111/j.1745-3984.1976.tb00016.x

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education, 21*(3), 187-206. doi: 10.1080/08957340802161741

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

Education, 71(3), 229-250. Retrieved from http://www.heldref.org/pubs/jxe/about.html

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003). *Separate versus concurrent calibration methods in vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Keller, L. A., Skorupski, W. P., Swaminathan, H., & Jodoin, M. G. (2004). *An evaluation of item response theory equating procedures for capturing changes in examinee distributions with mixed-format tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Kim, J. (2007). *A comparison of calibration methods and proficiency estimators for creating IRT vertical scales.* (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses. (ProQuest document ID: 1379532271).

Kim, S. H., & Cohen, A. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143. doi: 10.1177/01466216980222003

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*(1), 1-11. doi: 10.1111/j.1745-3984.1981.tb00838.x

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: Praeger Publishers.

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

Lee, W. C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412-432. doi: 10.1177/014662102237797

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83-102. doi: 10.1207/s15324818ame0601_5

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(5), 3-16. doi: 10.3102/0013189X031006003

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179-197. doi: 10.1111/j.1745-3984.1995.tb00462.x

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement, 16*, 247-260. doi: 10.1111/j.1745-3984.1979.tb00106.x

Lohman, D. F., & Hagen, E. P. (2002). *Cognitive abilities test. Form 6. Research Handbook*. Itasca, IL: Riverside Publishing.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179-193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139-160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling.* (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses. (ProQuest document ID: 1895633701).

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 25*, 373-383.

NCLB. (2001). Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200-219. Retrieved from http://www.jstor.org/stable/1435266

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359-368. doi: 10.1111/j.1745-3984.1980.tb00837.x

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156. Retrieved from http://www.jstor.org/stable/1164922

Phillips, S. E. (1986). The effects of the deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement, 23*(2), 107-118. doi: 10.1111/j.1745-3984.1986.tb00237.x

Pomplun, M., Omar, M. H., & Custer, M. (2004). A comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement, 64*(4), 600-616. doi: 10.1177/0013164403261761

Rentz, R. R., & Bashaw, W. L. (1977). The national reference scale for reading: An application of the Rasch model. *Journal of Educational Measurement, 14*(2), 161-179. doi: 10.1111/j.1745-3984.1977.tb00034.x

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10*(13), 1-4.

SAS Institute Inc. (2009). *SAS,* (Version 9.2). Cary, NC.

SAS Institute Inc. (n.d.). SAS(R) 9.2 Language Reference: Dictionary, Fourth Edition Retrieved March 29, 2011, from http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a001466748.htm

Shen, L. (1993). *Constructing a measure for longitudinal medical achievement studies by the Rasch model one-step equating*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*(1), 69-82. doi: 10.1177/014662168801200107

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement, 15*(1), 23-35. doi: 10.1111/j.1745-3984.1978.tb00053.x

Slinde, J. A., & Linn, R. L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement, 16*(3), 159-165. doi: 10.1111/j.1745-3984.1979.tb00097.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. doi: 10.1177/014662168300700208

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265-275. doi: 10.1111/j.1745-3984.1976.tb00017.x

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory*. [Computer software]. Chicago, IL: Scientific Software Inc.

Thurstone, L. L. (1925). A method of scaling psychological and educational test. *The Journal of Educational Psychology*, *16*(7), 433-451. Retrieved from http://www.apa.org/journals/edu/

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, No. 1*, Retrieved from http://www.springerlink.com/content/5p03860068725802/.

Tong, Y. (2005). *Comparisons of methodologies and results in vertical scaling for educational achievement tests.* (Doctoral Dissertation) Retrieved from ProQuest Dissertations and Theses. (ProQuest document ID: 913518251).

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*(2), 227-253. doi: 10.1080/08957340701301207

U.S. Department of Education. Lead & manage my school. Letters to chief state school officiers regarding an update on several NCLB cornerstones, Accessed in 2009 Retrieved from http://www2.ed.gov/admins/lead/account/cornerstones/index.html

U.S. Department of Education. Lead and manage my school: The standard and assessments peer review Program Overview, Accessed in 2011 Retrieved from http://www2.ed.gov/admins/lead/account/peerreview/index.html

U.S. Department of Education. (2003). No child left behind, accountability and adequate yearly progress (AYP), Accessed in 2011 Retrieved, from http://www2.ed.gov/admins/lead/account/ayp203/edlite-index.html

U.S. Department of Education. (January, 2009). Stronger accountability: Growth models, Accessed in 2009 Retrieved, from http://www.ed.gov/admins/lead/account/growthmodel/index.html

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*, 141-162. doi: 10.1111/j.1745-3984.2000.tb01080.x

Whitely, S. E., & Dawis, R. V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, *11*(3), 163-178. doi: 10.1111/j.1745-3984.1974.tb00988.x

Wilcox, R. R. (1977). Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. *Journal of Educational Statistics, 2*, 289-307. Retrieved from http://www.jstor.org/stable/1164810

Wilcox, R. R. (1981). Solving measurement problems with an answer-until-correct procedure. *Applied Psychological Measurement, 5*, 399-414. doi: 10.1177/014662168100500313

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*(2), 93-107. doi: 10.1111/j.1745-3984.1998.tb00529.x

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299-325. doi: 10.1111/j.1745-3984.1986.tb00252.x

Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*(4), 293-313. doi: 10.1111/j.1745-3984.1997.tb00520.x

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG*. [Computer software]. Chicago, IL: Scientific Software International Inc.

Appendix A: BILOG-MG and ICL command files[2]

[2] The example syntax file for ICL was included when you install the program by the creator of the ICL program and. The comments in the ICL file were created by the author of the program. The syntax files from BILOG-MG were created by Andrea J. Gotzmann and no additional comments are provided.

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'mathgr3_3.dat',
      NPARM = 3,
      SAVE;
>SAVE PARM = 'mathgr3_3.PAR',
     SCORE = 'mathgr3_3.SCO',
     PDISTRIB='mathgr3_3.PPD',
     EXPECTED='mathgr3_3.exp';
>LENGTH NITEMS = (42);
>INPUT  NTOTAL=42, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST1 TNAME = TEST1,;
>FORM1 LENGTH=21, INUMBER = (1(1)21);
>FORM2 LENGTH=21, INUMBER = (22(1)42);
(7A1,1X,I1,1X,21A1)
>CALIB NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR, PLOT=0.05,
NOFLOAT;
>SCORE METHOD=2, INFO=2, NOPRINT;
```

Figure A1. BILOG-MG initial run sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>COMMENT
>GLOBAL  DFNAME='rep1.dat', NPARM=3, SAVE;
>SAVE    SCORE='rep1.sco', PARM='rep1.par';
>LENGTH  NITEMS=207;
>INPUT   NTOT=207, NGROUPS=6, NFORMS=6, NIDCH=7;
>ITEMS   INUM=(1(1)207), INAME=(ITEM0001(1)ITEM00207);
>TEST1    TNAME=TEST1;
>FORM1 LENGTH=40, INUM=(1(1)40);
>FORM2 LENGTH=34, INUM=(41(1)74);
>FORM3 LENGTH=34, INUM=(75(1)108);
>FORM4 LENGTH=34, INUM=(109(1)142);
>FORM5 LENGTH=31, INUM=(143(1)173);
>FORM6 LENGTH=34, INUM=(174(1)207);
>GROUP1  GNAME='GRADE 3', LENGTH=40, INUM=(1(1)40);
>GROUP2  GNAME='GRADE 4', LENGTH=74, INUM=(1(1)74);
>GROUP3  GNAME='GRADE 5', LENGTH=68, INUM=(41(1)108);
>GROUP4  GNAME='GRADE 6', LENGTH=68, INUM=(75(1)142);
>GROUP5  GNAME='GRADE 7', LENGTH=65, INUM=(109(1)173);
>GROUP6  GNAME='GRADE 8', LENGTH=65, INUM=(143(1)207);
(7A1,I1,I1,1X,40A1)
>CALIB   NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR, PLOT=0.01,
NOFLOAT,
 REFERENCE=4;
>SCORE   METHOD=2, INFO=2, NOPRINT;
```

Figure A2. BILOG-MG Concurrent Calibration sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat', PRNAME='rep1.PRN',
      NPARM = 3,
      SAVE;
>SAVE PARM = 'rep1_1.PAR', SCORE='rep1.SCO', PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (74);
>INPUT  NTOTAL=74, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST TNAME = TEST1, FIX=(0(0)40, 1(0)34);
>FORM1 LENGTH=40, INUM=(1(1)40);
>FORM2 LENGTH=34, INUM=(41(1)74);
(7A1,1X,I1,1X,40A1)
>CALIB IDIST=1, NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR,
PLOT=0.01,
NOFLOAT, NOADJUST;
>QUAD POINTS=( -5.45000  -5.24487  -5.03974  -4.83462  -4.62949
             -4.42436  -4.21923  -4.01410  -3.80897  -3.60385
             -3.39872  -3.19359  -2.98846  -2.78333  -2.57821
             -2.37308  -2.16795  -1.96282  -1.75769  -1.55256
             -1.34744  -1.14231  -0.93718  -0.73205  -0.52692
             -0.32179  -0.11667   0.08846   0.29359   0.49872
              0.70385   0.90897   1.11410   1.31923   1.52436
              1.72949   1.93462   2.13974   2.34487   2.55000),
 WEIGHTS=( 0.000027723  0.000064470  0.000140651  0.000290691
0.000574843
  0.0010854  0.0019561  0.0033679  0.0055457  0.0087288
  0.0131277  0.0188542  0.0258582  0.0338773  0.0424667
  0.0511115  0.0593831  0.0669830  0.0734978  0.0783385
  0.0808699  0.0804123  0.0764270  0.0689102  0.0586244
  0.0469372  0.0353650  0.0251195  0.0168614  0.0107321
  0.0064973  0.0037482  0.0020676  0.0011046  0.0005539
  0.0002717  0.0001253  0.0000581  0.0000227  0.0000001);
>SCORE METHOD=2, IDIST=3, INFO=2, NOPRINT;
```

Figure A3. BILOG-MG Fixed Common Item Parameter Reading grade 4 sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat', PRNAME='rep1.PRN',
       NPARM = 3,
       SAVE;
>SAVE PARM = 'rep1_1.PAR', SCORE='rep1.SCO', PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (68);
>INPUT  NTOTAL=68, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST TNAME = TEST1, FIX=(0(0)34, 1(0)34);
>FORM1 LENGTH=34, INUM=(1(1)34);
>FORM2 LENGTH=34, INUM=(35(1)68);
(7A1,1X,I1,1X,34A1)
>CALIB IDIST=1, NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR,
PLOT=0.01,
NOFLOAT, NOADJUST;
>QUAD POINTS=( -4.75000  -4.54487  -4.33974  -4.13462  -3.92949
               -3.72436  -3.51923  -3.31410  -3.10897  -2.90385
               -2.69872  -2.49359  -2.28846  -2.08333  -1.87821
               -1.67308  -1.46795  -1.26282  -1.05769  -0.85256
               -0.64744  -0.44231  -0.23718  -0.03205   0.17308
                0.37821   0.58333   0.78846   0.99359   1.19872
                1.40385   1.60898   1.81410   2.01923   2.22436
                2.42949   2.63462   2.83974   3.04487   3.25000),
  WEIGHTS=(0.000031527  0.000071514  0.000154124  0.000315784
0.000617043
  0.0011518  0.0020548  0.0035034  0.0057079  0.0088852
  0.0132130  0.0187672  0.0254931  0.0332121  0.0416366
  0.0504106  0.0591811  0.0675524  0.0748961  0.0802781
  0.0826727  0.0813343  0.0761087  0.0675557  0.0568206
  0.0453017  0.0342853  0.0246775  0.0169271  0.0110855
  0.0069410  0.0041554  0.0023790  0.0013164  0.0006774
  0.0003503  0.0001572  0.0000768  0.0000322  0.0000129);
>SCORE METHOD=2, IDIST=3, INFO=2, NOPRINT;
```

Figure A4. BILOG-MG Fixed Common Item Parameter Reading grade 5 sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat', PRNAME='rep1.PRN',
      NPARM = 3,
      SAVE;
>SAVE PARM = 'rep1_1.PAR', SCORE='rep1.SCO',PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (65);
>INPUT  NTOTAL=65, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST TNAME = TEST1, FIX=(1(0)34);
>FORM1 LENGTH=34, INUM=(1(1)34);
>FORM2 LENGTH=31, INUM=(35(1)65);
(7A1,1X,I1,1X,34A1)
>CALIB IDIST=1, NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR,
PLOT=0.01,
NOFLOAT, NOADJUST;
>QUAD POINTS=(-3.30000  -3.09487  -2.88974  -2.68462  -2.47949
            -2.27436  -2.06923  -1.86410  -1.65897  -1.45385
            -1.24872  -1.04359  -0.83846  -0.63333  -0.42821
            -0.22308  -0.01795   0.18718   0.39231   0.59744
             0.80256   1.00769   1.21282   1.41795   1.62308
             1.82821   2.03333   2.23846   2.44359   2.64872
             2.85385   3.05897   3.26410   3.46923   3.67436
             3.87949   4.08462   4.28974   4.49487   4.70000),
  WEIGHTS=(0.000027433  0.000061894  0.000133515  0.000275347
0.000542707
  0.0010220  0.0018381  0.0031578  0.0051856  0.0081452
  0.0122337  0.0175594  0.0241190  0.0318221  0.0405119
  0.0499454  0.0596718  0.0689398  0.0767399  0.0819904
  0.0838033  0.0817278  0.0759005  0.0670518  0.0563451
  0.0450813  0.0343919  0.0250566  0.0174554  0.0116396
  0.0074282  0.0045318  0.0026615  0.0014709  0.0007999
  0.0003909  0.0001979  0.0000870  0.0000375  0.0000172);
>SCORE METHOD=2, IDIST=3, INFO=2, NOPRINT;
```

Figure A5. BILOG-MG Fixed Common Item Parameter Reading grade 7 sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat', PRNAME='rep1.PRN',
      NPARM = 3,
      SAVE;
>SAVE PARM = 'rep1_1.PAR', SCORE='rep1.SCO', PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (65);
>INPUT  NTOTAL=65, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST TNAME = TEST1, FIX=(1(0)31);
>FORM1 LENGTH=31, INUM=(1(1)31);
>FORM2 LENGTH=34, INUM=(32(1)65);
(7A1,1X,I1,1X,34A1)
>CALIB IDIST=1, NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR,
PLOT=0.01,
NOFLOAT, NOADJUST;
>QUAD POINTS=(  -2.50000  -2.29487  -2.08974  -1.88462  -1.67949
             -1.47436  -1.26923  -1.06410  -0.85897  -0.65385
             -0.44872  -0.24359  -0.03846   0.16667   0.37179
              0.57692   0.78205   0.98718   1.19231   1.39744
              1.60256   1.80769   2.01282   2.21795   2.42308
              2.62821   2.83333   3.03846   3.24359   3.44872
              3.65385   3.85897   4.06410   4.26923   4.47436
              4.67949   4.88462   5.08974   5.29487   5.50000),
  WEIGHTS=(0.000012639  0.000030242  0.000069700  0.000154861
0.000331192
  0.0006792  0.0013265  0.0024482  0.0042498  0.0069453
  0.0107421  0.0158056  0.0222240  0.0299806  0.0389455
  0.0488494  0.0591941  0.0691437  0.0775419  0.0831606
  0.0850526  0.0828407  0.0768270  0.0678876  0.0572213
  0.0460658  0.0354637  0.0261337  0.0184497  0.0124864
  0.0081005  0.0050402  0.0030066  0.0017198  0.0009377
  0.0004967  0.0002421  0.0001164  0.0000550  0.0000228);
>SCORE METHOD=2, IDIST=3, INFO=2, NOPRINT;
```

Figure A6. BILOG-MG Fixed Common Item Parameter Reading grade 8 sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat', PRNAME='rep1.PRN',
      NPARM = 3,
      SAVE;
>SAVE PARM = 'rep1_1.PAR', SCORE='rep1.SCO', PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (91);
>INPUT  NTOTAL=91, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST TNAME = TEST1, FIX=(0(0)42, 1(0)49);
>FORM1 LENGTH=42, INUM=(1(1)42);
>FORM2 LENGTH=49, INUM=(43(1)91);
(7A1,1X,I1,1X,49A1)
>CALIB IDIST=1, NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR,
PLOT=0.01,
 NOFLOAT, NOADJUST;
>QUAD POINTS=( -6.19000  -5.98487  -5.77974  -5.57462  -5.36949
             -5.16436  -4.95923  -4.75410  -4.54897  -4.34385
             -4.13872  -3.93359  -3.72846  -3.52333  -3.31821
             -3.11308  -2.90795  -2.70282  -2.49769  -2.29256
             -2.08744  -1.88231  -1.67718  -1.47205  -1.26692
             -1.06179  -0.85667  -0.65154  -0.44641  -0.24128
             -0.03615   0.16897   0.37410   0.57923   0.78436
              0.98949   1.19462   1.39974   1.60487   1.81000),
 WEIGHTS=( 0.000028396  0.000064789  0.000141006  0.000292593
0.000580170
  0.0010993  0.0019904  0.0034393  0.0056623  0.0088704
  0.0132243  0.0187992  0.0255623  0.0333466  0.0418392
  0.0506028  0.0591395  0.0669428  0.0735087  0.0783312
  0.0808283  0.0803249  0.0763033  0.0688420  0.0587864
  0.0474813  0.0362603  0.0260484  0.0174268  0.0108199
  0.0062871  0.0034657  0.0018377  0.0009412  0.0004700
  0.0002265  0.0001076  0.0000471  0.0000213  0.0000001);
>SCORE METHOD=2, IDIST=3, INFO=2, NOPRINT;
```

Figure A7. BILOG-MG Fixed Common Item Parameter Mathematics grade 4 sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat', PRNAME='rep1.PRN',
      NPARM = 3,
      SAVE;
>SAVE PARM = 'rep1_1.PAR', SCORE='rep1.SCO', PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (98);
>INPUT  NTOTAL=98, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST TNAME = TEST1, FIX=(0(0)49, 1(0)49);
>FORM1 LENGTH=49, INUM=(1(1)49);
>FORM2 LENGTH=49, INUM=(50(1)98);
(7A1,1X,I1,1X,49A1)
>CALIB IDIST=1, NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR,
PLOT=0.01,
 NOFLOAT, NOADJUST;
>QUAD POINTS=( -5.15000  -4.94487  -4.73974  -4.53462  -4.32949
             -4.12436  -3.91923  -3.71410  -3.50897  -3.30385
             -3.09872  -2.89359  -2.68846  -2.48333  -2.27821
             -2.07308  -1.86795  -1.66282  -1.45769  -1.25256
             -1.04744  -0.84231  -0.63718  -0.43205  -0.22692
             -0.02179   0.18333   0.38846   0.59359   0.79872
              1.00385   1.20897   1.41410   1.61923   1.82436
              2.02949   2.23462   2.43974   2.64487   2.85000),
  WEIGHTS=(0.000027603  0.000063689  0.000139567  0.000291426
0.000580207
  0.0011027  0.0019982  0.0034474  0.0056616  0.0088449
  0.0131578  0.0186727  0.0253472  0.0329944  0.0412887
  0.0498153  0.0581390  0.0658144  0.0723685  0.0772745
  0.0800072  0.0801464  0.0772805  0.0710208  0.0615299
  0.0498343  0.0375264  0.0262135  0.0170102  0.0103026
  0.0058624  0.0031558  0.0016175  0.0007937  0.0003745
  0.0001704  0.0000747  0.0000318  0.0000132  0.0000001);
>SCORE METHOD=2, IDIST=3, INFO=2, NOPRINT;
```

Figure A8. BILOG-MG Fixed Common Item Parameter Mathematics grade 5 sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat', PRNAME='rep1.PRN',
      NPARM = 3,
      SAVE;
>SAVE PARM = 'rep1_1.PAR', SCORE='rep1.SCO', PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (96);
>INPUT  NTOTAL=96, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST TNAME = TEST1, FIX=(1(0)47);
>FORM1 LENGTH=47, INUM=(1(1)47);
>FORM2 LENGTH=49, INUM=(48(1)96);
(7A1,1X,I1,1X,49A1)
>CALIB IDIST=1, NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR,
PLOT=0.01,
 NOFLOAT, NOADJUST;
>QUAD POINTS=(-3.28000  -3.07487  -2.86974  -2.66462  -2.45949
            -2.25436  -2.04923  -1.84410  -1.63897  -1.43385
            -1.22872  -1.02359  -0.81846  -0.61333  -0.40821
            -0.20308   0.00205   0.20718   0.41231   0.61744
             0.82256   1.02769   1.23282   1.43795   1.64308
             1.84821   2.05333   2.25846   2.46359   2.66872
             2.87385   3.07897   3.28410   3.48923   3.69436
             3.89949   4.10462   4.30974   4.51487   4.72000),
  WEIGHTS=(0.000019093  0.000045866  0.000105087  0.000229192
0.000477669
  0.0009464  0.0017741  0.0031363  0.0052354  0.0082766
  0.0124350  0.0178100  0.0244152  0.0321528  0.0407164
  0.0495775  0.0581801  0.0660846  0.0728586  0.0779909
  0.0808695  0.0808415  0.0774588  0.0706937  0.0610247
  0.0494753  0.0375046  0.0265709  0.0176491  0.0110509
  0.0065645  0.0037231  0.0020262  0.0010625  0.0005381
  0.0002637  0.0001239  0.0000565  0.0000251  0.0000104);
>SCORE METHOD=2, IDIST=3, INFO=2, NOPRINT;
```

Figure A9. BILOG-MG Fixed Common Item Parameter Mathematics grade 7 sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat', PRNAME='rep1.PRN',
       NPARM = 3,
       SAVE;
>SAVE PARM = 'rep1_1.PAR', SCORE='rep1.SCO', PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (98);
>INPUT  NTOTAL=98, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST TNAME = TEST1, FIX=(1(0)49);
>FORM1 LENGTH=49, INUM=(1(1)49);
>FORM2 LENGTH=49, INUM=(50(1)98);
(7A1,1X,I1,1X,49A1)
>CALIB IDIST=1, NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR,
PLOT=0.01,
 NOFLOAT, NOADJUST;
>QUAD POINTS=(  -1.99000  -1.78487  -1.57974  -1.37462  -1.16949
                -0.96436  -0.75923  -0.55410  -0.34897  -0.14385
                 0.06128   0.26641   0.47154   0.67667   0.88179
                 1.08692   1.29205   1.49718   1.70231   1.90744
                 2.11256   2.31769   2.52282   2.72795   2.93308
                 3.13821   3.34333   3.54846   3.75359   3.95872
                 4.16385   4.36897   4.57410   4.77923   4.98436
                 5.18949   5.39462   5.59974   5.80487   6.01000),
  WEIGHTS=(0.000024297  0.000057703  0.000129812  0.000276630
0.000559936
  0.0010748  0.0019569  0.0033821  0.0055521  0.0086540
  0.0128047  0.0180457  0.0243997  0.0318283  0.0401262
  0.0489599  0.0578417  0.0661600  0.0733693  0.0789577
  0.0821299  0.0820336  0.0789577  0.0705298  0.0599264
  0.0477699  0.0357942  0.0253593  0.0171034  0.0110451
  0.0068547  0.0040989  0.0023621  0.0013125  0.0006992
  0.0003615  0.0001750  0.0000824  0.0000385  0.0000154);
>SCORE METHOD=2, IDIST=3, INFO=2, NOPRINT;
```

Figure A10. BILOG-MG Fixed Common Item Parameter Mathematics grade 8 sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL DFNAME = 'rep1.dat',
       NPARM = 3,
       SAVE;
>SAVE PARM = 'rep1.PAR',
      SCORE = 'rep1.SCO', PDISTRIB='REP1.PPD';
>LENGTH NITEMS = (74);
>INPUT  NTOTAL=74, NFORM=2, NALT=4, NIDCHAR=7;
>ITEMS;
>TEST1 TNAME = TEST1,;
>FORM1 LENGTH=40, INUMBER = (1(1)40);
>FORM2 LENGTH=34, INUMBER = (41(1)74);
(7A1,1X,I1,1X,40A1)
>CALIB NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR, PLOT=0.01,
NOFLOAT;
>SCORE METHOD=2, INFO=2, NOPRINT;
```

Figure A11. BILOG-MG Test Characteristic Curve sample syntax file

```
BILOG-MG syntax
25/01/2010 12:42:30 PM
>GLOBAL  DFNAME='rep1.dat', NPARM=3, SAVE;
>SAVE    SCORE='rep1.sco', PARM='rep1.par';
>LENGTH  NITEMS=145;
>INPUT   NTOT=145, NGROUPS=2, NFORMS=3, NIDCH=7;
>ITEMS   INUM=(1(1)145), INAME=(ITEM0001(1)ITEM0145);
>TEST1    TNAME=TEST1;
>FORM1 LENGTH=47, INUM=(1(1)47);
>FORM2 LENGTH=49, INUM=(48(1)96);
>FORM3 LENGTH=49, INUM=(97(1)145);
>GROUP1  GNAME='GRADE 7', LENGTH=96, INUM=(1(1)96);
>GROUP2  GNAME='GRADE 8', LENGTH=98, INUM=(48(1)145);
(7A1,I1,I1,1X,49A1)
>CALIB   NQPT=40, CYCLES=500, CRIT=0.01, TPRIOR, PLOT=0.01,
NOFLOAT,
 REFERENCE=1;
>SCORE   METHOD=2, INFO=2, NOPRINT;
```

Figure A12. BILOG-MG Hybrid Test Characteristic Curve sample syntax file

```
output -log_file rep1.log
allocate_items_dist 40
options -default_prior_a none
options -default_prior_b none
options -default_prior_c none

read_examinees rep1.dat 40i1
read_item_param rep1.par
set estep [new_estep]
estep_compute $estep 1 1
delete_estep $estep

set eapfile [open rep1.theta w]

# Write EAP and MLE estimates and number correct for each examinee on
# a separate line of the output file
for {set i 1} {$i <= [num_examinees]} {incr i} {
        # compute number correct
        set resp [examinee_responses $i]
        set numcorrect 0
        foreach r $resp {
                if {$r > 0} then {incr numcorrect}
        }
                # get examinee posterior mean (EAP estimate)
        set eap [examinee_posterior_mean $i]

        # get examinee MLE estimate
        set mle [examinee_theta_MLE $i -10.0 10.0]
        # Write EAP and MLE estimates and number correct. The first
        # argument to the format command indicates that the second and
        # third arguments to the format command will be written as
        # floating-point numbers with 6 digits after the decimal point and
        # that the fourth argument will be written as an integer, with
        # a tab character separating the numbers.
        puts $eapfile [format "%.6f\t%.6f\t%d" $eap $mle $numcorrect]
}

# close output file
close $eapfile

# end of run
release_items_dist
```

Figure A13. ICL theta scoring sample syntax file

Appendix B: Results Tables Scale Score Mean and Standard Deviations, Decision Accuracy and Consistency[3], Conditional Standard Error, and Root-Mean-Squared-Difference

[3] The decision accuracy and decision consistency minimum value for the present study was set at 0.80, as 0.80 represents 80% both accurate and consistent classification and the values below 0.80 are bolded in the tables.

Table B1.

Scale Score mean and standard deviations for Reading normal conditions

	1,500							
	CC		FCIP		TCC		HCC	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	407.64	42.33	400.02	42.66	390.94	43.25	394.92	41.17
4	442.36	41.06	434.85	41.68	428.67	42.33	430.92	41.68
5	467.87	45.13	463.47	44.93	463.46	46.96	464.02	45.70
6	495.92	46.42	499.97	46.35	499.97	46.35	499.20	46.62
7	521.06	43.51	529.13	43.77	529.75	42.35	527.96	42.84
8	551.07	38.17	559.85	38.23	557.68	36.59	558.65	37.33
	3,000							
	CC		FCIP		TCC		HCC	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	407.39	42.10	399.31	42.93	391.13	42.96	395.33	40.66
4	442.46	40.77	434.43	41.76	428.99	41.90	431.14	41.26
5	468.22	44.76	463.77	44.65	463.72	46.76	464.31	45.33
6	495.79	46.54	500.02	46.36	500.02	46.36	499.19	46.69
7	520.77	43.90	529.18	43.79	529.39	42.52	527.27	42.98
8	550.60	39.02	559.87	38.34	559.89	36.67	557.59	37.99

Table B2.

Scale Score mean and standard deviations for Reading skewed conditions

	1,500							
	CC		FCIP		TCC		HCC	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	412.25	43.05	400.66	45.57	396.34	42.60	397.02	42.28
4	429.49	41.20	434.28	42.40	429.49	42.24	430.29	42.73
5	465.66	46.20	464.40	44.02	465.66	46.49	464.77	46.59
6	499.82	47.68	499.82	45.83	499.82	45.83	499.29	47.39
7	525.68	43.95	528.76	42.55	525.68	41.65	526.27	43.15
8	552.14	38.09	558.04	35.83	552.14	35.39	554.15	36.98

	3,000							
	CC		FCIP		TCC		HCC	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	411.25	43.02	400.74	45.84	397.14	42.23	397.61	41.76
4	443.02	41.01	433.74	42.44	429.91	41.81	430.45	42.20
5	469.45	45.96	464.73	43.65	466.12	46.30	464.83	46.22
6	495.28	47.79	499.76	45.82	499.76	45.82	499.22	47.48
7	517.06	44.52	529.02	42.47	525.04	42.14	525.63	43.46
8	543.55	38.62	559.68	35.42	552.34	35.55	554.17	37.19

Table B3

Scale Score mean and standard deviations for Mathematics normal conditions

				1,500				
	CC		FCIP		TCC		HCC	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	371.09	43.80	367.29	41.95	366.68	42.84	368.26	41.16
4	412.82	40.36	406.94	40.14	406.82	40.72	407.39	39.79
5	454.45	41.17	452.81	40.04	452.30	42.70	451.88	41.08
6	496.87	46.85	500.05	47.33	500.05	47.33	499.27	47.01
7	528.54	48.58	531.46	47.08	533.39	48.27	530.70	48.01
8	589.58	55.55	590.82	49.53	593.15	53.69	590.95	54.06
				3,000				
	CC		FCIP		TCC		HCC	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	370.10	43.53	367.09	41.89	367.55	42.58	368.14	40.90
4	412.08	40.39	407.44	40.32	407.57	40.51	407.13	39.77
5	454.42	41.27	453.20	40.13	452.87	42.59	452.38	41.05
6	497.01	46.93	499.85	47.32	499.85	47.32	499.07	47.12
7	529.14	48.65	531.59	47.25	533.28	48.42	531.25	47.92
8	590.76	55.40	591.77	49.10	592.71	54.18	591.31	53.95

Table B4

Scale Score mean and standard deviations for Mathematics skewed conditions

	1,500							
	CC		FCIP		TCC		HCC	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	371.40	44.90	368.71	42.91	373.67	40.64	368.65	42.10
4	406.73	40.56	404.34	40.22	406.73	38.22	404.35	39.73
5	451.47	41.82	450.81	38.75	451.47	42.02	450.95	41.62
6	499.90	48.10	499.90	47.32	499.90	47.32	498.55	47.88
7	535.22	50.19	532.24	46.82	535.22	49.17	532.95	49.35
8	597.15	55.99	593.68	47.14	597.15	53.73	595.74	54.39
	3,000							
	CC		FCIP		TCC		HCC	
Grade	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	372.44	44.36	370.30	42.80	375.00	39.67	368.71	41.64
4	410.81	40.09	407.03	39.90	407.43	37.34	404.31	39.29
5	453.64	41.71	450.75	38.67	451.49	41.86	450.63	41.62
6	496.18	48.27	499.69	47.40	499.69	47.40	498.52	47.92
7	528.77	50.79	532.95	46.91	534.34	49.85	531.36	49.11
8	593.81	56.76	597.78	46.16	596.63	54.88	592.53	54.69

Table B5.

Decision Accuracy and Consistency for Vertical Scaling for Reading, normal distribution, 3,000 sample size

Accuracy					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.820	0.827	**0.586**	**0.775**
	upper	0.893	0.911	0.931	0.933
4	lower	0.823	**0.750**	**0.798**	**0.755**
	upper	0.935	0.950	0.958	0.957
5	lower	**0.736**	**0.718**	**0.724**	**0.720**
	upper	0.941	0.953	0.944	0.949
6	lower	**0.675**	**0.687**	**0.687**	**0.683**
	upper	0.955	0.945	0.945	0.946
7	lower	**0.645**	**0.702**	**0.708**	**0.690**
	upper	0.988	0.978	0.980	0.982
8	lower	**0.644**	**0.723**	**0.727**	**0.704**
	upper	0.996	0.991	0.992	0.993

Consistency					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.824	**0.792**	**0.797**	0.802
	upper	0.936	0.945	0.954	0.956
4	lower	0.858	0.834	0.822	0.826
	upper	0.945	0.955	0.963	0.963
5	lower	0.856	0.847	0.846	0.848
	upper	0.938	0.948	0.939	0.944
6	lower	0.848	0.853	0.853	0.852
	upper	0.948	0.939	0.939	0.939
7	lower	0.813	0.830	0.830	0.826
	upper	0.980	0.966	0.970	0.972
8	lower	0.818	0.845	0.842	0.837
	upper	0.993	0.983	0.985	0.987

Table B6.

Decision Accuracy and Consistency for Vertical Scaling for Reading, skewed distribution, 1,500 sample size

Accuracy					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.840	0.873	0.814	0.830
	upper	0.869	0.903	0.935	0.936
4	lower	0.814	**0.764**	0.804	0.807
	upper	0.955	0.969	0.971	0.973
5	lower	**0.753**	**0.735**	**0.736**	**0.740**
	upper	0.951	0.968	0.956	0.962
6	lower	**0.696**	**0.702**	**0.702**	**0.702**
	upper	0.976	0.968	0.968	0.967
7	lower	**0.662**	**0.713**	**0.700**	**0.696**
	upper	0.997	0.992	0.994	0.993
8	lower	**0.626**	**0.733**	**0.690**	**0.700**
	upper	1.000	0.999	0.999	0.999

Consistency					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.885	0.862	0.861	0.866
	upper	0.902	0.917	0.938	0.936
4	lower	0.897	0.876	0.861	0.868
	upper	0.938	0.957	0.964	0.963
5	lower	0.877	0.864	0.866	0.867
	upper	0.930	0.954	0.937	0.943
6	lower	0.869	0.873	0.873	0.875
	upper	0.958	0.949	0.949	0.946
7	lower	0.827	0.851	0.843	0.847
	upper	0.994	0.985	0.989	0.987
8	lower	0.812	0.863	0.839	0.850
	upper	0.999	0.997	0.999	0.998

Table B7.

Decision Accuracy and Consistency for Vertical Scaling for Reading, skewed distribution, 3,000 sample size

		Accuracy			
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.820	0.866	0.859	0.868
	upper	0.875	0.900	0.934	0.938
4	lower	0.811	**0.762**	**0.791**	**0.793**
	upper	0.959	0.969	0.971	0.975
5	lower	**0.752**	**0.734**	**0.732**	**0.736**
	upper	0.952	0.968	0.954	0.962
6	lower	**0.696**	**0.700**	**0.700**	**0.700**
	upper	0.975	0.967	0.967	0.966
7	lower	**0.657**	**0.715**	**0.686**	**0.691**
	upper	0.997	0.992	0.995	0.993
8	lower	**0.622**	**0.746**	**0.685**	**0.703**
	upper	1.000	0.998	0.999	0.999
		Consistency			
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.883	0.862	0.862	0.866
	upper	0.904	0.917	0.939	0.938
4	lower	0.896	0.874	0.861	0.868
	upper	0.941	0.957	0.965	0.965
5	lower	0.876	0.863	0.866	0.867
	upper	0.930	0.954	0.935	0.943
6	lower	0.869	0.872	0.872	0.874
	upper	0.958	0.948	0.948	0.945
7	lower	0.827	0.851	0.842	0.845
	upper	0.994	0.984	0.990	0.987
8	lower	0.811	0.867	0.840	0.849
	upper	0.999	0.997	0.999	0.997

Table B8.

Conditional Standard Error for Reading, normal distribution, 3,000 sample size

Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.229	0.239	0.251	0.227
	upper	0.414	0.434	0.495	0.474
4	lower	0.209	0.227	0.225	0.220
	upper	0.388	0.441	0.470	0.469
5	lower	0.289	0.290	0.319	0.304
	upper	0.698	0.711	0.705	0.723
6	lower	0.316	0.305	0.305	0.310
	upper	0.654	0.651	0.651	0.646
7	lower	0.420	0.371	0.350	0.392
	upper	0.882	0.752	0.728	0.817
8	lower	0.514	0.464	0.410	0.465
	upper	1.041	0.953	0.855	0.934

Table B9.

Conditional Standard Error for Reading, skewed distribution, 1,500 sample size

Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.217	0.248	0.230	0.229
	upper	0.461	0.557	0.591	0.567
4	lower	0.207	0.240	0.236	0.238
	upper	0.429	0.603	0.641	0.595
5	lower	0.343	0.330	0.369	0.366
	upper	0.774	0.883	0.862	0.843
6	lower	0.377	0.354	0.354	0.360
	upper	0.743	0.816	0.816	0.749
7	lower	0.487	0.421	0.418	0.460
	upper	1.153	1.002	1.013	1.067
8	lower	0.627	0.534	0.525	0.570
	upper	1.398	1.265	1.215	1.224

Table B10.

Conditional Standard Error for Reading, skewed distribution, 3,000 sample size

Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.218	0.252	0.226	0.224
	upper	0.473	0.581	0.601	0.570
4	lower	0.205	0.244	0.233	0.233
	upper	0.429	0.630	0.647	0.585
5	lower	0.345	0.330	0.375	0.367
	upper	0.798	0.927	0.893	0.875
6	lower	0.382	0.366	0.366	0.373
	upper	0.742	0.869	0.869	0.782
7	lower	0.501	0.427	0.437	0.476
	upper	1.161	1.020	1.054	1.091
8	lower	0.633	0.524	0.545	0.582
	upper	1.354	1.237	1.235	1.218

Table B11.

Root-Mean-Squared-Difference between Vertical Scaling procedures for Reading, normal distribution, 3,000 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	9.102	16.620	12.368	9.219	5.695	5.393
4	8.693	13.810	11.674	6.605	5.027	2.472
5	4.897	5.571	4.039	4.252	2.125	2.980
6	4.710	4.710	4.001	0.000	1.097	1.097
7	8.956	9.163	6.759	1.982	3.563	3.358
8	9.606	9.769	7.575	2.742	3.735	3.613

Table B12.

Root-Mean-Squared-Difference between Vertical Scaling procedures for Reading, skewed distribution, 1,500 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	13.520	16.907	16.110	7.052	7.586	4.568
4	10.784	15.224	14.410	6.650	6.996	3.933
5	6.409	5.607	5.176	5.440	4.185	4.251
6	5.636	5.636	4.849	0.000	2.496	2.496
7	12.443	10.311	9.336	5.291	4.504	5.295
8	15.034	10.049	10.867	7.337	4.758	4.829

Table B13.

Root-Mean-Squared-Difference between Vertical Scaling procedures for Reading, skewed distribution, 3,000 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	12.041	14.766	14.044	6.546	6.428	2.764
4	10.280	14.000	13.139	5.570	4.988	2.170
5	6.121	5.211	4.774	5.320	3.707	3.880
6	5.564	5.564	4.592	0.000	2.395	2.395
7	12.662	9.085	9.003	4.308	5.158	3.759
8	16.700	9.713	11.523	7.661	7.101	4.578

Table B14.

Decision Accuracy and Consistency for Vertical Scaling for Mathematics, normal distribution, 3,000 sample size

		Accuracy			
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	**0.519**	**0.465**	**0.481**	**0.498**
	upper	0.920	0.934	0.930	0.937
4	lower	0.818	0.842	0.833	0.834
	upper	0.951	0.959	0.959	0.963
5	lower	0.818	0.818	**0.799**	0.806
	upper	0.921	0.930	0.918	0.929
6	lower	**0.673**	**0.683**	**0.683**	**0.680**
	upper	0.926	0.915	0.915	0.919
7	lower	**0.685**	**0.707**	**0.713**	**0.703**
	upper	0.929	0.928	0.916	0.926
8	lower	**0.670**	**0.698**	**0.686**	**0.677**
	upper	0.870	0.905	0.869	0.877

		Consistency			
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.816	0.815	0.814	0.816
	upper	0.947	0.955	0.953	0.957
4	lower	0.830	0.817	0.820	0.820
	upper	0.955	0.962	0.962	0.965
5	lower	0.884	0.881	0.873	0.879
	upper	0.937	0.944	0.937	0.942
6	lower	0.840	0.844	0.844	0.843
	upper	0.942	0.935	0.935	0.937
7	lower	0.843	0.848	0.849	0.847
	upper	0.939	0.939	0.932	0.937
8	lower	0.889	0.890	0.891	0.889
	upper	0.913	0.930	0.912	0.916

Table B15.

Decision Accuracy and Consistency for Vertical Scaling for Mathematics, skewed distribution, 1,500 sample size

Accuracy					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	**0.582**	**0.507**	**0.738**	**0.545**
	upper	0.912	0.930	0.930	0.935
4	lower	0.864	0.855	0.867	0.854
	upper	0.979	0.983	0.985	0.988
5	lower	0.802	0.805	**0.789**	**0.789**
	upper	0.945	0.963	0.936	0.954
6	lower	**0.694**	**0.692**	**0.692**	**0.694**
	upper	0.944	0.924	0.924	0.936
7	lower	**0.698**	**0.720**	**0.727**	**0.718**
	upper	0.940	0.946	0.919	0.936
8	lower	**0.705**	**0.726**	**0.725**	**0.715**
	upper	0.866	0.930	0.866	0.869
Consistency					
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.853	0.849	0.860	0.852
	upper	0.934	0.946	0.944	0.947
4	lower	0.869	0.853	0.858	0.860
	upper	0.965	0.972	0.976	0.979
5	lower	0.906	0.902	0.893	0.901
	upper	0.927	0.951	0.934	0.937
6	lower	0.873	0.870	0.870	0.874
	upper	0.928	0.926	0.926	0.924
7	lower	0.877	0.878	0.879	0.881
	upper	0.922	0.932	0.912	0.917
8	lower	0.911	0.910	0.913	0.912
	upper	0.877	0.922	0.879	0.880

Table B16.

Decision Accuracy and Consistency for Vertical Scaling for Mathematics, skewed distribution, 3,000 sample size

		Accuracy			
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	**0.615**	**0.559**	0.814	**0.545**
	upper	0.913	0.927	0.932	0.938
4	lower	0.847	0.872	0.863	0.865
	upper	0.978	0.979	0.985	0.989
5	lower	0.803	0.805	**0.789**	**0.786**
	upper	0.944	0.962	0.936	0.955
6	lower	**0.694**	**0.691**	**0.691**	**0.695**
	upper	0.944	0.924	0.924	0.936
7	lower	**0.688**	**0.723**	**0.719**	**0.709**
	upper	0.943	0.941	0.918	0.943
8	lower	**0.701**	**0.754**	**0.719**	**0.698**
	upper	0.866	0.914	0.860	0.888

		Consistency			
Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.856	0.853	0.865	0.853
	upper	0.933	0.944	0.946	0.949
4	lower	0.871	0.856	0.858	0.859
	upper	0.964	0.968	0.977	0.981
5	lower	0.906	0.901	0.892	0.900
	upper	0.926	0.951	0.935	0.913
6	lower	0.872	0.869	0.869	0.874
	upper	0.929	0.926	0.926	0.923
7	lower	0.875	0.878	0.878	0.880
	upper	0.924	0.929	0.912	0.924
8	lower	0.911	0.916	0.913	0.909
	upper	0.876	0.913	0.876	0.889

Table B17.

Conditional Standard Error for Mathematics normal distribution, 3,000 sample size

Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.284	0.271	0.278	0.256
	upper	0.491	0.503	0.491	0.490
4	lower	0.181	0.190	0.169	0.169
	upper	0.237	0.259	0.242	0.252
5	lower	0.212	0.217	0.207	0.214
	upper	0.335	0.361	0.327	0.346
6	lower	0.196	0.197	0.197	0.197
	upper	0.292	0.285	0.285	0.290
7	lower	0.229	0.204	0.210	0.219
	upper	0.365	0.347	0.340	0.345
8	lower	0.226	0.180	0.208	0.212
	upper	0.383	0.347	0.352	0.367

Table B18.

Conditional Standard Error for Mathematics skewed distribution, 1,500 sample size

Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.311	0.295	0.253	0.279
	upper	0.658	0.687	0.605	0.667
4	lower	0.183	0.204	0.183	0.174
	upper	0.294	0.362	0.373	0.378
5	lower	0.211	0.217	0.221	0.215
	upper	0.388	0.481	0.454	0.412
6	lower	0.207	0.238	0.238	0.215
	upper	0.347	0.446	0.446	0.381
7	lower	0.251	0.224	0.243	0.237
	upper	0.464	0.483	0.484	0.434
8	lower	0.260	0.199	0.246	0.242
	upper	0.490	0.502	0.496	0.468

Table B19.

Conditional Standard Error for Mathematics skewed distribution, 3,000 sample size

Grade	Cut-Score	CC	FCIP	TCC	HCC
3	lower	0.299	0.289	0.242	0.274
	upper	0.638	0.678	0.592	0.664
4	lower	0.181	0.203	0.180	0.168
	upper	0.292	0.363	0.383	0.386
5	lower	0.210	0.219	0.225	0.216
	upper	0.387	0.490	0.465	0.416
6	lower	0.208	0.242	0.242	0.213
	upper	0.346	0.451	0.451	0.366
7	lower	0.255	0.229	0.256	0.236
	upper	0.465	0.494	0.512	0.433
8	lower	0.264	0.188	0.261	0.247
	upper	0.490	0.474	0.516	0.486

Table B20.

Root-Mean-Squared-Difference between Vertical Scaling procedures for Mathematics, normal distribution, 3,000 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	10.300	4.179	4.237	10.475	10.058	3.219
4	6.095	5.624	6.018	4.947	4.934	1.708
5	3.257	3.641	2.350	4.568	2.845	3.282
6	3.660	3.660	3.093	0.000	1.077	1.077
7	4.817	5.457	3.037	2.687	3.673	3.878
8	7.569	3.412	3.102	6.101	6.078	2.427

Table B21.

Root-Mean-Squared-Difference between Vertical Scaling procedures for Mathematics, skewed distribution, 1,500 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	9.868	6.522	4.739	11.221	9.326	6.283
4	7.093	5.131	6.309	5.670	5.303	3.457
5	4.975	4.705	2.656	5.168	3.939	4.120
6	5.003	5.003	3.304	0.000	2.617	2.617
7	5.653	6.303	3.282	4.347	4.662	4.162
8	10.375	4.510	2.881	8.705	9.017	2.687

Table B22.

Root-Mean-Squared-Difference between Vertical Scaling procedures for Mathematics, skewed distribution, 3,000 sample size

Grade	CC vs FCIP	CC vs TCC	CC vs HCC	FCIP vs TCC	FCIP vs HCC	TCC vs HCC
3	11.154	6.645	5.074	12.295	10.811	7.389
4	5.873	5.517	7.281	5.310	5.893	4.210
5	5.159	4.833	3.237	5.022	4.001	4.175
6	5.072	5.072	3.344	0.000	2.738	2.738
7	6.783	6.727	3.608	3.615	4.506	4.645
8	12.148	4.156	3.347	9.689	10.822	4.572