

Smarter Balanced Assessment Consortium: 2013-14 Technical Report

- Validity
- Item and Test Development
 - Pilot Test and Field Test
- Achievement Level Setting

January 14, 2016



Chapter 1 Introduction	19
Overview	19
Technical Report Approach	21
NCLB Peer Review Guidelines and Established Standards	21
Overview and Background of the Smarter Balanced Theory of Action	23
Seven Principles of Smarter Balanced Underlying the Theory of Action	23
Purposes for the Smarter Balanced Assessment System	24
Figure 1. Overview of Smarter Balanced Theory of Action	26
Figure 2. Smarter Balanced Development Timeline	27
Chapter Overview.....	27
Acknowledgments.....	29
Smarter Balanced Work Groups	29
Outside Groups and Organizations that Collaborated with the Smarter Balanced Assessment System	29
2014 Technical Advisory Committee.....	29
Contributors to the Accessibility Accommodations Framework.....	30
Other Acknowledgments.	30
References.....	31
Chapter 2: Validity	33
Introduction.....	33
Essential Validity Elements for Summative and Interim Assessments.....	34
Table 1. Synopsis of Essential Validity Evidence Derived from <i>Standards</i> (AERA et al., 1999, p. 17).	35
Careful Test Construction.....	35
Adequate Measurement Precision (Reliability).....	36
Appropriate Test Administration.....	37
Appropriate Scoring	37
Accurate Scaling and Linking.....	37
Appropriate Standard Setting.....	38
Attention to Fairness, Equitable Participation, and Access.....	39
Validating “On-Track/Readiness”	39
Adequate Test Security.....	41
Summary of Essential Validity Evidence based on the Smarter Pilot- and Field Tests	41
Table 2. Essential Validity Evidence for the Summative and Interim Assessments for Careful Test Construction.....	42

Table 3. Essential Validity Evidence for the Summative and Interim Assessments for Adequate Measurement Precision (Reliability).	43
Table 4. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Test Administration.	44
Table 5. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Scoring.	44
Table 6. Essential Validity Evidence for the Summative and Interim Assessments for Accurate Scaling and Linking.	45
Table 7. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Standard Setting.	46
Table 8. Essential Validity Evidence for the Summative and Interim Assessments for Attention to Fairness, Equitable Participation and Access.	47
Table 9. Essential Validity Evidence for the Summative and Interim Assessments for Validating “On-Track/Readiness”	47
Table 10. Essential Validity Evidence for the Summative and Interim Assessments for Adequate Test Security.	48
The <i>Standards’</i> Five Primary Sources of Validity Evidence	48
Purposes of the Smarter Balanced System for Summative, Interim, and Formative Assessments	50
Table 11. Validity Framework for Smarter Balanced Summative Assessments	52
Table 12. Validity Framework for Smarter Balanced Interim Assessments.	53
Evidence Using the Five Primary Sources of Validity Framework	53
Table 13. Listing of Evidence Type, Evidence Source, and Primary Validity Source for Summative, Interim, and Formative Test Purposes.	54
Conclusion for Field Test Validity Results	60
References	61
American Institutes for Research (2014b). Smarter Balanced Scoring Specification: 2014–2015 Administration	61
Chapter 3: Item Development	65
Item and Task Specifications	65
The Item/task Pool	67
Item Writing	67
Training	68
Educator Participation	69
State-Managed Item Development	69

Item Reviews..... 69

Item Scoring..... 71

Composition of the Item Pool..... 71

 Table 1 Item Types found in the Consortium’s Item Pool. 71

 Table 2 Total Number of CAT Items and Performance Tasks (PT) developed by Grade and Content Area
..... 73

 Table 3. Mathematics Specifications and Archetype Delivery 74

 Table 4. English/Language Arts/Literacy Specifications and Archetype Delivery 75

Use of Systems and Tagging..... 75

Training Activities..... 76

 Virtual Training of Educator Item and Task Writers..... 77

 Continued Training during Development 77

 Certification and Management..... 77

 Table 5. Sample Certification Set..... 78

Item/Task Set Evaluation..... 78

 Role of Item Quality Review Panel..... 78

 Educator Recruitment Activities 79

 Table 6. Recruiting Activities and Timeline 80

Item development process 80

 Table 7. Additional item writing, development, review and scoring documentation 83

Chapter 4 Test Design..... 84

 Figure 1. Relationships among Smarter Balanced Content 84

A Brief Description of Smarter Balanced Content Structure..... 84

 Table 1. Major Domains Identified for ELA and Mathematics. 85

Synopsis of Assessment System Components..... 85

Evidence-Centered Design in Constructing Smarter Balanced Assessments 86

Content Alignment in Smarter Balanced Test Design 87

Test Blueprints..... 88

Summative Assessment..... 89

 Figure 2. Blueprint for grade 6 showing detailed content structure (Assessment Targets), page 1 of 2
..... 91

 Figure 3. Blueprint for grade 6 showing detailed content structure (Assessment Targets), page 2 of 2
..... 92

CAT and Performance Task Test Components	93
Operational Adaptive Test Design	93
Expansion of the Item Pool	94
Performance Task Design	95
Test Scoring	96
Field Test Delivery Modes	97
Measurement Models (IRT) Adopted.....	97
Interim Assessment.....	98
ELA/Literacy ICA Blueprints	99
Mathematics ICA Blueprints	99
Table 2. Summary of Interim Test Features for ICAs and IABs.....	101
Table 3. Summary of ELA Interim Assessment Blocks.....	102
Table 4. Summary of Mathematics Interim Assessment Blocks.....	103
Table 5. High School Mathematics Assessment Blocks	104
Pool analysis and adequacy: Background and Recommendations	104
Simulations Studies for 2014-15 operational summative tests	107
Test Design Specifications and Outcomes.....	108
References.....	113
Chapter 5 Test Fairness.....	115
Introduction.....	115
Definitions for Validity, Bias, Sensitivity, and Fairness.	115
The Smarter Balanced Accessibility and Accommodations Framework	117
Figure 1. Conceptual Framework for Usability, Accessibility, and Accommodations.	120
Meeting the Needs of Traditionally Underrepresented Populations.....	120
The Individual Student Assessment Accessibility Profile (ISAAP).....	121
Usability, Accessibility, and Accommodations Guidelines: Intended Audience and Recommended Applications.....	123
Guidelines for Accessibility for English Language Learners.	124
Fairness as a Lack of Measurement Bias: Differential Item Functioning Analyses	126
Test Fairness and Implications for Ongoing Research.....	126
Internal Structure.	127
Response Processes.	127
Relationships with Other Variables.	127
Test Consequences.....	128

References.....	129
Chapter 6 Pilot Test and Special Studies (Dimensionality Analysis and IRT Model Choice)	132
Pilot Data Collection Design.....	133
Table 1. Total Number of CAT Components and Performance Tasks (PT).....	134
Table 2. ELA/literacy Grades 3 to 10 Pilot Test CAT Component Blueprint.	135
Table 3. ELA/literacy Grade 11 Pilot Test CAT Component Blueprint.	135
Table 4. Mathematics Grades 3 to 11 Pilot Test CAT Component Blueprint.	136
Vertical Linking Item Assignment	136
Figure 1. Summary of Vertical Articulation of Test Content by Grade.....	137
Pilot Test Sampling Procedures.....	137
Sampling Consideration for the Pilot Test.....	137
Test Administration and Sample Size Requirements.....	138
Table 5. Targeted Student Sample Size by Content Area and Grade for the Pilot Test.	139
Pilot Sampling Considerations.....	139
Sampling Procedures	142
Table 6. Approximate Sample Sizes by Content Area and State, the Sample Target and the Number Obtained for the Pilot Test.	144
Table 7. ELA/literacy Student Population and Sample Characteristics (Percentages).	145
Table 8. Mathematics Student Population and Sample Characteristics (Percentages)	146
Pilot Classical Test Results	147
Pilot Classical Item Flagging Criteria	148
Description of Pilot Classical Statistics Evaluated	149
Pilot Results	150
Table 9. Summary of Number of Pilot Test Items and Students Obtained.	151
Table 10. Overview of ELA/literacy CAT Component Statistics.....	152
Table 11. Overview of Mathematics Component Statistics.	152
Table 12. Student Testing Durations in Days (Percentage Completion).	153
Table 13. Summary of Reliability and Difficulty for CAT Administrations.	154
Table 14. Description of Item Flagging for Selected-response Items.....	155
Table 15. Description of Item Flagging for Constructed-response Items.....	156
Table 16. Number of Items Flagged for ELA/literacy by Selected- and Constructed-response.	157
Table 17. Number of Items Flagged for Mathematics by Selected- and Constructed-response.	157

Table 18. Definition of Focal and Reference Groups..... 158

Table 19. DIF Categories for Selected-Response Items..... 159

Table 20. DIF Categories for Constructed-Response Items..... 159

Table 21. Number of DIF Items Flagged by Item Type and Subgroup (ELA/literacy, Grades 3 to 7).... 160

Table 22. Number of DIF Items Flagged by Item Type and Subgroup (ELA/literacy, Grades 8 to 11)... 161

Table 23. Number of C DIF Items Flagged by Item Type and Subgroup (Mathematics, Grades 3 to 7).162

Table 24. Number of C DIF Items Flagged by Item Type and Subgroup (Mathematics, Grades 8 to 11).
..... 163

Dimensionality Study.....164

 Rationale and Approach.....164

 Factor Models.....164

 Figure 2. An Example of the Bifactor Model with Four Minor Factors Corresponding to Claims. 165

 Table 25. Summary of MIRT Analysis Configuration Showing Number of Content, Grades and MIRT
 Models..... 166

MIRT Scaling Models.....166

Software and System Requirements.....167

Evaluation of the Number and Types of Dimensions and MIRT Item Statistics.....167

 Table 26. Models and Fit Measures for ELA/literacy Within Grade..... 169

 Table 27. Models and Fit Measures for Mathematics Within Grade..... 171

 Table 28. Models and Fit Measures for ELA/literacy Across Adjacent Grades. 173

 Table 29. Models and Fit Measures for Mathematics Across Adjacent Grades..... 177

MIRT Item Statistics and Graphs.....181

Discussion and Conclusion.....182

 Figure 3. Item Vector Plot for ELA/literacy Grade 3 (Within Grade)..... 184

 Figure 4. Item Vector Plot for ELA/literacy Grade 4 (Within Grade)..... 184

 Figure 5. Item Vector Plot for ELA/literacy Grade 5 (Within Grade)..... 185

 Figure 6. Item Vector Plot for ELA/literacy Grade 6 (Within Grade)..... 185

 Figure 7. Item Vector Plot for ELA/literacy Grade 7 (Within Grade)..... 186

 Figure 8. Item Vector Plot for ELA/literacy Grade 8 (Within Grade)..... 186

 Figure 9. Item Vector Plot for ELA/literacy Grade 9 (Within Grade)..... 187

 Figure 10. Item Vector Plot for ELA/literacy Grade 10 (Within Grade)..... 187

 Figure 11. Item Vector Plot for ELA/literacy Grade 11 (Within Grade)..... 188

 Figure 12. Item Vector Plot for ELA/literacy Grades 3 and 4 (Across Grades)..... 188

Figure 13. Item Vector Plot for ELA/literacy Grades 4 and 5 (Across Grades)	189
Figure 14. Item Vector Plot for ELA/literacy Grades 5 and 6 (Across Grades)	189
Figure 15. Item Vector Plot for ELA/literacy Grades 6 and 7 (Across Grades)	190
Figure 16. Item Vector Plot for ELA/literacy Grades 7 and 8 (Across Grades)	190
Figure 17. Item Vector Plot for ELA/literacy Grades 8 and 9 (Across Grades)	191
Figure 18. Item Vector Plot for ELA/literacy Grades 9 and 10 (Across Grades)	191
Figure 19. Item Vector Plot for ELA/literacy Grades 10 and 11 (Across Grades)	192
Figure 20. Item Vector Plots for the Subset of ELA/literacy Grades 3 and 4 Vertical Linking Items .	192
Figure 21. Item Vector Plots for the Subset of ELA/literacy Grades 4 and 5 Vertical Linking Items .	193
Figure 22. Item Vector Plots for the Subset of ELA/literacy Grades 5 and 6 Vertical Linking Items .	193
Figure 23. Item Vector Plots for the Subset of ELA/literacy Grades 6 and 7 Vertical Linking Items .	194
Figure 24. Item Vector Plots for the Subset of ELA/literacy Grades 7 and 8 Vertical Linking Items .	194
Figure 25. Item Vector Plots for the Subset of ELA/literacy Grades 8 and 9 Vertical Linking Items .	195
Figure 26. Item Vector Plots for the Subset of ELA/literacy Grades 9 and 10 Vertical Linking Items	195
Figure 27. Item Vector Plots for the Subset of ELA/literacy Grades 10 and 11 Vertical Linking Items	196
Figure 28. Item Vector Plot for Mathematics Grade 3 (Within Grade)	196
Figure 29. Item Vector Plot for Mathematics Grade 4 (Within Grade)	197
Figure 30. Item Vector Plot for Mathematics Grade 5 (Within Grade)	197
Figure 31. Item Vector Plot for Mathematics Grade 6 (Within Grade)	198
Figure 32. Item Vector Plot for Mathematics Grade 7 (Within Grade)	198
Figure 33. Item Vector Plot for Mathematics Grade 8 (Within Grade)	199
Figure 34. Item Vector Plot for Mathematics Grade 9 (Within Grade)	199
Figure 35. Item Vector Plot for Mathematics Grade 10 (Within Grade)	200
Figure 36. Item Vector Plot for Mathematics Grade 11 (Within Grade)	200
Figure 37. Item Vector Plot for Mathematics Grades 3 and 4 (Across Grades)	201
Figure 38. Item Vector Plot for Mathematics Grades 4 and 5 (Across Grades)	201
Figure 39. Item Vector Plot for Mathematics Grades 5 and 6 (Across Grades)	202
Figure 40. Item Vector Plot for Mathematics Grades 6 and 7 (Across Grades)	202
Figure 41. Item Vector Plot for Mathematics Grades 7 and 8 (Across Grades)	203
Figure 42. Item Vector Plot for Mathematics Grades 8 and 9 (Across Grades)	203
Figure 43. Item Vector Plot for Mathematics Grades 9 and 10 (Across Grades)	204

Figure 44. Item Vector Plot for Mathematics Grades 10 and 11 (Across Grades) 204

Figure 45. Item Vector Plot for the Subset of Mathematics Grades 3 and 4 (Vertical Linking Items)
..... 205

Figure 46. Item Vector Plot for the Subset of Mathematics Grades 4 and 5 (Vertical Linking Items)
..... 205

Figure 47. Item Vector Plot for the Subset of Mathematics Grades 5 and 6 (Vertical Linking Items)
..... 206

Figure 48. Item Vector Plot for the Subset of Mathematics Grades 6 and 7 (Vertical Linking Items)
..... 206

Figure 49. Item Vector Plot for the Subset of Mathematics Grades 7 and 8 (Vertical Linking Items)
..... 207

Figure 50. Item Vector Plot for the Subset of Mathematics Grades 8 and 9 (Vertical Linking Items)
..... 207

Figure 51. Item Vector Plot for the Subset of Mathematics Grades 9 and 10 (Vertical Linking Items)
..... 208

Figure 52. Item Vector Plot for the Subset of Mathematics Grades 10 and 11 (Vertical Linking Items)
..... 208

Figure 53. Clustering of Item Angle Measures for Grades 3 to 5, ELA/literacy (within grade) 209

Figure 54. Clustering of Item Angle Measures for Grades 6 to 8, ELA/literacy (within grade) 210

Figure 55. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (within grade) 211

Figure 56. Clustering of Item Angle Measures for Grades 3 to 6, ELA/literacy (across grades)..... 212

Figure 57. Clustering of Item Angle Measures for Grades 6 to 9, ELA/literacy (across grades)..... 213

Figure 58. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (across grades)..... 214

Figure 59. Clustering of Item Angle Measures for Grades 3 to 6, ELA/literacy (vertical linking) 215

Figure 60. Clustering of Item Angle Measures for Grades 6 to 9, ELA/literacy (vertical linking) 216

Figure 61. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (vertical linking) 217

Figure 62. Clustering of Item Angle Measures for Grades 3 to 5, Mathematics (within grade) 218

Figure 63. Clustering of Item Angle Measures for Grades 6 to 8, Mathematics (within grade) 219

Figure 64. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (within grade) 220

Figure 65. Clustering of Item Angle Measures for Grades 3 to 6, Mathematics (across grades)..... 221

Figure 66. Clustering of Item Angle Measures for Grades 6 to 9, Mathematics (across grades)..... 222

Figure 67. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (across grades).... 223

Figure 68. Clustering of Item Angle Measures for Grades 3 to 6, Mathematics (vertical linking) 224

Figure 69. Clustering of Item Angle Measures for Grades 6 to 9, Mathematics (vertical linking) 225

Figure 70. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (vertical linking) .. 226

Item Response Theory (IRT) Model Comparison227

IRT Data Step.....229

Table 30. Number of Items Dropped from the Calibration (On-grade). 231

Table 31. Number of Constructed-response and Performance tasks with Collapsed Score Levels (On-grade)..... 231

Table 32. Number of Constructed-response and Performance tasks with Collapsed Score Levels for ELA/literacy (Detail)..... 232

Table 33. Number of Constructed-response with Collapsed Score Levels for Mathematics (Detail)... 233

Table 34. Number of ELA/literacy and Mathematics Items in the IRT Calibration. 234

Table 35. Descriptive Statistics for Number of Students per Item for ELA/literacy and Mathematics. 235

IRT Model Calibration235

IRT Model Fit Comparison.....237

Table 36. Summary of G^2 Statistics of On-Grade ELA/literacy Items across 1PL, 2PL, and 3PL IRT Models. 238

Table 37. Summary of G^2 Statistics of On-Grade Mathematics Items across 1PL, 2PL, and 3PL IRT Models..... 238

Guessing Evaluation.....239

Table 38. Summary of Guessing Parameter Estimates for On-Grade ELA/literacy Items..... 239

Table 39. Summary of Guessing Parameter Estimates for On-Grade Mathematics Items..... 240

Common Discrimination Evaluation241

Table 40. Summary of 2PL/GPC Slope and Difficulty Estimates and Correlations for ELA/literacy. 242

Table 41. Summary of 2PL/GPC Slope and Difficulty Estimates and Correlations for Mathematics. ... 247

Evaluation of Ability Estimates.....251

Table 42. ELA/literacy Correlations of Ability Estimates across Different Model Combinations..... 251

Table 43. Mathematics Correlations of Ability Estimates across Different Model Combinations..... 253

IRT Model Recommendations.....255

Figure 71. Scatter Plot of ELA/literacy 2PL/GPC Slope and Difficulty Estimates by Item Type, Score Category and Claim..... 256

Figure 72. Scatter Plot of Mathematics 2PL/GPC Slope and Difficulty Estimates by Item Type, Score Category, and Claim..... 261

Figure 73. ELA/literacy Scatter Plots of Theta Estimates across Different Model Combinations 266

Figure 74. Mathematics Scatter Plots of Theta Estimates Across Different Model Combinations... 271

References.....	276
Chapter 7 Field Test Design, Sampling, and Administration	280
Introduction.....	280
Field Test Data Collection Design.....	281
Table 1. Field Test Data Collection Design for ELA/literacy and Mathematics.	282
Field Test Delivery Modes	282
CAT (LOFT) Administration.	283
Performance Task Administration	283
Field Test Design	284
Table 2. Field Test Design for the Item and Performance Task Pools.	286
Numbers and Characteristics of Items and Students Obtained in the Field Test.....	287
Table 3. Number of Field Test Vertical Scaling Items Obtained by Type for ELA/literacy.....	287
Table 4. Number of Field Test Vertical Scaling Items Obtained by Type for Mathematics.....	288
Table 5. Number of Field Test Item Pool Calibration (Overall Total) Obtained by Claim for ELA/literacy.	288
Table 6. Number of Field Test Item Pool Calibration (Overall Total) Obtained by Claim for Mathematics.	288
Figure 1. Distributions of Number of Items per Student in the Vertical Scaling (ELA/literacy)	290
Figure 2. Distributions of Number of Items per Student in the Vertical Scaling (Mathematics)	291
Figure 3. Distributions of Number of Items per Student in the Item Pool Calibration Sample (ELA/literacy)	292
Figure 4. Distributions of Number of Items per Student in the Item Pool Calibration Sample (Mathematics)	293
Linking PISA and NAEP Items onto the Smarter Balanced Assessments.....	294
Table 7. Comparison of Features across the Smarter Balanced, NAEP, and PISA Assessment Programs.	295
Field Test Student Sampling Design	296
Defining the Target Population	297
State Participation Conditions.....	298
Table 8. State Participation and Sample Acquisition Conditions.	299
Technical Sampling Characteristics.....	299
Detailed Sampling Procedures.....	300
Sampling Results.....	302

Table 9. Sample Size (Percent) for ELA/literacy and Mathematics by Grade and Smarter Balanced State for Vertical Scaling..... 303

Table 10. Student Demographic Characteristics (in Percentages) for ELA/literacy by Grade for Vertical Scaling..... 304

Table 11. Student Demographic Characteristics (in Percentages) for Mathematics by Grade for Vertical Scaling..... 305

Table 12. Sample Size (Percents) for ELA/literacy and Mathematics by Grade and Smarter Balanced State for the Item Pool Calibration..... 306

Table 13. Student Demographic Characteristics (in Percentages) for ELA/literacy by Grade for the Item Pool Calibration. 307

Table 14. Student Demographic Characteristics (in Percentages) for Mathematics by Grade for the Item Pool Calibration. 308

Field Test Administration and Security.....309

Table 15. Comparison of Features for the Training and Practice Tests..... 310

Table 16. Expected Testing Times for Smarter Balanced Field Tests. 312

Table 17. Distribution of Test Duration in Minutes for the ELA/literacy CAT for the Item Pool Calibration Administration. 314

Table 18. Distribution of Test Duration in Minutes for the ELA/literacy Performance Task for the Item Pool Calibration. 315

Table 19. Distribution of Test Duration in Minutes for the Mathematics CAT for the Item Pool Calibration. 317

Table 20. Distribution of Test Duration in Minutes for the Mathematics Performance Tasks. 318

Universal Tools, Designated Supports, and Accommodations320

Table 21. Definitions for Universal Tools, Designated Supports, and Accommodations..... 320

Test Security321

Table 22. Definitions for Three Levels of Test Security Incidents. 322

References.....323

Chapter 8 Field Test Data Steps and Classical Test Analyses.....324

Introduction.....324

Data Inclusion/Exclusion Rules for Items and Students325

Table 1. Summary of Students Excluded and Resulting Sample Size. 326

Table 2. Summary of ELA Item Exclusions (Item Pool Calibration) by Type. 328

Table 3. Summary of Mathematics Item Exclusions (Item Pool Calibration) by Type..... 328

Item Pool Composition (Vertical Scaling and Item Pool Calibration Steps)	328
Table 4. Summary of ELA Vertical Scaling Items by Purpose and Type.....	329
Table 5. Summary of Vertical Scale Mathematics Items by Purpose and Type.	330
Table 6. Number of On-grade Vertical Scaling Items by Content Area and Characteristics.	331
Table 7. Number of Off-grade Vertical Linking Items by Content Area and Characteristics.....	332
Table 8. Number of On-grade Calibration Items by Content Area and Characteristics.	333
Classical Item and Test Analysis.....	333
Item Difficulty.....	334
Item Discrimination.	335
Distractor Analysis.....	336
Reliability Analyses.....	337
Item Flagging Criteria for Content Data Review.....	337
Table 9. Item Flagging Based on Classical Statistics and Judgmental Review.	338
Table 10. Summary of Vertical Scaling Items with Flags (ELA).....	339
Table 11. Summary of Vertical Scaling Items with Flags (Mathematics).....	341
Table 12. Summary of Item Flags for the Item Pool Calibration (ELA).....	343
Table 13. Summary of Items with Flags for the Item Pool Calibration (Mathematics).....	344
Field Test Classical Results.....	345
Vertical Scaling Results: Classical Item and Test Statistics.	345
Table 14. Number of Items, Average Item Difficulty, and Discrimination for ELA Vertical Scaling Items.	346
Table 15. Number of Items, Average Item Difficulty, and Discrimination for Mathematics Vertical Scaling Items.....	347
Figure 1. P-value Plots for Vertical Linking Items (ELA) (AIS is used here as association between p- values).....	348
Figure 2. P-value Plots for Vertical Linking Items (Mathematics) (AIS is used here as association between p-values).....	349
Table 16. Pearson Correlation between CAT and Performance Tasks for the Vertical Scaling.....	350
Table 17. Number of Items, Average Item Difficulty, and Discrimination for the for Item Pool Calibration.	351
Table 18. Pearson Correlations between CAT and Performance Tasks for the Item Pool Calibration.	352
Table 19. Reliability and SEM of Performance Tasks for the Item Pool Calibration.....	353
Figure 3. Distributions of Total Raw Scores as a Percentage of the Corresponding Maximum Possible Score for the Item Pool Calibration Sample (ELA).....	355

Figure 4. Distributions of Total Raw Scores as a Percentage of the Corresponding Maximum Possible Score for the Item Pool Calibration Sample (Mathematics)..... 356

Subgroup Analysis of Test Difficulty for the Item Pool Calibration.357

 Table 20. Summary of Average Test Difficulty by Subgroup for ELA..... 357

 Table 21. Summary of Average Test Difficulty by Subgroup for Mathematics 361

Differential Item Functioning (DIF) Analyses for the Calibration Item Pool.....364

 Table 22. Definition of Focal and Reference Groups..... 365

 Table 23. DIF Flagging Logic for Selected-Response Items..... 365

 Table 24. DIF Flagging Logic for Constructed-Response Items..... 366

 Table 25. Number of DIF Items Flagged by Category (ELA, Grades 3 to High School). 368

 Table 26. Number of DIF Items Flagged by Category (Mathematics, Grades 3 to High School)..... 369

Prospective Evidence in Support of Rater Agreement370

 Monitoring Scoring Processes.370

 Monitoring Raters and Associated Statistics.....370

 Monitoring Automated Scoring and Associated Statistics.....371

External Assessments: NAEP and PISA.....371

 Table 27. Comparison of Features across the Smarter Balanced, NAEP, and PISA Assessment Programs 372

 PISA Overview.373

 NAEP Overview.....373

 Results374

 Table 28. Number of Items, Average Item Difficulty, and Discrimination for NAEP and PISA Items... 375

 Figure 5. Comparison of NAEP Item Difficulty and Values Obtained from Smarter Balanced Samples 376

References.....377

Chapter 9 Field Test IRT Scaling and Linking Analyses378

 Introduction.....378

 Figure 1. Major Goals and Activities for Field Test Statistical Analysis 379

 Horizontal Scaling: IRT Calibration for a Single Grade379

 Vertical Scaling: Linking Across Multiple Grades.....380

 Figure 2. Summary of Field-Test Vertical Linking Item Configuration..... 383

 Vertical Scale Linking Design.....383

 IRT Preprocessing and Item Inclusion/Exclusion Criteria.....384

Table 1. Number of Items by Type in the Vertical Linking Design..... 385

Table 2. Unique Number of CAT Items and Performance Tasks (PTs) Administered and the Survivorship for Vertical Scaling..... 386

Table 3. Summary of ELA/literacy and Mathematics Items by Purpose and Claim. 387

Table 4. Summary of ELA/literacy by Type and Purpose..... 388

Table 5. Summary of Mathematics by Type and Purpose..... 389

IRT Models and Software389

 Figure 3. Sample ICC Plot for a Dichotomous Item Demonstrating Good Fit..... 391

 Figure 4. Sample ICC Plot for a Dichotomous Item Demonstrating Poor Fit..... 391

 Figure 5. Sample ICC Plot for a Polytomous Item Demonstrating Good Fit..... 392

 Figure 6. Sample ICC Plot for a Polytomous Item Demonstrating Poor Fit 392

Item Fit.392

Vertical Linking Via Stocking-Lord.393

Evaluation of Vertical Anchor Item Stability.....394

 Table 6. Example of STUIRT Linking Methods and Output..... 394

Vertical Scale Evaluation.....394

Horizontal and Vertical Scaling Results395

 Table 7. Summary of Classical Statistics by Purpose for ELA/literacy. 396

 Table 8. Summary of Classical Statistics by Purpose for Mathematics. 397

 Table 9. Summary of Item Parameter Estimates for Horizontal Calibration Step..... 397

 Figure 7. ELA/literacy Item Fit Chi-Square Plots (Vertical Scaling) 398

 Figure 8. Mathematics Item Fit Chi-Square Plots (Vertical Scaling) 399

 Table 10. Summary of Likelihood Ratio χ^2 Test Statistics by Grade and Content Area. 400

 Table 11. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 3 to 4..... 401

 Figure 9. Comparison of ELA/literacy a - and b -parameter estimates for Linking Grade 3 to 4 401

 Table 12. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 4 to 5..... 402

 Figure 10. Comparison of ELA/literacy a - and b -parameter estimates for Linking Grade 4 to 5 402

 Table 13. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 5 to 6..... 403

 Figure 11. Comparison of ELA/literacy a - and b -parameter estimates for Linking Grade 5 to 6 403

Table 14. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 7 to 6..... 404

 Figure 12. Comparison of ELA/literacy a - and b -parameter estimates for linking Grade 7 to 6..... 404

Table 15. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 8 to 7..... 405

 Figure 13. Comparison of ELA/literacy a - and b -parameter estimates for Linking Grade 8 to 7..... 405

Table 16. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: High School to Grade 8..... 406

 Figure 14. Comparison of ELA/literacy a - and b -parameter estimates for Linking High School to Grade 8 406

Table 17. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 3 to 4..... 407

 Figure 15. Comparison of Mathematics a - and b -parameter estimates for Linking Grade 3 to 4 407

Table 18. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 4 to 5..... 408

 Figure 16. Comparison of Mathematics a - and b -parameter estimates for Linking Grade 4 to 5 408

Table 19. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 5 to 6..... 409

 Figure 17. Comparison of Mathematics a - and b -parameter estimates for Linking Grade 5 to 6 409

Table 20. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 7 to 6..... 410

 Figure 18. Comparison of Mathematics a - and b -parameter estimates for Linking Grade 7 to 6 410

Table 21. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 8 to 7..... 411

 Figure 19. Comparison of Mathematics a - and b -parameter estimates for Linking Grade 8 to 7 411

Table 22. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: High School to Grade 8. 412

 Figure 20. Comparison of Mathematics a - and b -parameter estimates for Linking High School to Grade 8 412

Table 23. Vertical Linking Transformation Constants from the Stocking-Lord Procedure..... 413

 Figure 21. Distribution of WRMSD for ELA/literacy (Vertical Linking Items) 414

 Figure 22. Distribution of WRMSD for Mathematics (Vertical Linking Items)..... 415

 Figure 23. ELA/literacy Cumulative Distributions of Student Ability across Grades 416

 Figure 24. Mathematics Cumulative Distributions of Student Ability across Grades 417

Table 24. Summary of Vertically Scaled Student Ability Estimates and Effect Size. 418

Figure 25. ELA/literacy Student Ability Distributions Across Grades 3 to High School 419

Figure 26. Mathematics Student Ability Distributions Across Grades 3 to High School 419

Figure 27. Boxplots of Theta Estimates across Grade Level for ELA/literacy 420

Figure 28. Boxplots of Theta Estimates Across Grade Level for Mathematics..... 421

Figure 29. ELA/literacy Test Information by Score Level and Combined for Grades 3 to 6 422

Figure 30. ELA/literacy Test Information by Score Level and Combined for Grades 7 to High School
..... 423

Figure 31. ELA/literacy Total Test Information for Grades 3 to High School..... 424

Figure 32. Mathematics Test Information and Score Level and Combined for Grades 3 to 6 425

Figure 33. Mathematics Test Information and Score Level and Combined for Grades 7 to High School
..... 426

Figure 34. Mathematics Total Test Information for Grades 3 to High School..... 427

Figure 35. IRT Standard Error Plots for ELA/literacy Grades 3 to High School (HS) 427

Figure 36. IRT Standard Error Plots for Mathematics Grades 3 to High School (HS) 428

Establishing the Minimum and Maximum Scale Score.....429

Table 25. Lowest and Highest Obtainable Theta Values and Resulting Theta Scale Summary..... 429

Cross-validation of Vertical Linking Results.....429

Figure 37. Cross-validation of Vertical Linking Results Comparing cumulative frequency distributions
of theta (EAP) for ELA and mathematics obtained from the CRESST cross-validation 431

Calibration Step for Item Pool.....432

Table 26. Distribution of Student Observations per Item in the Field Test Pool. 432

Table 27. Summary of IRT Item Parameter Estimates for the Field Test Item Pool..... 433

Figure 38. Comparison of Student Proficiency Estimates (theta) for the Vertical Scaling (Achievement
Level Setting Sample) and the Item Pool Calibrations Step for ELA/literacy 434

Figure 39. Comparison of Student Proficiency Estimates (theta) for the Vertical Scaling (Achievement
Level Setting Sample) and the Item Pool Calibrations Step for Mathematics 435

Table 28. Distributions of ELA/literacy Theta Estimates and Conditional Standard Error of
Measurement. 436

Table 29. Distributions of Mathematics Theta Estimates and Conditional Standard Error of
Measurement. 436

References.....437

Chapter 10 Achievement Level Setting.....440

Background.....	440
Achievement Level Setting Process	440
The Bookmark Procedure	441
Creating ordered item booklets.	441
Item mapping.	443
External data.....	443
Typical application of the Bookmark procedure.	444
Software development.	445
Table 1 Software Development Timeline.	446
The home page.....	446
Figure 1. Home Page With One Instruction Bar Expanded.....	447
Figure 2. List of Resources Accessible From Home Page.....	447
Figure 3. Sample Item Map.	448
Figure 4. Sample OIB Page With Selected-Response Item.	449
Figure 5. Item Information Page.....	450
Figure 6. OIB Page For Constructed-Response Item.	451
Figure 7. OIB Page Showing Links to Performance Task, Sample Student Response, and Rubric. ...	452
Figure 8. Comment.	453
Figure 9. Set Bookmark Dropdown Box in the Item Map.....	454
Figure 10. Submitting Bookmarks.	454
Design and Implementation of the Online Panel.....	455
Online panel activities.	455
Conduct and results of the online panel.	455
Table 2. Numbers of Online Panelists, by Role, Grade, and Subject.....	457
Table 3. Impact of Online Panel Bookmark Placements: Percent of Students At or Above Level 3	457
Design and Implementation of the In-Person Workshop.....	458
Table 4. In-Person Workshop Panelists by Subject and Grade.....	459
Table 5. High-Level Agenda for Each In-Person Workshop.	459
Recruitment and selection of panelists.	459
Preparation of materials.....	460
Training of facilitators and table leaders.....	460
Orientation and training.....	461

Round-by-round item review and discussion.....	461
Table 6. Results of Round 1 of Bookmark Placement (Entries are Median Page Numbers).....	462
Table 7. Results of Round 2 of Bookmark Placement (Entries are Median Page Numbers).....	462
Table 8. Results of Round 3 of Bookmark Placement (Entries are Median Page Numbers).....	464
Table 9. Round 3 Cut Score Recommendations: Scale Score Cuts and % At or Above	464
Table 10. Round 3 Questionnaire Results: Confidence in Cut Scores Recommended (Discounting Blanks)	465
Table 11. Summary of Round 3 Evaluation Responses (Discounting Blanks).....	465
Data analysis and reporting.....	466
Design and Implementation of the Cross-Grade Review Committee	466
Figure 11. Cross-Grade Review Graphic.....	468
Table 12. Cross-Grade Review Results	468
Approval by Chiefs	469
Table 13. Final Cut Scores Approved By Chiefs, With Impact Data.....	471
Achievement Level Descriptors	472
Threshold ALDs.....	472
Comparison of final cuts to recommended cuts.....	473
Table 14. Comparison of Final Cuts to Those Recommended by the Cross-Grade Review Committees.	473
ALD review.....	474
Findings and recommendations.....	474
Range ALDs.....	474
Reporting ALDs	474
Long Range Validity Agenda for Performance Level Cut Scores	475
Validation Studies with Internal Variables	476
Validation Studies with External Variables	477
Organization and Implementation of Studies	479
Table 15. Validation Study Implementation Timeline.....	479
References.....	482
Change Log	484

Chapter 1 Introduction

Overview

The Smarter Balanced Assessment Consortium’s (Smarter Balanced) vision for a new generation assessment system—one that includes a set of balanced components that can be adapted to meet students’ needs across participating states. This is rooted in the need for valid, reliable, and fair assessments of the deep disciplinary understanding and higher-order thinking skills that are increasingly demanded by a knowledge-based global economy. This vision also is based on the belief that assessment must support ongoing improvements in instruction and promote meaningful learning experiences for students that lead to outcomes valued by all stakeholders. The overarching goal of Smarter Balanced is to ensure that all students leave high school prepared for postsecondary success in college or a career through a planned sequence of educational experiences and opportunities. To meet this goal, with support from institutions of higher education (IHEs) and workplace representatives, the Consortium built on the strong foundation in each participating state to create a high quality, balanced, multistate assessment system based on the Common Core State Standards (CCSS) in English language arts/literacy (ELA/Literacy) and mathematics. The role of the Consortium in this process was to guide the development and implementation of an assessment system that reshapes educational practice in participating states in strategic ways and leads to improved learning outcomes for students. Smarter Balanced provides options for customizable system components while also ensuring comparability of high-stakes summative test results across states. In addition, the Consortium is committed to creating a policy environment that fosters innovation while supporting the development of accountability systems that incentivize the right behaviors for students, teachers, and administrators and avoid ones that run counter to Smarter Balanced goals. The comprehensive assessment system proposed by the Consortium calls for strategic use of a variety of item types and performance events to measure the full range of the Common Core State Standards and to ensure accurate assessment of all students, including students with disabilities, English language learners, and low- and high-performing students. Smarter Balanced implemented a system that contains the following features

- assessing Common Core State Standards based computer adaptive summative and interim assessments that make use of technology-enhanced item types and human-scored performance events,
- interim/benchmark assessments that provide more flexible and in-depth and/or midcourse information about what students know and can do in relation to the Common Core State Standards,
- research-supported, resource-based, instructionally sensitive tools, processes, and practices developed by state educators that can be used formatively at the classroom level to improve teaching and increase learning,
- focused ongoing support to teachers through professional development opportunities and exemplary resource materials linked to the Common Core State Standards,

- online reporting system that enables educators' secure access to key information about student progress toward college- and career-readiness and about specific strengths and limitations in what students know and are able to do at each grade level, and
- cross-state communications to inform Stakeholders about Smarter Balanced activities and to ensure a common focus on the goal of college- and career-readiness for all students.

A key component of college- and career-readiness is the ability to integrate knowledge and skills across multiple content standards. Smarter Balanced addressed this capacity using new item types and performance tasks. Performance assessment emphasize the application of knowledge and skills and incorporate a range of non-selected-response tasks to evidence about students' abilities to solve substantive, meaningful problems. Performance assessments also give students opportunities to demonstrate their ability to find and organize information to solve problems, undertake research, frame and conduct investigations, analyze and synthesize data, and apply learning to novel situations. Smarter Balanced performance tasks involve interaction of students with stimulus materials and/or engagement in a problem solution, ultimately leading to an exhibition of the students' application of knowledge and skills, often in writing. Stimuli include a variety of information forms (e.g., reading & graphics) as well as an assignment or problem situation. As a result, performance tasks are an integral part of the Smarter Balanced test design. Further, performance tasks allow some types of peer-group work and collaboration with the teacher or other students prior to the actual scored assessment.

The innovative and efficient use of technology serves as a central feature of this balanced assessment system. Some central notions concerning technology use are that

1. the Smarter Balanced system uses computer adaptive testing to increase the precision and efficiency of the tests.,
2. the expanded use of technology enables the Consortium's goals of developing innovative and realistic item types that ensure measurement of student achievement across a wide performance continuum and provides efficiencies and enhancements for teacher and administrator professional development and capacity building at the local level.
3. through the use of an interoperable electronic platform and leveraging of cross-state resources, Smarter Balanced delivers assessments and produces both standardized and customizable reports that are cost effective, timely, and useful for a range of audiences in tracking and analyzing the progress toward college- and career-readiness of individual students, student subgroups, classrooms, schools, districts, and states.

In summary, the Smarter Balanced learning and assessment system is grounded in a sound theory of action. This system promotes research-supported classroom practice and incorporates a balanced set of technology-enabled tools, innovative assessments, and state-of-the-art classroom support mechanisms that are intended to work coherently to facilitate teaching and learning. Over time, with a purposeful governing structure and Institutes of Higher Education in participating states, this assessment system will affect the improve teaching and learning consistent with Smarter Balanced's theory of action described below.

Technical Report Approach

The intent of this report is to provide comprehensive and detailed evidence in support of the validity of Smarter Balanced assessment program. Integral to this description is a discussion of validity and the Smarter Balanced test validation process. At the outset, it should be recognized that this process of demonstrating evidence in support of validity is an ongoing process. Validity information is provided here is primarily from the initial Pilot Test and the later Field Test phases. The Field Test reflects the final item statistics of record, reporting scale, and achievement levels for ELA/Literacy and mathematics.

To inform the Consortium, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999; 2014) hereafter referred to as the *Standards*) was used as the foundation for developing the necessary validity evidence. The 2014 version of the *Standards* was published just as this report was in the process of being finalized. Occasionally, this report referenced the 2014 *Standards* as well. Also referenced is the U.S. Department of Education (DOE)'s *Standards and Assessments Peer Review Guidance* (2009b), which stipulated the requirements for assessment programs to receive federal approval under the No Child Left Behind (NCLB) legislation. This information is consistent with the Joint Committee on Standards for Educational Evaluation's (JCSEE) *Program Evaluation Standards* (Yarbrough, Shulha, Hopson, & Caruthers, 2011) and the *Guiding Principles for Evaluators* (American Evaluation Association, 2004) which state "evaluators aspire to construct and provide the best possible information that might bear on the value of whatever is being evaluated" (p. 1). With respect to Smarter Balanced, this information is necessary for understanding the degree to which the Consortium is meeting its goals, and in some cases, what further tasks remain to improve the system as it evolves operationally.

NCLB Peer Review Guidelines and Established Standards

Among the principles underlying the Smarter Balanced theory of action is the adherence "to established professional standards" (Smarter Balanced, 2010, p. 33). In addition to adhering to the AERA et al. (1999; 2014) *Standards*, the Consortium will also meet the requirements of the U.S. Department of Education peer review process for NCLB assessments. Although these requirements were suspended as they undergo revision (Delisle, 2012), they remain important because they reflected the Department's most recent standards for ensuring quality and equity in statewide assessment programs. Validity evidence and the ongoing research agenda should incorporate the guidance provided in the *Standards and Assessments Peer Review Guidance* (U.S. Department of Education, 2009b). There is a great deal of overlap between the AERA et al. (1999; 2014) *Standards* and the U.S. Department of Education's *Peer Review Guidance*. However, the *Guidance* stipulates many important requirements. In particular, to meet these requirements the validity evidence and the ongoing research agenda should include the following

- evidence concerning the purpose of an assessment system and studies that support the validity of using results from the assessment system based on their stated purpose and use,
- strong correlations of test and item scores, with relevant measures of academic achievement and weak correlations with irrelevant characteristics, such as demographics (i.e., convergent and discriminant validity),
- investigations regarding whether the assessments produce the intended consequences;

- documentation of the definitions for cut scores and the rationale and procedures for establishing them,
- evidence concerning the precision of the cut scores and consistency of student classification,
- evidence of sufficient levels of reliability for the overall population and for each targeted subpopulation,
- evidence of content alignment over time through quality control reviews,
- evidence of comprehensive alignment and measurement of the full range of content standards, Depth of Knowledge, and cognitive complexity,
- evidence that the assessment plan and test specifications describe how all content standards are assessed and how the domain is sampled that lead to valid inferences about student performance on the standards, both individually and aggregated,
- scores that reflect the full range of achievement standards,
- documentation that describes how the assessments consist of a *coherent* system across grades and subjects including studies establishing vertical scales, and
- identification of how assessments provide information on the progress of students.

The overlap of this evidence with the AERA et al. (1999; 2014) *Standards* is large, and the anticipated revisions to this guidance will likely retain many of these key features. For example, in the temporary suspension of peer review, the U.S. Department of Education reiterated the following desired characteristics for a high-quality assessment system as

- valid, reliable, and fair for its intended purposes and measures student knowledge and skills against college- and career-ready standards,
- covering the full range of those standards, including standards against which student achievement has traditionally been difficult to measure,
- appropriate, eliciting complex student demonstrations or applications of knowledge and skills,
- providing an accurate measure of student achievement across the full performance continuum, including for high- and low-achieving students,
- assessing all students, including English Language Learners and students with disabilities,
- making provisions for alternate assessments based on grade-level academic achievement standards or alternate assessments based on alternate academic achievement standards for students with the most significant cognitive disabilities, consistent with 34 C.F.R. § 200.6(a)(2),
- providing an accurate measure of student growth over a full academic year or course,
- providing student achievement data and student growth data that can be used to determine whether individual students are college- and career-ready or on track to being college- and career-ready,
- producing data, including student achievement data and student growth data, that can be used to inform determinations of school effectiveness for purposes of accountability under Title I, individual principal and teacher effectiveness for purposes of evaluation; principal and teacher professional development and support needs; and program improvement.

These characteristics of high-quality assessment systems were given consideration in the development of the Smarter Balanced Assessment System to ensure that evidence was provided meets these high

standards. The Theory of Action and primary purposes and goals of Smarter Balanced are briefly described below.

Overview and Background of the Smarter Balanced Theory of Action

The Smarter Balanced Assessment Consortium supports the development and implementation of learning and assessment systems to reshape education in participating states in order to improve student outcomes. Through expanded use of technology and targeted professional development, the Consortium's Theory of Action calls for integration of learning and assessment systems, leading to more informed decision-making and higher-quality instruction and ultimately increasing the number of students who are well prepared for college and careers.

The ultimate goal of Smarter Balanced is to ensure that all students leave high school prepared for postsecondary success in college or a career through increased student learning and improved teaching. This approach suggests that enhanced learning will result from high-quality assessments that support ongoing improvements in instruction and learning and which are educative for students, parents, teachers, school administrators, members of the larger public, and policymakers. Meeting this goal will require reform and coordination of many elements across the education system. This goal includes but is not limited to a quality assessment system that strategically “balances” summative, interim, and formative components (Darling-Hammond & Pecheone, 2010). An assessment system is required that provides valid measurement across the full range of common rigorous academic standards, including assessment of deep disciplinary understanding and higher-order thinking skills that are increasingly demanded by a knowledge-based economy, and by the establishment of clear, internationally benchmarked performance expectations. Other elements that are outside the intended Smarter Balanced scope-of-work, but not outside its influence, are comprehensive professional development and valid accountability measures.

Seven Principles of Smarter Balanced Underlying the Theory of Action

The Smarter Balanced assessment is guided by a set of seven principles shared by systems in high-achieving nations and a number of high-achieving states in the U.S.

1. Assessments are grounded in a thoughtful, standards-based curriculum and managed as part of an integrated system of standards, curriculum, assessment, instruction, and teacher development. Curriculum and assessments are organized around a well-defined set of learning progressions along multiple dimensions within subject areas. Formative and interim/benchmark assessments and associated support tools are conceptualized in tandem with summative assessments; all of them are linked to the Common Core State Standards and supported by a unified technology platform.
2. Assessments produce evidence of student performance on challenging tasks that evaluate the Common Core State Standards. Instruction and assessments seek to teach and evaluate knowledge and skills that generalize and can transfer to higher education and multiple work domains. These assessments emphasize deep knowledge of core concepts and ideas within and across the disciplines—along with analysis, synthesis, problem solving, communication, and critical thinking—thereby requiring a focus on complex performances as well as on specific concepts, facts, and skills.
3. Teachers are integrally involved in the development and scoring of assessments. While many assessment components are efficiently scored with computer assistance, teachers must also be

involved in the formative and summative assessment systems so that they understand and can teach in a manner that is consistent with the full intent of the standards while becoming more skilled in their own classroom assessment practices.

4. The development and implementation of the assessment system is a state-led effort with a transparent and inclusive governance structure. Assessments are structured to improve teaching and learning. Assessments as, of, and for learning are designed to develop understanding of learning standards, what constitutes high-quality work, to what degree is growth occurring, and what is needed for further student learning.
5. Assessment, reporting, and accountability systems provide useful information on multiple measures that is educative for all Stakeholders. Reporting of assessment results is timely and meaningful—offering specific information about areas of performance so that teachers can follow up with targeted instruction, students can better target their own efforts, and administrators and policymakers can fully understand what students know and can do—in order to guide curriculum and professional development decisions.
6. Design and implementation strategies adhere to established professional standards. The development of an integrated, balanced assessment system is an enormous undertaking, requiring commitment to established quality standards in order for the system to be credible, fair, and technically sound. Smarter Balanced continues to be committed to developing an assessment system that meets critical elements required by US DOE Peer Review, relying heavily on the *Standards* as its core resource for quality design. Other key sources of professional standards that guide Smarter Balanced include a reasoning-from-evidence approach (National Research Council, 2001; Mislevy, Almond, & Lukas, 2004), *An Introduction to the Operational Best Practices for Statewide Large-Scale Assessment Programs* (Association of Test Publishers, Council of Chief State School Officers, 2010), and the American National Standards Institute (ANSI) endorsed *Student Evaluation Standards*, *Program Evaluation Standards*, and *Personnel Evaluation Standards* (Joint Committee on Standards for Educational Evaluation, 1994, 2002, 2008, respectively).

Purposes for the Smarter Balanced Assessment System

The Smarter Balanced purpose statements refer to three categories: (a) summative assessments, (b) interim assessments, and (c) formative assessment resources.

The purposes of the Smarter Balanced summative assessments are to provide valid, reliable, and fair information concerning

- students' ELA/literacy and mathematics achievement with respect to the Common Core State Standards as measured by the ELA/literacy and mathematics summative assessments in grades 3 to 8 and high school,
- whether grade 11 students have sufficient academic proficiency in ELA/literacy and mathematics to be ready to take credit-bearing, transferable college courses after completing their high school coursework,
- measurement of students' status prior to grade 11 to determine whether they have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on track for achieving college and/or career readiness,

- measurement of students' annual growth toward college- and career-readiness in grade 11 ELA/literacy and mathematics,
- how instruction can be improved at the classroom, school, district, and state levels,
- students' ELA/literacy and mathematics proficiency for federal accountability purposes and potentially for state and local accountability systems, and
- equitable achievement for all students and subgroups of students in ELA/literacy and mathematics.

The purposes of the Smarter Balanced interim assessments are to provide valid, reliable, and fair information about

- student progress toward mastery of the skills in ELA/literacy and mathematics measured by the summative assessment,
- student performance at the Claim or cluster of Assessment Targets so teachers and administrators can track student progress throughout the year and adjust instruction accordingly,
- individual and group (e.g., school, district) performance at the Claim level in ELA/literacy and mathematics to determine whether teaching and learning are on target,
- teacher-moderated scoring of performance events as a professional development vehicle to enhance teacher capacity to evaluate student work aligned to the standards, and
- student progress toward the mastery of skills measured in ELA/literacy and mathematics across all students and subgroups.

The purposes of the Smarter Balanced formative assessment resources are to provide measurement tools and resources to

- improve teaching and learning,
- provide resources to teachers to help them monitor their students' progress throughout the school year,
- illustrate how teachers and other educators can use assessment data to engage students in monitoring their own learning, and
- help teachers and other educators align instruction, curricula, and assessments,
- assist teachers and other educators in using the summative and interim assessments to improve instruction at the individual and classroom levels,
- offer professional development and resources for how to use assessment information to improve teacher decision-making in the classroom.

The primary rationale of the Smarter Balanced assessments is that these aspects can interact to improve the intended student outcomes (i.e., college- and career-readiness). While there are many ways in which the Smarter Balanced assessment system can be deployed, one possible connection among these assessment components is presented in Figure 1.

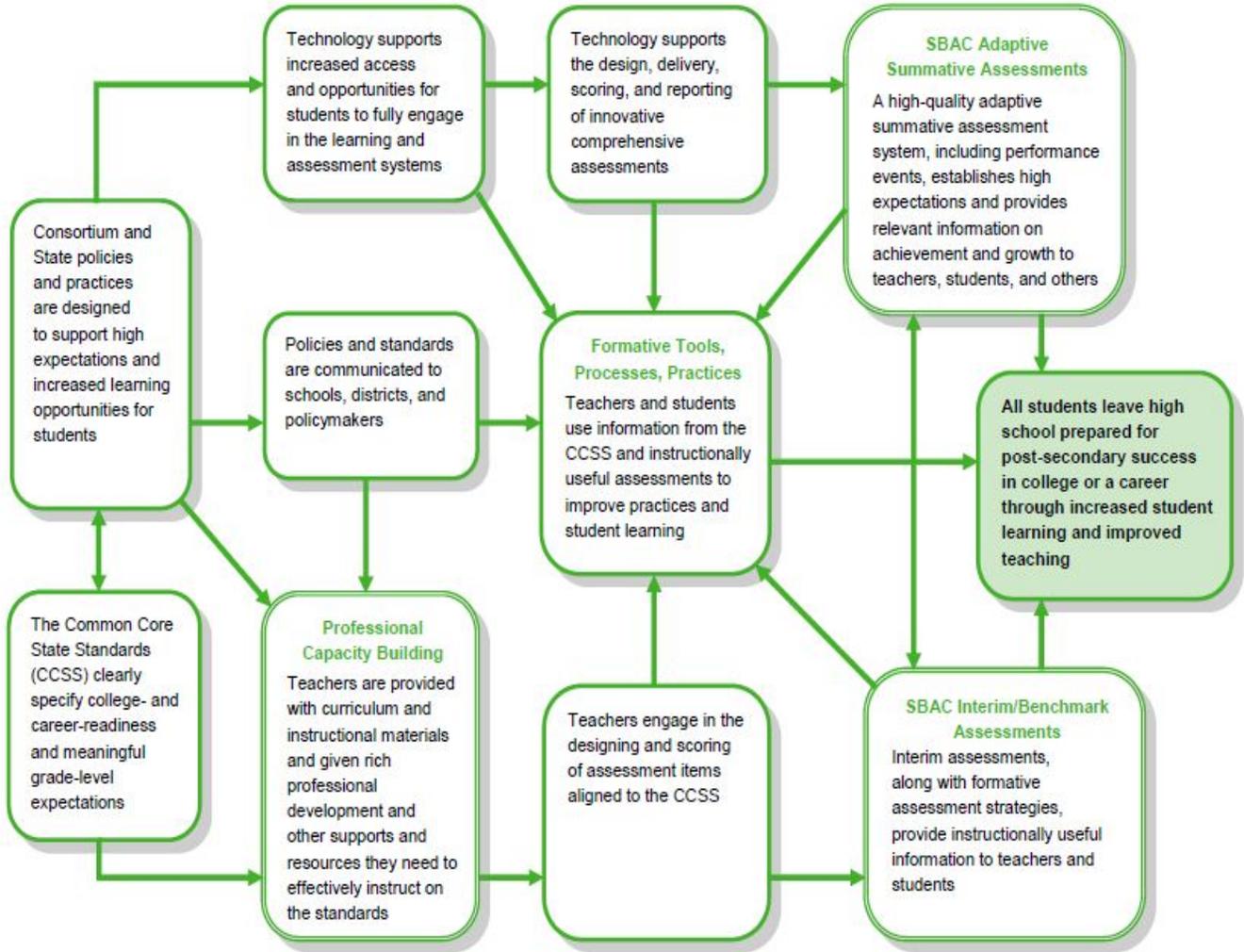


Figure 1. Overview of Smarter Balanced Theory of Action

The Smarter Balanced timeline that references critical aspects of development of the system, such as achievement level descriptors, item development and acceptance, Pilot and Field Test windows, and standard setting events, is given below.

Timeline

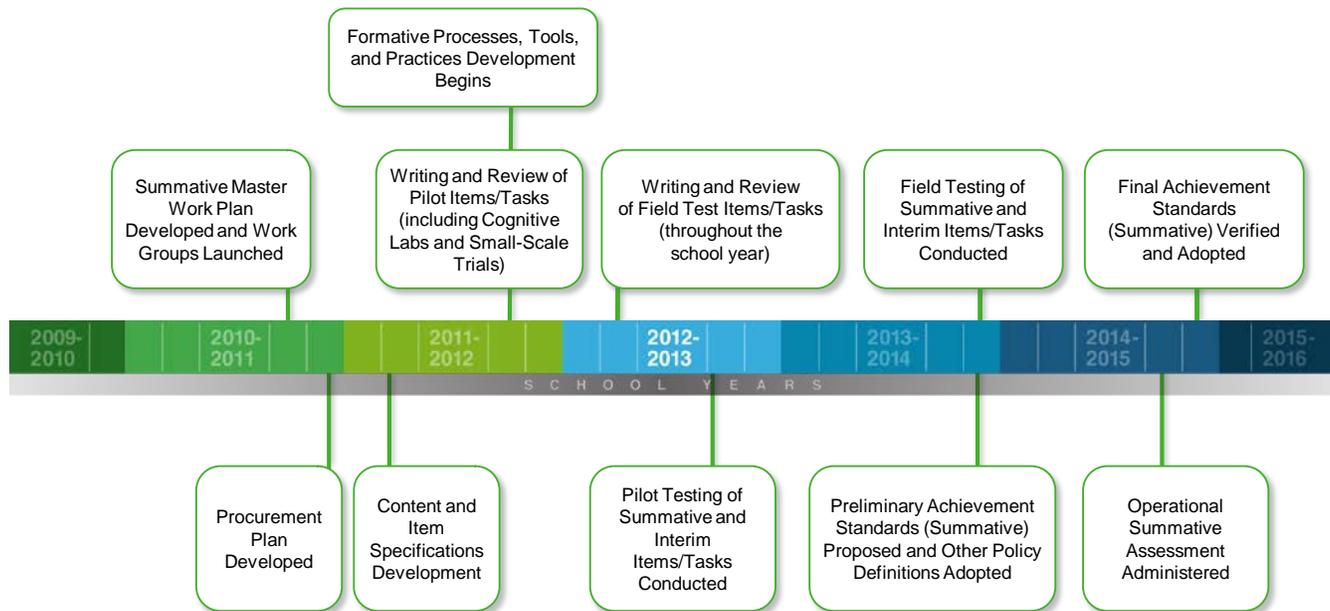


Figure 2. Smarter Balanced Development Timeline

Chapter Overview

Brief synopses of the other chapters contained in the Smarter Balanced Technical Report are given below in order to direct further review.

Chapter 2 Validity Various types of evidence are offered in support of validity in an on-going process of test validation. These sources are validity evidence from the *Standards* are based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. As appropriate, some future sources of research and associated evidence are suggested. Since the evidence provided is consistent with a Field Test, essential elements necessary for validity are presented.

Chapter 3 Content Development The process of item and task development is described that presents evidence related to content related validity. A major component for item development was the content alignment studies exploring innovative item and task types writing items intended to measure the Common Core of State Standards. The rationale for and the development process for new types of machine scored and listening items, and classroom-based performance task are described.

Chapter 4 Test Design This chapter provides information pertaining to the content validity of the Smarter Balanced assessment system. Test design is primarily deals with test philosophy and purposes, the intended student population and other technical characteristics. There are two interrelated parts to the test design presented here. More traditionally, the first section pertains to the test blueprints and the entailing specifications such as the use of evidence centered design in Smarter Balanced test development. Secondly, the rationale for performance and CAT assessment components are presented. Summative and interim blueprints are presented here. Central organizing principles for the design and use of interim assessments are also given. Since CAT is an important part of the test design, recommended approaches for evaluating this test delivery approach and the accompanying item pool are presented.

Chapter 5 Test Fairness Test fairness concerns whether score interpretations are valid for all relevant subgroups that minimizes construct irrelevant variance. The evidence for test fairness can consist of logical (e.g., bias review of items) or can be statistical in nature (e.g., differential item functioning, DIF). This chapter primarily presents Smarter Balanced Conceptual Framework for Usability, Accessibility, and Accommodations as well as DIF analysis.

Chapter 6 Pilot Test The Pilot Test design is described, sampling, and the subsequent data analysis. Two outcomes of the Pilot were a dimensionality study for each content area and a study on Item Response Theory (IRT) model choice. The outcomes from these two studies were recommendations for the use of unidimensional models and the two-parameter and generalized partial credit models for IRT scaling.

Chapter 7 Field Test Design Sampling and Administration The Field Test design is described that described the Linear-on-the-Fly test delivery to students that consisted of CAT items and the administration of performance tasks. Two major scaling phases are described that consisted of the IRT vertical scaling and the later horizontal calibration and linking of the much larger item pool. The sampling design is described and the associated decisions. Basic elements of the test administration and security are described.

Chapter 8 Data Step and Classical Test Analysis After the Field Test administration was completed, a data step was performed prior to classical analysis. The data step included decisions concerning the item and student inclusion/exclusion logic and the outcomes from the classical item and test analysis such as item p-values (difficulty).

Chapter 9 Field Test Scaling and Linking Analysis Construction of the vertical scale using the IRT models (i.e., 2PL/GPCM) is presented. Extensive description of the vertical scaling and supporting evidence is given. The outcomes from the second calibration of the item pool are given in terms of its IRT characteristics.

Chapter 10 Achievement Level Setting The chapter describes the procedures used for the Smarter Balanced achievement level setting in the fall of 2014 and the outcomes in terms of the cut scores established.

Acknowledgments

Smarter Balanced Work Groups

The multifaceted nature of this project required that these contracts be organized into larger cross-disciplinary Smarter Balanced work groups that consisted of:

- Accessibility and Accommodations Work Group (A&A)
- Test Administration/Student Access Work Group (TASA)
- Item Development/Performance Tasks Work Group (ID/PT)
- Validation and Psychometrics/Test Design Work Group (V&P/TD)
- Formative Assessment Practices and Professional Learning/Transition to Common Core Work Group (FAPPL/TCC)
- Technology Approach/Reporting Work Group (Tech/RPT)

Outside Groups and Organizations that Collaborated with the Smarter Balanced Assessment System

Below is a partial list of individuals and groups that contributed time and expertise to the consortium.

2014 Technical Advisory Committee.

- | | |
|--------------------------------|--|
| • Jamal Abedi, Ph.D. | <i>UC Davis/CRESST</i> |
| • Randy Bennett, Ph.D. | <i>ETS</i> |
| • Derek C. Briggs, Ph.D. | <i>University of Colorado</i> |
| • Gregory J. Cizek, Ph.D. | <i>University of North Carolina</i> |
| • David T. Conley, Ph.D. | <i>University of Oregon</i> |
| • Linda Darling-Hammond, Ph.D. | <i>Stanford University</i> |
| • Brian Gong, Ph.D. | <i>The Center for Assessment</i> |
| • Edward Haertel, Ph.D. | <i>Stanford University</i> |
| • Joan Herman, Ph.D. | <i>UCLA/CRESST</i> |
| • G. Gage Kingsbury, Ph.D. | <i>Psychometric Consultant</i> |
| • James W. Pellegrino, Ph.D. | <i>University of Illinois, Chicago</i> |
| • W. James Popham, Ph.D. | <i>UCLA, Emeritus</i> |
| • Joseph Ryan, Ph.D. | <i>Arizona State University</i> |
| • Martha Thurlow, Ph.D. | <i>University of Minnesota/NCEO</i> |

Contributors to the Accessibility Accommodations Framework.

In February 2012, the Smarter Balanced Assessment Consortium Accessibility and Accommodations Work Group began work on developing the Accessibility and Accommodations Framework. The primary goal of this effort was to develop uniform accessibility and accommodation policies and guidelines that will be adopted and used by all Smarter Balanced states. Recognizing the diversity in policies and practices that currently exist across states, the legal issues that must be addressed by the policies, the mixed research findings regarding many accommodation practices, and the differences in opinion regarding accommodation policies, the work group undertook an iterative process designed to gather input from a large and diverse audience. This effort began by contracting with Measured Progress and its partners, who included:

- Members of the Measured Progress Innovation Lab who conducted work in accessibility in digital environments, developed the Accessible Test Design model, and were leaders in developing the Accessible Portable Item Protocol (APIP) Standard,
- Experts at Educational Testing Service who have conducted a variety of studies on test accommodations and accessibility for students with disabilities and for students who are English language learners, and who have developed industry-recognized guidelines for accessibility in the context of assessment,
- Experts at the George Washington University Center for Equity and Excellence in Education, who are nationally recognized experts in accessible assessment for students who are English language learners and who have worked with several states to develop policies on test accommodations for students who are English language learners, and
- Experts affiliated with the National Center on Educational Outcomes who have conducted extensive reviews of state-test accommodation policies, worked with the Assessing Special Education Students (ASES) work group of the Council of Chief State School Officers (CCSSO) to develop test accommodation policies, and closely monitored research on test accommodations.

In addition to these partners, an expert panel was formed composed of the following members:

- Jamal Abedi -assessment of English language learners,
- Martha Thurlow -assessment of students with disabilities,
- Sheryl Lazarus -test accommodations for students with disabilities,
- Stephanie Cawthon -accommodations for students who communicate in American Sign Language
- Richard Jackson -accommodations for students with visual impairments,
- Rebecca Kopriva -assessment of students who are English language learners, and
- Stephen Sireci -validity of test accommodations.

Other Acknowledgments.

This technical report leveraged the *Smarter Balanced Comprehensive Research Agenda* by Stephen G. Sireci (2012) as the primary validity framework and sources of evidence. Input was provided on critical aspects of the program and this report by the Smarter Balanced Technical Advisory Committee.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.
- Darling-Hammond, L., & Pecheone, R. (2010). *Developing an Internationally Comparable Balanced Assessment System that Supports High-Quality Learning*. Retrieved from <http://www.k12center.org/publications.html>.
- Delisle, D.S. (2012, December). Letter to Chief State School Officers. Washington, DC: U.S. Department of Education.
- American Evaluation Association (2004). *Guiding Principles for Evaluators*. Washington, DC: Author.
- Mislevy, R.J., Almond, R.G., & Lukas, J. (2004). *A Brief Introduction to Evidence-Centered Design*. (CSE Technical Report 632). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST).
- National Research Council. (2001). *Educating Children With Autism*. Committee on Educational Interventions for Children with Autism. Division of Behavioral and Social Sciences and Autism. Washington, DC: National Academy Press.
- No Child Left Behind Act of 2001*, 20 U.S.C. 70 6301
- Smarter Balanced Assessment Consortium (2010, June 23). Race to the Top Assessment Program Application for New grants: Comprehensive Assessment Systems, CFDA Number: 84.395B. OMB Control Number 1810-0699.
- U.S. Department of Education (2009a, November). *Race to the Top Program Executive Summary*. Washington, DC: Author.
- U.S. Department of Education (2009b, January). Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001. [Revised December 21, 2007 to include Modified academic achievement standards. Revised with technical edits, January 12, 2009] Washington, DC: Author.
- U.S. Department of Education (2010, September). U.S. secretary of education Duncan announces winners of competition to improve student assessments. Downloaded October 3, 2012 from <http://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse>.
- Joint Committee on Standards for Educational Evaluation, American Psychological Association Student Evaluation Standards, Program Evaluation Standards, and Personnel Evaluation Standards (2008).

Sireci, S. G. (2012). Smarter Balanced Assessment Consortium: Comprehensive Research Agenda. Report Prepared for the Smarter Balanced Assessment Consortium.

Standards and Assessments Peer Review Guidance (U.S. Department of Education, 2009b).

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., and Caruthers, F. A. (2011). *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users* (3rd ed.). Thousand Oaks, CA: Sage. 34 C.F.R. § 200.6(a)(2)

An Introduction to the Operational Best Practices for Statewide Large-Scale Assessment Programs (2013). ATP& CCSSO.

Chapter 2: Validity

Introduction

Validity refers to the degree to which each interpretation or use of a test score is supported by the accumulated evidence (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999; 2014; ETS, 2002). It constitutes the central notion underlying the development, administration, and scoring of a test and the uses and interpretations of test scores. Validation is the process of accumulating evidence to support each proposed score interpretation or use. This validation process does not rely on a single study or gathering one type of evidence. Rather, validation involves multiple investigations and different kinds of supporting evidence (AERA, APA, & NCME, 1999; 2014; Cronbach, 1971; ETS, 2002; Kane, 2006). It begins with test design and is implicit throughout the entire assessment process, which includes item development and field-testing, analyses of items, test scaling, and linking, scoring, and reporting. This chapter provides an evaluative framework for the validation of the Smarter Balanced Assessment System. It points the reader to supporting evidence in other parts of this technical report and other sources that seek to demonstrate that the Smarter Balanced Assessment System adheres to guidelines for fair and high quality assessment. Since many aspects of the program were still under development at the time of this report, additional research that further supports the Smarter Balanced goals is mentioned as appropriate throughout this chapter.

This chapter is organized primarily around the principles prescribed by AERA, APA, and NCME's *Standards for Educational and Psychological Testing* (1999; 2014) and the Smarter *Balanced Assessment Consortium: Comprehensive Research Agenda* (Sireci, 2012), both of which serve the primary sources for this chapter. The *Standards* are considered to be "the most authoritative statement of professional consensus regarding the development and evaluation of educational and psychological tests" (Linn, 2006, p. 27) currently available. As this report and the associated validation was nearing completion, the 2014 *Standards* were published. The basic notions of validity described in the 1999 *Standards* are consistent with those entailed in the 2014 *Standards*. The 2014 *Standards* differ from earlier ones in the emphasis given to the increased prominence of technology in testing, such as computer adaptive testing (CAT) and automated scoring. CAT methodology and automated scoring approaches are both important components of the Smarter Balanced assessments. The use of the *Standards* in this chapter refers to the 2014 version unless the 1999 edition is specifically referenced.

The validity evidence presented in this technical report was collected in the context of two phases that consisted of a pilot test and field test prior to any operational administration. As a result, many critical elements of the program were being developed simultaneously within a short time span late in 2014. The validity evidence is intended to provide the best possible information for both understanding the degree to which the Smarter Balanced Consortium is meeting its goals consistent with completion of the Field Test phase, as well as the steps needed to be undertaken to improve the system as it evolves operationally.

Two types of overlapping validity frameworks are presented. The first validity framework corresponds to the essential validity elements (AERA et al. 1999, p.17). This essential validity information is more consistent with the types of evidence typically reported for many large-scale educational assessment programs. These essential validity elements present a more traditional synopsis of validity evidence, which form the basis for the evidence demonstrated for the Smarter Balanced Field Test to date and the initial operational administrations. The second more comprehensive validity framework cross-references Smarter Balanced test purposes against the *Standards'* five primary sources of validity evidence. These five sources of validity evidence consist of (1) test content, (2) response processes,

(3) internal structure, (4) relations to other variables, and (5) consequences of testing. Evidence in support of the five sources of validity will need to be addressed more fully in the course of ongoing Smarter Balanced research. The essential validity elements form a subset and sample the five sources of validity evidence. The essential validity framework is presented first followed by the five primary sources of validity.

Essential Validity Elements for Summative and Interim Assessments

The *Standards* describe the process of validation that consists of developing a sufficiently convincing argument, based on empirical evidence, that the interpretations and actions based on test scores are sound. Kane (1992, 2006) characterized this process as a validity argument, which is consistent with the validation process described by the 1999 *Standards*.

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. . . . Ultimately, the validity of an intended interpretation . . . relies on all the available evidence relevant to the technical quality of a testing system (AERA et al., 1999, p. 17).

The 1999 *Standards* describe these essential validity elements as “evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees.” Some modifications were made to the original 1999 specification of these essential elements. Careful attention to fairness for all examinees was changed to attention to fairness, equitable participation, and access. Validating on-track/readiness and test security were also added as essential elements. Although the 1999 *Standards* mention “reliability,” the more general term of “precision” is used instead to underscore the need to conceptualize measurement error with other frameworks such as item response theory and generalizability theories. Table 1 presents a brief description of this essential validity evidence. Many of these essential validity elements fall under the validity evidence based on test content (e.g., careful test construction) and internal structure (adequate score reliability, scaling, equating). The types of evidence listed in Table 1 will reemerge when considering the five specific validity sources, which represent the full validity framework. This overlap underscores the fundamental nature of these elements for supporting the use of Smarter Balanced assessments for their intended purposes. Table 1 is followed by a brief description of the potential types of evidence associated with each essential element.

Table 1. Synopsis of Essential Validity Evidence Derived from *Standards* (AERA et al., 1999, p. 17).

Essential Element	Type of Associated Validation Evidence
Careful Test Construction	Examination of test development steps, including construct definition (test specifications and blueprints), item writing, content review, item analysis, alignment studies, and other content validity studies; review of technical documentation such as IRT scaling.
Adequate Measurement Precision (Reliability)	Analysis of test information, conditional standard errors of measurement, generalizability studies, decision accuracy and consistency, and reliability estimates.
Appropriate Test Administration	Review of test administration procedures, including protocols for test irregularities; use of and appropriate assignment of test accommodations.
Appropriate Scoring	Review of scoring procedures (hand-scored, automated), rater agreement analyses, machine/human comparisons (if relevant), generalizability studies, and fairness for subgroups. <i>Test Scoring Specifications</i> (AIR, 2014b).
Accurate Scaling and Equating	Documentation of test design, IRT model choice, scaling and equating procedures, IRT residuals, validating vertical scaling assumptions, third-party verification of horizontal and vertical equating.
Appropriate Standard Setting	Comprehensive standard-setting documentation provided, in Chapter 10, including procedural, internal, and external validity evidence for all achievement-level standards.
Attention to Fairness, Equitable Participation and Access	Review of accommodation policies, implementation of accommodations, sensitivity review, DIF analyses, differential predictive validity analyses, qualitative and statistical analyses of accommodated tests; analysis of participation rates, translations, and other policies.
Validating “On-track/Readiness”	Examining relationships with external variables as well as evidence from internal structure.
Adequate Test Security	Analysis of data integrity policies, test security procedures, monitoring of test administrations, analysis of suspected cheating behavior, item exposure, review of anomalous CAT results.

Careful Test Construction. Validity evidence of careful test construction can derive from a comprehensive inspection of the test development process that reviews all test development

activities. The audit encompasses descriptions of testing purposes, operational definitions of the constructs measured, item development procedures, content reviews, alignment studies, sensitivity and bias reviews, pilot testing, item and DIF analyses, item calibration, item selection, scoring rubrics for constructed-response items, and assembly of tests and clarity of test instructions. For adaptive assessments, the adequacy of the item selection algorithm, particularly in delivering tests that conform to the blueprint, should also be reviewed.

The degree to which the test specifications for the assessment sufficiently reflect the Common Core State Standards and the degree to which the relative weights of the cells in the test specifications reflect the corresponding emphases in the Common Core State Standards should be evaluated (Mislevy & Riconscente, 2006). This entails the use of traditional content validity studies (e.g., Crocker, Miller, & Franks, 1989) and alignment studies (Bhola, Impara, & Buckendahl, 2003; Martone & Sireci, 2009; Porter & Smithson, 2002; Rothman, 2003; Webb, 2007). To evaluate the appropriateness of the test specifications, the process by which the specifications were developed need to be reviewed to ensure that all member states had input and that there was consensus regarding the degree to which the test specifications targeted for the assessment represented the Common Core State Standards. To evaluate the degree to which the summative assessments adequately represent the test specifications requires recruitment and training of qualified and independent subject matter experts in ELA/literacy and mathematics to review the Common Core State Standards in conjunction with the test specifications and Smarter Balanced test items. At least two hypothesized aspects of the assessments can be validated using the content experts. First, the items can be evaluated to ensure that they were appropriately assessing the Common Core of State Standards as intended. Second, the items are measuring the breadth of higher- and lower-order cognitive skills (i.e., Depth of Knowledge) that they are intended to measure. Popham (1992) suggested a criterion of 7 out of 10 subject matter experts (SMEs) rating an item congruent with its standard to confirm that the item fits the standard. Several statistics have been proposed for evaluating item-standard congruence such as Hambleton's (1980) item-objective congruence index and Aiken's (1980) content validity index. In addition, Penfield and Miller (2004) established confidence intervals for subject matter experts mean ratings for content congruence.

Adequate Measurement Precision (Reliability). The notion of measurement precision extends the notion of reliability beyond a descriptive statistic for a test. It refers to the amount of expected variation in a test score or a classification based on a test score. Examples of this type of information include estimates of score reliability, standard errors of measurement, item and test information functions, conditional standard error functions, and estimates of decision accuracy and consistency. Estimates of score reliability typically include internal consistency estimates based on a single test administration (coefficient alpha, stratified alpha, IRT marginal reliability). Generalizability studies that focus on specific facets of measurement are important for identifying the sources of measurement error.

For expected operational adaptive test delivery with multiple content and psychometric constraints, simulations play an important role in evaluating the operational adaptive algorithm and delivery system and the evaluation of measurement error. Test information functions, recovery of simulated examinee ability, and analysis of bias and error are all highly interrelated and can be addressed collectively. All test scores include an error component, the size of which generally varies across test takers. Differences in precision across score ranges are ignored by overall measures of precision that, like test reliability, are aggregated across score levels. However, IRT provides a related pair of test precision measures that are specific to, or conditional on, score level. Both the test information function and the inversely related conditional standard error measure test-precision level across the score scale. (The conditional standard error function is the inverse of the square root of the test information function.) In a simulation environment, the score bias function measures the extent to which score estimates converge to their true values. The smaller the bias and error, the better the

test administration and scoring procedures recover simulated examinee ability. Even if the goal is to measure each student to some fixed criteria for test information/conditional standard error, test precision can vary not just across proficiency levels but also across test takers at the same level of proficiency. Certain students are more easily assessed compared with other ones. Students who respond predictably (as the underlying item-response model expects them to) will be more easily measured by an adaptive test than those who respond in unanticipated ways. Predictable students are well targeted early in a test and are typically presented a series of highly discriminating items. Those students that respond in more unexpected ways will be more difficult to target and less likely to receive informative tests. Much of this inconsistency is unavoidable. However, test administration procedures may differ in the extent to which each test taker is measured on the targeted precision. It should be noted that exceeding the precision target is almost as undesirable as falling short. Measuring some test takers more precisely than necessary wastes resources (in the form of item exposures) that could be used more productively with other test takers.

Appropriate Test Administration. Evidence in this category involves review of test administration manuals and other aspects of the test administration processes. This review can include a review of the materials and processes associated with both standard and accommodated test administrations. Observations of test administrations and a review of proctor and test irregularity reports can be inspected. The policies and procedures for granting and providing accommodations to students with disabilities and English language learners can be reviewed, and case studies of accommodated test administrations should be selected and reviewed to evaluate the degree to which the policies and procedures were followed.

Evidence in this category should also confirm that the routing of test content to students during the linear-on-the-fly and adaptive administration is performing according to expectations and that all computerized scoring programs are accurate. Monitoring of item exposure rates is important as well. The *Standards* (2014 p. 43) point out that one way to evaluate administrations using computer adaptive techniques is to use simulations with known parameters to estimate reliability/precision.

Appropriate Scoring. Validity evidence to confirm that the scoring of Smarter Balanced assessments is appropriate should include a review of scoring documentation. The *Standards* (p. 92) state that such documentation should be presented in sufficient detail and clarity to maximize the accuracy of scoring and the processes for selecting, training, and qualifying scorers. The scoring processes should also include monitoring of all aspects of rater agreement. If any assessments are scored locally, the degree to which the scorers are trained, and the accuracy of their scores, should also be documented. Generalizability studies that quantify various sources of measurement error also provide important evidence, such as characterizing the student by task interactions on performance tasks. For automated scoring, descriptions of the methods used for scoring should be described as well as the development methods that were utilized, such as natural language processing and training and validations studies.

Accurate Scaling and Linking. Scaling and linking are essential activities for producing valid scores and score interpretations for the Smarter Balanced assessments. Scaling activities include item calibration and creation of a standardized scale on which scores are reported. A sound scaling and linking design and representative student samples are critical precursors to conducting the scaling analysis. A sound linking design includes criteria such as content-representative and blueprint-conforming test forms being administered, particularly with respect to common/anchor linking items. Evaluating the adequacy of these scaling and linking activities includes steps that confirm the hypothesized dimensionality of the assessments and the viability of a single construct (dimension) across grades, the performance of different IRT scaling models, scrutiny of the linking results, and potentially examining the invariance of the equating across subgroups of students. A major

assumption for the use of more traditional scaling methods is that a given test is essentially unidimensional, consisting of a major dimension along with some minor ones. The nature of the change over grade levels is characterized in the common items given across grade levels that are used to construct the vertical scale. The viability of a vertical scale depends on this major dimension being consistent across levels of the test. The influence of the minor dimensions will determine how the construct shifts across grade levels. Since the ELA/literacy and mathematics assessments were vertically linked across grades, evidence concerning the nature of change in the construct over levels of the test and its plausibility are necessary. A “cross validation study,” where an independent third party replicates the scaling and linking, provides an important validity check on the accuracy of the equating. Once the calibrated item pool is available, a choice of IRT scoring methods is necessary, such as maximum likelihood estimation (Thissen & Wainer, 2001), that forms the basis for achievement level reporting.

Appropriate Standard Setting. When achievement-level (i.e., proficiency) standards are set on tests, scale scores often become less important than the proficiency classifications students receive which are the central focus of many accountability systems. There are many different methods for setting standards, but regardless of the method used, there must be sufficient validity evidence to support the classification of students into achievement levels. The Smarter Balanced summative assessments used achievement levels, some of which signified “on track” to “college readiness” (grades 3-8) or “college ready” (grade 11). An additional element was the articulation of cut points across grade levels in the context of a vertical scale. The primary assumption here is that the cut points increase across grade levels in a logical progression that reflects increased levels of achievement that ultimately culminate in “readiness” in grade 11. Articulated achievement levels means the proficiency cut scores maintain some consistent level of stringency or pattern across grades.

Gathering and documenting validity evidence for standards set on educational tests can be categorized into three categories—procedural, internal, and external (Kane, 1994; 2001). Procedural evidence for standard setting “focuses on the appropriateness of the procedures used and the quality of the implementation of these procedures” (Kane, 1994, p. 437). The selection of qualified standard-setting panelists, appropriate training of panelists, clarity in defining the tasks and goals of the study, appropriate data collection procedures, and proper implementation of the method are all examples of procedural evidence. Internal evidence for evaluating standard-setting studies focuses on the expected consistency of results if the study was replicated. A primary criterion is the standard error of the cut score. However, calculation of this standard error is difficult due to dependence among panelists’ ratings and practical factors (e.g., time and expense in conducting independent replications). Oftentimes, evaluations of the variability across panelists within a single study and the degree to which this variability decreases across subsequent rounds of the study are presented as internal validity evidence. However, as Kane (2001) pointed out:

A high level of consistency across participants is not to be expected and is not necessarily desirable; participants may have different opinions about performance standards. However, large discrepancies can undermine the process by generating unacceptably large standard errors in the cut scores and may indicate problems in the training of participants (p. 73).

In addition to simply reporting the standard error of the cut score, Kane (2001) suggested that consistency can be evaluated across independent panels, subgroups of panelists, or assessment tasks (e.g., item formats), or by using generalizability theory to gauge the amount of variability in panelists’ ratings attributed to these different factors. Another source of internal validity evidence proposed by Kane was to evaluate the performance of students near the cut score on specific items to see if their performance was consistent with the panelists’ predictions. External validity evidence

for standard setting involves studying the degree to which the classifications of students based on test scores are consistent with other measures of their achievement in the same subject area. External validity evidence includes classification consistency across different standard-setting methods applied to the same test, to tests of mean differences across examinees classified in different achievement levels on other measures of achievement, and the degree to which external ratings of student performance are congruent with their test-based achievement-level classifications. External validity evidence is particularly important for validating the “college and career readiness” standards set on the summative assessments. A number of measures for determining college readiness already exists. The degree to which the constructs measured by these external assessments overlap with the Smarter Balanced summative assessments and the degree to which their definitions of readiness are similar (or different) should be addressed by Smarter Balanced.

Attention to Fairness, Equitable Participation, and Access. Chapter 3 of the *Standards* (p. 49-70) addresses fairness in testing. The intent of the Smarter Balanced system is to provide additional flexibility and remove construct-irrelevant barriers that prevent students from taking the test or demonstrating their best performance. Construct irrelevant barriers can be minimized through test design and testing adaptations. Evidence-centered design, item specifications, usability, accessibility, and accommodations guidelines, bias and sensitivity guidelines, and reviews by content developers are all used to develop items and tasks that ensure the targeted constructs are measured accurately. A critical aspect of access is the ability to deliver items, tasks, and the collection of student responses in a way that maximizes validity for each student. Equitable participation and access ensures that all students can take the test in a way that allows them to comprehend and respond appropriately. This includes, but is not limited to, English Language Learners (ELLs), students with disabilities, and ELLs with disabilities. The *Standards* also specify an aspect of fairness as a lack of measurement bias. Characteristics of items that are construct irrelevant can affect the performance by members of some identifiable subgroups, which is called differential item functioning. Many methods exist for investigating differential item functioning statistically. For an item exhibiting DIF, additional investigation is required in order to conclude it is biased.

Validating “On-Track/Readiness” “On-track” denotes notions concerning expectations and adequate levels of growth being demonstrated. Studies related to expected growth will be conducted after two iterations of operational testing have been conducted. States use a variety of growth models, and the Consortium is not recommending or discouraging any specific model. Growth is defined as improvement in performance for a given group of students over time—such as improvement from grade 6 to grade 7. Vertical scales can facilitate the measurement of student growth and permit direct comparisons of change using scale scores (or status) across different grades or change within a year.

College and career readiness may have substantial overlap since employers and colleges have similar perspectives on the level of knowledge and skills required for entry (Achieve, 2004; ACT, 2006). However, others have argued the benchmarks for college and career readiness will be very different (Camara, 2013; Loomis, 2011). On-track for college readiness implies the acquisition of knowledge and the mastery of specific skills deemed important as students progress through elementary, middle, and high school that are stipulated in the Common Core State Standards. Validity studies such as content alignment can be used to confirm that the Smarter Balanced assessments are targeting the correct Common Core State Standards and adequately represent these standards. However, studies of this type do not confirm that the Common Core State Standards actually contain the appropriate knowledge and skills to support college and career readiness (Sireci, 2012). At the higher education level, Conley, Drummond, de Gonzalez, Rooseboom, and Stout (2011) conducted a national survey of postsecondary institutions to evaluate the degree to which the grade 11 Common Core State Standards contain the knowledge and skills

associated with college readiness. They found that most of approximately 2,000 college professors rated the Common Core State Standards as highly important for readiness in their courses. Similarly, Vasavada, Carman, Hart, and Luisser (2010) found strong alignment between College Board assessments of college readiness and the Common Core State Standards. The additional evidence required for readiness is evidence that these standards reflect the appropriate prerequisite skills in mathematics and ELA/literacy that are needed to bypass remedial college courses and to successfully begin postsecondary education or a career. Other validity evidence based on test content is the content overlap (alignment) studies that will be undertaken to gauge the similarity of knowledge and skills measured across the Summative assessments and external assessments that are used to evaluate the readiness standards (Sireci, 2012). Postsecondary admissions tests (e.g., ACT, SAT) and college placement tests (e.g., Accuplacer, Advanced Placement, & Compass) can be used in concurrent and predictive validity studies. This requires the overlap in the skills measured to be identified to derive the proper inferences.

The degree to which other measures of college readiness benchmarks are consistent with the Smarter Balanced readiness standards can be examined. Camara (2013) listed seven criteria that have been or could be used for setting or evaluating college readiness benchmarks on the Smarter Balanced assessments. These criteria are:

- persistence to second year;
- graduation or completion of a degree or certification program;
- time to degree completion (e.g., six years to earn a bachelor's degree);
- placement into college credit courses;
- exemption from remediation courses;
- college grades in specific courses; and
- college grade-point average.

Validity evidence based on relations to other external variables for the purpose of classifying students as college ready can involve both correlation type studies and classification consistency analyses.

The college and career readiness standard is intentionally integrated with the “on-track” standards set at the lower grade levels with the intended consequence that the system better prepares students for college or careers by the time they graduate high school. These college and career readiness outcomes can be appraised using trends in college completion and remedial course enrollments over time, and by surveying secondary and postsecondary educators about students’ proficiencies. The recommended studies based on test consequences for college and career readiness purposes should include teacher surveys regarding changes in student achievement and preparedness over time and changes in their instruction over time. Students can be surveyed regarding college and career aspirations. Student and teacher samples that are representative at the state level would suffice for these studies. Validity evidence based on the consequences of the college and career readiness standard should involve analysis of secondary and postsecondary enrollment and persistence, changes in course-taking patterns over time, and teacher retention for teachers in mathematics and ELA/literacy.

Studies of the relationship of Smarter scores to college course enrollment, grades and course completion will be conducted as students using the Consortium tests enter college. The Consortium has created career cluster readiness frameworks Smarter Balanced, (2013) to inform alignment of test content to career-related skills and to aid in score interpretation.

Adequate Test Security. Test security is a prerequisite to validity. As described by NCME (2012), “When cheating occurs, the public loses confidence in the testing program and in the educational system which may have serious educational, fiscal, and political consequences.” Threats to test security include cheating behaviors by students, teachers, or others who have unwarranted access to testing materials. A lack of test security may result in the exposure of items before tests are administered, students copying or sharing their answers, or changing students’ answers to test questions in fixed-form tests. Many proactive steps can be taken to reduce, eliminate, and evaluate cheating. The first step is to keep confidential test material secure and have solid procedures in place for maintaining the security of paper and electronic materials. The NCME (2012) document on data integrity outlined several important areas of test security. These areas include procedures that should be in place before, during, and after testing. The activities prior to testing include securing the development and delivery of test materials. During testing, activities include adequate proctoring to prevent cheating, imposters, and other threats. After testing, checking social media for item content and the forensic analysis of students’ responses and answer changes and aberrant score changes over time are also beneficial. The goal of these security activities is to ensure that test data are “free from the effects of cheating and security breaches and represent the true achievement measures of students who are sufficiently and appropriately engaged in the test administration” (NCME, 2012, p. 3).

The evaluation of the test security procedures of the assessments involved a review of the test security procedures and data forensics by Smarter Balanced. The NCME (2012) document on test data integrity suggests that security policies should address the following:

... staff training and professional development, maintaining security of materials and other prevention activities, appropriate and inappropriate test preparation and test administration activities, data collection and forensic analyses, incident reporting, investigation, enforcement, and consequences. Further, the policy should document the staff authorized to respond to questions about the policy and outline the roles and responsibilities of individuals if a test security breach arises. The policy should also have a communication and remediation response plan in place (if, when, how, who) for contacting impacted parties, correcting the problem and communicating with media in a transparent manner (p. 4).

With adaptive test administration, the probability of students receiving the same items at similar times is low, and the probability of answer copying is very low. However, consistent with other CAT programs, item exposure rates should be carefully monitored on an ongoing basis. Rules for rotating items out of the summative assessment with comparatively high exposure rates are needed. Due to the nature of performance tasks that are more memorable and subject to practice effects, they will need to be replaced or transitioned frequently.

Summary of Essential Validity Evidence based on the Smarter Pilot- and Field Tests

Other chapters of the Smarter Balanced technical report describe the evidence and studies performed to date for the Smarter Balanced Assessment Field Test. Tables 2 to 10 list the essential validity elements and associated evidence types for each one. For example, Table 2 presents the essential validity element for “Careful Test Construction. It lists the types of validation evidence associated with that element in terms of a short label, and provides the associated evidence source. For the evidence source, the chapter and section in parenthesis refers to this Technical Report. When appropriate, other Smarter Balanced documentation or reports are listed in italics. The reader will need to make a judgment as to the importance of the essential validity element presented and the number, quality, and types of supporting evidence.

Table 2. Essential Validity Evidence for the Summative and Interim Assessments for Careful Test Construction.

Evidence Type	Evidence Source
Theory of Action/testing purposes clearly stated	<i>Smarter Balanced Theory of Action</i> , Introduction (Theory of Action)
Evidence centered design implemented	Test Design, (Evidence Centered Design), <i>Smarter Balanced Bibliography</i> , <i>General Item Specifications</i>
Test specifications sufficiently documented	Test Design, (Operational Summative Assessment Blueprints and Specifications), <i>Performance Task Specification</i> , <i>Mathematics Performance Task Specifications</i>
Construct definition	Test Design, (Operational Summative Assessment Blueprints and Specifications), <i>ELA/literacy Content Specifications</i> , <i>Mathematics Content Specifications</i>
Item writers appropriately recruited and trained	<i>Mathematics Performance Task Specifications</i>
Items adhere to item writing style guidelines	<i>General Item Specifications</i>
Items reviewed for content quality and technical adequacy	Field Test Data Step and Classical Test Analysis (Item Flagging Criteria for Content Data Review)
Content validity/alignment studies	<i>General Item Specifications</i> ; <i>Smarter Balanced Assessment Consortium Alignment Study</i> , <i>HumRRO</i>
Sensitivity reviews	Test Fairness, (Definitions for Validity, Bias, Sensitivity, and Fairness), <i>Standards for Educational and Psychological Testing</i> , <i>Smarter Balanced Bias and Sensitivity Guidelines</i> , <i>ETS Guidelines for Fairness Review of Assessments</i>
Test booklets conform to test blueprints	Field Test Design, Sampling, and Administration (Numbers and Characteristics of Items and Students Obtained in the Field Test), Field Test Design, Sampling, and Administration (Field Test Delivery Modes), <i>Smarter Balanced Adaptive Item Selection Algorithm Design Report</i> , <i>General Item Specifications</i> , <i>Simulations studies from AIR and CRESST</i> .
Data review (Classical)	Pilot Study, (Pilot Classical Test Results); Field technical review.
Item selection/delivery based on content criteria	Test Design, (Operational Summative Assessment Blueprints and Specifications), Field Test Design,

Evidence Type	Evidence Source
	Sampling, and Administration, (Field Test Delivery Modes), <i>Adaptive Selection Algorithm</i> (Cohen & Albright, 2014)
IRT Item calibration	Pilot Test, (Dimensionality Study), Pilot Test, (IRT Model Comparison), Field Test IRT Scaling and Linking Analyses, (All sections)

Table 3. Essential Validity Evidence for the Summative and Interim Assessments for Adequate Measurement Precision (Reliability).

Evidence Type	Evidence Source
Test reliability (Internal Consistency)	Data Step and Classical Test Analysis (Field Test Results); Simulation studies by AIR and CRESST.
IRT item fit	Field Test IRT Scaling and Linking Analyses (Horizontal and Vertical Scaling Results), Pilot Test, (Item Response Theory [IRT] Model Comparison)
Conditional standard error of measurement (CSEM) for ability	Field Test IRT Scaling and Linking Analyses (Horizontal and Vertical Scaling Results) Simulation studies by AIR and CRESST.
Standard error of measurement (Classical)	Data Step and Classical Test Analysis (Field Test Results)
IRT test information	Field Test IRT Scaling and Linking Analyses (Horizontal and Vertical Scaling Results)
Generalizability studies	
Cut-score decision consistency and accuracy	Simulation studies by AIR and CRESST.

Table 4. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Test Administration.

Evidence Type	Evidence Source
Availability of training and practice modules	<i>Test Administration Manual</i>
Clearly defined instructions	<i>Test Administration Manual, Field Test Design, Sampling and Administration, (Field Test Administration and Security), Smarter Balanced Technical Specifications Manual, Calculator Availability Information for 2014 Field Test</i>
Test delivery system functioned as expected	<i>Smarter Balanced “Tests of the Test” Successful: Field Test Provides Clear Path Forward; Simulation studies by AIR and CRESST.</i>

Table 5. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Scoring

Evidence Type	Evidence Source
Sufficient levels of rater agreement	<i>Pilot Test: (Evaluation of Reliability and Validity for Automated Scoring Models in the Pilot), Field Test: Automated Scoring Research Studies</i>
Scoring rubrics for constructed-response items are reviewed	<i>Smarter Balanced Scoring Guide for Selected Short-Text Mathematics Items (Field Test 2014), Performance Tasks Specifications</i>
Adaptive item selection algorithm documented	<i>Smarter Balanced Adaptive Item Selection Algorithm Design Report (Cohen & Albright, 2014)</i>
Content conforming tests are delivered	Field Test Design, Sampling and Administration (Numbers and Characteristics of Items and Students Obtained in the Field Test)
Rationale, development, and rater agreement for automated scoring	<i>Smarter Balanced Pilot Automated Scoring Research Studies, Field Test: Automated Scoring Research Studies</i>

Table 6. Essential Validity Evidence for the Summative and Interim Assessments for Accurate Scaling and Linking.

Evidence Type	Evidence Source
Sample is representative	Field Test Design, Sampling and Administration, (Sampling Results)
Rationale for IRT model choice is provided	Pilot Test (Dimensionality Study), Pilot Test (Item Response Theory Model Comparison)
Calibration and linking design is appropriate	Field Test IRT Scaling and Linking Analyses, (Vertical Scaling: Linking Across Multiple Grades), Field Test Design, Sampling and Administration, (Field Test), Field Test Design, Sampling and Administration (Field Test Student Sampling Design)
Accurate IRT horizontal and vertical scaling met	Field Test IRT Scaling and Linking Analyses, (All Sections)
Accurate equating methods applied	Field Test Design, Sampling and Administration, (Field Test), Field Test Design, Sampling, and Administration (Field Test Student Sampling Design), Field Test IRT Scaling and Linking Analyses, (Assumptions and Interpretive Cautions Concerning Vertical Scales) Field Test IRT Scaling and Linking Analyses, (Horizontal and Vertical Scaling Results)
Assumptions for establishing vertical scale are met	Field Test IRT Scaling and Linking Analyses, (Assumptions and Interpretive Cautions Concerning Vertical Scales), Field Test IRT Scaling and Linking Analyses, (Vertical Linking Procedures)

Table 7. Essential Validity Evidence for the Summative and Interim Assessments for Appropriate Standard Setting.

Evidence Type	Evidence Source
Justification of standard setting method(s)	Achievement Level Setting (The Bookmark Procedure), <i>Achievement Level Setting Plan</i>
Panelist recruitment and training	Achievement Level Setting (Recruitment and selection of panelists)
Clarity of goals/tasks	<i>Achievement Level Setting Plan</i>
Clear achievement level descriptors	<i>Initial Achievement Level Descriptors and College Content-Readiness Policy, Achievement Level Setting (Achievement Level Descriptors), Interpretation and Use of Scores and Achievement Levels</i>
Appropriate data collection	Achievement Level Setting (The Bookmark Procedure), <i>Achievement Level Setting Plan</i>
Implementation	Achievement Level Setting (All Sections), <i>Achievement Level Setting Plan</i>
Panelist confidence	Achievement Level Setting (Round-by-round item review and discussion)
Sufficient documentation	Achievement Level Setting (All Sections), <i>Achievement Level Setting Plan</i>
Sufficient inter-panelist consistency	Achievement Level Setting (Round-by-round item review and discussion)
Across grade level articulation	Achievement Level Setting (Design and Implementation of the Cross-Grade Review Committee)
Reasonableness of achievement standards	Achievement Level Setting (All Sections), <i>Statements of Support: Achievement Level Setting, Achievement Level Descriptors and College Content-Readiness, Achievement Level Setting Statements of Support</i>

Table 8. Essential Validity Evidence for the Summative and Interim Assessments for Attention to Fairness, Equitable Participation and Access.

Evidence Type	Evidence Source
DIF Analysis	Field Test Datastep and Classical Test Analysis (Differential Item Functioning (DIF) Analyses for the Calibration Item Pool)
Equitable Participation	<i>Usability, Accessibility, and Accommodations Guidelines</i>
Universal Design, Assessment Supports, and Accommodations	<i>Usability, Accessibility, and Accommodations Guidelines; Test Fairness (Usability, Accessibility, and Accommodations Guidelines: Intended Audience and Recommended Applications), General Item Specifications, Signing Guidelines, Tactile Accessibility Guidelines</i>
Support for English Language Learners	<i>Guidelines for Accessibility for English Language Learners</i>
Bias and Sensitivity	<i>Smarter Balanced Assessment Consortium: Bias and Sensitivity Guidelines, Test Fairness (Definitions for Validity, Bias, Sensitivity, and Fairness)</i>

Table 9. Essential Validity Evidence for the Summative and Interim Assessments for Validating “On-Track/Readiness”.

Evidence Type	Evidence Source
Relationships with External Variables/Tests	Field Test Design, Sampling, and Administration (Linking PISA and NAEP to Smarter Balanced Assessments), Data Step and Classical Test Analysis (Field Test Results)
Operational definition of college content-readiness	<i>Study of the Relationship Between the Early Assessment Program and the Smarter Balanced Field Tests, ELA/literacy Achievement Level Descriptors and College Content-Readiness Policy, Mathematics Achievement Level Descriptors and College Content-Readiness Policy, Reaching the Goal: The Applicability and Importance of the Common Core State Standards to College and Career Readiness</i>

Table 10. Essential Validity Evidence for the Summative and Interim Assessments for Adequate Test Security.

Evidence Type	Evidence Source
Test security procedures and exceptions escalation documented	<i>Online Field Test Administration Manual for Spring 2014 Field Tests of English Language Arts/Literacy and Mathematics</i> , Field Test Design, Sampling, and Administration, (Field Test Administration and Security), <i>The Smarter Balanced Technology Strategy Framework and Testing Device Requirements</i>

The *Standards'* Five Primary Sources of Validity Evidence

The five sources of validity evidence serve as organizing principles and represent a comprehensive framework for evaluating validity for Smarter Balanced. These sources of validity evidence are intended to emphasize different aspects of validity. However, since validity is a unitary concept, they do not constitute distinct types of validity. These five sources of validity evidence consist of (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing. They are briefly described below:

1. Validity evidence based on *test content* refers to traditional forms of content validity evidence, such as the rating of test specifications and test items (Crocker, Miller, & Franks, 1989; Sireci, 1998), as well as “alignment” methods for educational tests that evaluate the interactions between curriculum frameworks, testing, and instruction (Rothman, Slattery, Vranek, & Resnick, 2002; Bholá, Impara & Buckendahl, 2003; Martone & Sireci, 2009). The degree to which (a) the Smarter Balanced test specifications captured the Common Core State Standards and (b) the items adequately represent the domains delineated in the test specifications, were demonstrated in the alignment studies. The major assumption here is that the knowledge, skills, and abilities measured by the Smarter Balanced assessments are consistent with the ones specified in the Common Core State Standards. Administration and scoring can be considered as aspects of content-based evidence. With computer adaptive testing, an extra dimension of test content is to ensure that the tests administered to students conform to the test blueprint.
2. Validity evidence based on *response processes* refers to “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA et al., 1999 p. 12). This evidence might include documentation of such activities as
 - interviewing students concerning their responses to test items (i.e., speak alouds);
 - systematic observations of test response behavior;
 - evaluation of the criteria used by judges when scoring performance tasks, analysis of student item-response-time data, features scored by automated algorithms; and
 - evaluation of the reasoning processes students employ when solving test items (Emberetson, 1983; Messick, 1989; Mislevy, 2009).

This type of evidence was used to confirm that the Smarter Balanced assessments are measuring the cognitive skills that are intended to be the objects of measurement and that students are using these targeted skills to respond to the items.

3. Validity evidence based on *internal structure* refers to statistical analyses of item and score subdomains to investigate the primary and secondary (if any) dimensions measured by an assessment. Procedures for gathering such evidence include factor analysis or multidimensional IRT scaling (both exploratory and confirmatory). With a vertical scale, a consistent primary dimension or construct shift across the levels of the test should be maintained. Internal structure evidence also evaluates the “strength” or “salience” of the major dimensions underlying an assessment using indices of measurement precision such as test reliability, decision accuracy and consistency, generalizability coefficients, conditional and unconditional standard errors of measurement, and test information functions. In addition, analysis of item functioning using Item Response Theory (IRT) and differential item functioning (DIF) fall under the internal structure category. For Smarter Balanced, a dimensionality study was conducted in the Pilot Test to determine the factor structure of the assessments and the types of scales developed as well as the associated IRT models used to calibrate them.
4. Evidence based on *relations to other variables* refers to traditional forms of criterion-related validity evidence such as concurrent and predictive validity, as well as more comprehensive investigations of the relationships among test scores and other variables such as multitrait-multimethod studies (Campbell & Fiske, 1959). These external variables can be used to evaluate hypothesized relationships between test scores and other measures of student achievement (e.g., test scores and teacher grades), the degree to which different tests actually measure different skills and the utility of test scores for predicting specific criteria (e.g., college grades). This type of evidence is essential for supporting the validity of certain inferences based on scores from the Smarter Balanced assessments for certifying college and career readiness, which is one of the primary test purposes. A subset of students who took NAEP and PISA items also took Smarter Balanced items and performance tasks. A summary of the resulting item performance for NAEP, PISA, and all Smarter Balanced items was conducted.
5. Finally, evidence based on *consequences of testing* refers to the evaluation of the intended and unintended consequences associated with a testing program. Examples of evidence based on testing consequences include investigations of adverse impact, evaluation of the effects of testing on instruction, and evaluation of the effects of testing on issues such as high school dropout rates. With respect to educational tests, the *Standards* stress the importance of evaluating test consequences. For example, they state,

When educational testing programs are mandated . . . the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the use of the test, both intended and unintended, should also be examined by the test user (AERA et al., 1999, p. 145).

Investigations of testing consequences relevant to the Smarter Balanced goals include analyses of students’ opportunity to learn with regard to the Common Core State Standards, and analyses of changes in textbooks and instructional approaches. Unintended consequences, such as changes in instruction, diminished morale among teachers and students, increased pressure on students leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging, can be evaluated.

Purposes of the Smarter Balanced System for Summative, Interim, and Formative Assessments

The Smarter Balanced purpose statement refers to three categories consisting of summative, interim, and formative assessment resources. To derive the statements of purpose listed below, panels consisting of Smarter Balanced leadership, including the Executive Director, Smarter Balanced staff, Dr. Stephen Sireci and key personnel from Consortium states were convened.

The purposes of the Smarter Balanced summative assessments are to provide valid, reliable, and fair information concerning:

- 1) Students' ELA/literacy and mathematics achievement with respect to those CCSS measured by the ELA/literacy and mathematics summative assessments.
- 2) Whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on track for achieving college readiness.
- 3) Whether grade 11 students have sufficient academic proficiency in ELA/literacy and mathematics to be ready to take credit-bearing college courses.
- 4) Students' annual progress toward college and career readiness in ELA/literacy and mathematics.
- 5) How instruction can be improved at the classroom, school, district, and state levels.
- 6) Students' ELA/literacy and mathematics proficiencies for federal accountability purposes and potentially for state and local accountability systems.
- 7) Students' achievement in ELA/literacy and mathematics that is equitable for all students and subgroups of students.

Providing valid, reliable, and fair information about students' ELA/literacy and mathematics achievement with respect to the Common Core State Standards as measured by the summative assessments is central. Validity evidence to support this purpose derives from at least three sources—test content, internal structure, and response processes. With respect to test content, evidence confirming that the content of the assessments adequately represents the Common Core State Standards to be measured in each grade and subject area is essential. Content domain representation and congruence to the Common Core State Standards must be substantiated. Validity evidence based on internal structure involves analysis of item response data to confirm that the dimensionality of the data matches the intended structure and supports the scores that are reported. Measures of reliability, test information, and other aspects of measurement precision are also relevant. Validity evidence based on response processes should confirm that the items designed to measure higher-order cognitive skills are tapping into those targeted skills.

The Smarter Balanced assessments focus on the provision of valid, reliable, and fair information concerning whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on-track for achieving college readiness. Secondly, the intent is to provide valid, reliable, and fair information about whether grade 11 students have sufficient academic proficiency in ELA and mathematics to be ready to take credit-bearing college courses or are career-ready. These two purpose statements reflect that the Smarter Balanced summative assessments will be used to classify students into achievement levels. Before grade 11, one achievement level will be used at each grade to signal whether students are “on-track” to college or career readiness. At grade 11, the achievement levels will include a “college and career readiness” category. These classification decisions require validation that can be derived from four sources—test content, internal structure, relations with external variables, and testing consequences.

The purposes of the Smarter Balanced interim assessments are to provide valid, reliable, and fair information about:

- 1) Student progress toward mastery of the skills in ELA/literacy and mathematics measured by the summative assessment.
- 2) Student performance at the claim or cluster of Assessment Targets so teachers and administrators can track student progress throughout the year and adjust instruction accordingly.
- 3) Individual and group (e.g., school, district) performance at the claim level in ELA/literacy and mathematics to determine how teaching and learning can best be targeted.
- 4) Student progress toward the mastery of skills measured in ELA/literacy and mathematics across all students and subgroups of students.

The Smarter Balanced interim assessments differ from the summative assessments in that they are optional, non-secure components that can be administered multiple times within a school year and are designed to provide information at a finer level of detail with respect to students' strengths and weaknesses in relation to the Common Core State Standards. The interim assessments are intended to help teachers focus assessment on the most relevant aspects of classroom instruction at a particular point in time. They are also intended to play a role in professional development, particularly in cases in which teachers can determine how scoring rubrics align with the content standards and have the opportunity to score student responses to items.

The purposes of the Smarter Balanced formative assessment resources are to provide measurement tools and resources to:

- 1) Improve teaching and learning.
- 2) Monitor student progress throughout the school year.
- 3) Help teachers and other educators align instruction, curricula, and assessment.
- 4) Assist teachers and other educators in using the summative and interim assessments to improve instruction at the individual student and classroom levels.
- 5) Illustrate how teachers and other educators can use assessment data to engage students in monitoring their own learning.

The Formative Assessment Resources are not assessments per se, and so the evidence in support of their intended purposes extends beyond the five sources of validity evidence and requires a program evaluation approach. Tables 11 and 12 illustrate the validation framework for the summative and interim assessments by cross-referencing the purpose statements for each component with the five sources of validity evidence. The check marks in the cells indicate the type of evidence that could be used for validating a specific purpose. While this presentation is general, it is useful for understanding which sources of validity evidence are most important for specific test purposes. For example, for purposes related to providing information about students' knowledge and skills, validity evidence based on test content is critical. For purposes related to classifying students into achievement categories such as "on-track" or "college-ready", validity evidence based on internal structure is needed since evidence of this type also relies on sufficient level of decision consistency and accuracy being demonstrated.

Table 11. Validity Framework for Smarter Balanced Summative Assessments.

Purpose	Source of Validity Evidence				
	Content	Internal Structure	Relations with External Variables	Response Processes	Test Consequences
Report achievement with respect to the CCSS* as measured by the ELA/literacy and mathematics summative assessments	✓	✓	✓	✓	
Assess whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on track for college readiness	✓	✓	✓		✓
Assess whether grade 11 students have sufficient academic proficiency in ELA/literacy and mathematics to be ready to take credit-bearing college courses	✓	✓	✓		✓
Measure students' annual progress toward college and career readiness in ELA/literacy and mathematics	✓	✓	✓		✓
Inform how instruction can be improved at the classroom, school, district, and state levels	✓				✓
Report students' ELA/literacy and mathematics proficiency for Federal accountability purposes and potentially for state and local accountability systems	✓	✓	✓		✓
Assess students' achievement in ELA/literacy and mathematics in a manner that is equitable for <i>all</i> students and subgroups of students	✓	✓	✓	✓	✓

Note: *CCSS denotes Common Core State Standards.

Table 12. Validity Framework for Smarter Balanced Interim Assessments.

Purpose	Source of Validity Evidence				
	Content	Internal Structure	Relations with External Variables	Response Processes	Test Consequences
Assess student mastery of the skills and knowledge measured in ELA/literacy and mathematics	✓	✓		✓	
Assess students' performance at the claim level or finer so teachers and administrators can track student progress throughout the year and adjust instruction accordingly	✓	✓			✓
Assess individual and group (e.g., school, district) performance at the claim level in ELA/literacy and mathematics to determine whether teaching and learning are on target	✓	✓	✓		✓
Measure student progress toward the mastery of skills measured in ELA/literacy and mathematics across all subgroups	✓	✓	✓	✓	✓

Evidence Using the Five Primary Sources of Validity Framework

Table 13 lists the evidence type, the associated Smarter Balanced evidence demonstrated to date (or not), and the relevancy of the *Standards* five primary validity evidence sources. It further cross-classifies each piece of evidence with the Smarter Balanced Summative, Interim, and Formative Test Purposes. The evidence demonstrated lists the relevant chapter of the Technical Report or the relevant external documents. Even if several sources of evidence are presented, the reader must still make a judgment whether sufficient supporting evidence has been offered. For instance, the reader may question if, for vertical scaling, there is an increase in the difficulty of the assessments as the grade level increases, with generally greater student proficiency demonstrated in higher grades relative to lower grades. In the case where the evidence source in the table is blank simply means no evidence is available to date. The full complement of validity evidence is ambitious in both its scope and the resources required to fulfill it. The table is also useful in that it serves to identify current gaps in the ongoing validity argument and identify the sorts of evidence needed going forward, as proposed in the Smarter Balanced comprehensive research agenda (Sireci, 2012).

Table 13. Listing of Evidence Type, Evidence Source, and Primary Validity Source for Summative, Interim, and Formative Test Purposes.

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative						
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5		
Content validity and alignment	Test Design	1	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	
Sensitivity and bias review	Test Design	1	✓					✓	✓				✓							
Evidence Centered Design	Test Design, <i>General Item Specifications</i>	1, 2	✓	✓	✓	✓					✓	✓	✓							
Subdomain scores (e.g., claims)	Test Design	1, 3										✓	✓							
Scoring (raw scores)	Datastep and Classical Test Analysis	1, 3	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	
Standard setting	Achievement Level Setting	1, 3, 4	✓	✓	✓	✓	✓	✓	✓				✓		✓			✓		
Test construction practices	Test Design	1, 3	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓				✓	✓	✓
Fairness	Test Fairness	1, 2, 3, 4, 5	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	
Scope and sequence of curriculum		1, 5					✓					✓	✓		✓	✓	✓	✓	✓	
Test administration	Field Test Design, Sampling, and Administration	1, 5	✓						✓				✓		✓					

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative				
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5
Equitable participation & access	Test Fairness, Field Test Design, Sampling, and Administration, and <i>Usability, Accessibility, and Accommodations Guidelines</i>	1, 5	✓						✓			✓					✓	✓
Test accommodations	Test Design, Test Fairness, <i>Usability, Accessibility, and Accommodations Guidelines</i>	1, 5	✓					✓	✓			✓						
Formative resources development and implementation		1, 5												✓	✓	✓	✓	✓
Cognitive skills, think-aloud protocols		2	✓				✓	✓			✓							
Item response time		2	✓				✓	✓			✓							
Horizontal and vertical scales	Field Test IRT Scaling and Linking Analyses	3		✓	✓	✓		✓	✓		✓							
Decision consistency and accuracy	Achievement Level Setting	3		✓	✓	✓		✓	✓		✓		✓					

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative					
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	
IRT fit analysis	Field Test IRT Scaling and Linking Analyses	3		✓		✓		✓	✓		✓								
Reliability and standard error estimation	Field Test IRT Scaling and Linking Analyses	3	✓	✓	✓	✓	✓	✓	✓		✓	✓							
	and Datastep and Classical Test Analysis																		
Reliability of aggregate statistics		3	✓					✓	✓				✓				✓		
Generalizability studies		3		✓	✓	✓		✓											
Item parameter drift		3		✓	✓	✓					✓								
Test dimensionality	Pilot Test	3	✓			✓	✓				✓	✓	✓						
CAT algorithm		2, 3		✓	✓							✓	✓						
Mode comparability		3			✓														
Automated scoring	Pilot Test, <i>Field Test: Automated Scoring Research Studies</i>	3	✓								✓								

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative					
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	
Invariance of test structure		3	√					√	√				√						
Test security	Field Test Design, Administration, and Sampling	3, 4	√	√	√			√			√								
Convergent/discriminant validity		3, 4	√					√			√	√	√						
Differential item functioning	Test Fairness	3, 5	√					√					√						
Sensitivity to instruction		4	√	√	√	√	√	√	√		√	√	√	√					
Criterion-related validation of on-track		4		√															
Criterion-related studies of change in achievement/growth		4		√		√		√	√		√	√	√						
Criterion-related validation of readiness		4		√	√	√		√			√								
Differential predictive validity		4	√					√	√				√						
Group differences	Test Fairness, Datastep and Classical Test Analysis	4		√	√	√		√	√				√						
Classroom artifacts		4, 5					√					√			√	√	√	√	√

Evidence Type	Evidence	Primary Validity Source	Summative							Interim				Formative					
	Source		1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	
Perspective of postsecondary educators		5			✓	✓													
College enrollment, dropout, courses taken		5					✓		✓				✓						
Teacher morale/perception of test utility		5					✓	✓			✓	✓	✓	✓		✓		✓	✓
Teacher perception on changes in student learning		5		✓	✓	✓	✓		✓		✓	✓	✓	✓		✓	✓		✓
Student perspective		5			✓			✓											
Educator interviews, and focus groups		5		✓	✓														
Score report utility and clarity		5	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓
Score report usage rates		5					✓								✓	✓	✓	✓	✓
Follow-up on specific student decisions		5		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓		✓
Interim usage statistics		5									✓	✓	✓	✓					
High efficacy users of interim assessments		5									✓	✓	✓	✓					
Formative usage statistics		5														✓	✓	✓	✓
Collaborative leadership		5														✓		✓	✓

Evidence Type	Evidence		Summative							Interim				Formative				
	Source	Primary Validity Source	1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5
Utility of formative assessment		5												✓	✓	✓	✓	✓
Formative assessment student perception		5												✓	✓	✓	✓	✓
Parent perception of formative assessment		5												✓				✓
Critique of Theory of Action		5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Comparison with NAEP, PISA, TIMSS		1, 3, 4, 5																
Summary of validity evidence supporting seven Theory of Action principles		5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note: Primary Validity Source: 1=Test Content, 2=Response Processes, 3=Internal Structure, 4=Relations to other variables, 5=Testing consequences

Conclusion for Field Test Validity Results

Validation is an ongoing, essentially perpetual endeavor in which additional evidence can be provided but one can never absolutely “assert” an assessment is perfectly valid (Haertel, 1999). This is particularly true for the many purposes typically placed on tests. Program requirements are often subject to change and the populations assessed change over time. Nonetheless, at some point decisions must be made regarding whether sufficient evidence exists to justify the use of a test for a particular purpose. A review of the purpose statements and the available validity evidence determines the degree to which the principles outlined here have been realized. Most of this report has focused on describing the essential validity elements that partially provide this necessary evidence. The existing evidence was organized into an essential validity framework that can be used to evaluate whether professional testing standards consistent with a Field Test have been met. The essential validity elements presented here constitute critical evidence and are elements that are “relevant to the technical quality of a testing system” (AERA et al., 1999, p. 17). The evidence in support of these essential elements highlighted here referenced the relevant information from the other chapters of this technical report or referenced specific Smarter Balanced supporting documents, the products of other Smarter Balanced workgroups or outside groups. The types of evidence presented here are more consistent with those supporting a Field Test prior to operational administration. Many types of evidence from external sources could not reasonably be collected when so many parts of the program were being developed simultaneously.

The second validity framework consisting of the five sources of validity evidence represents a comprehensive agenda that entails a host of longer-range validation studies. At this juncture, a few potentially important types of validity activities are anticipated. An important area of research is the relationship of Smarter Balanced with other important national and international large-scale assessment programs, such as NAEP, TIMSS, and PISA. This is important to establish the technical properties and rigor of the Smarter Balanced assessments. Most important is the validation of the measurement of college and career readiness, which entails collecting various types of criteria from sources outside the Smarter Balanced assessment system. In considering potential validity studies that will be important in the future, establishing research support systems to enable these activities for outside investigators will have lasting benefits and ones that will augment the validity of Smarter Balanced.

References

- Abedi, J. & Ewers, N. (2013). *Accommodations for English Language Learners and Students with Disabilities: A Research-Based Decision Algorithm*. Smarter Balanced Assessment Consortium.
- Achieve, Inc. (2004). Ready or not: creating a high school diploma that counts. American Diploma Project. Washington, DC: Achieve, Inc. Retrieved February 11, 2008, from www.achieve.org/files/ADPreport_7.pdf.
- ACT (2006). *Ready for college, ready for work. Same or different?* Iowa City, IA: Author.
- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40, 955-959.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Institutes for Research (2014a). *Technical Specifications Manual for Online Testing For the Spring 2014 Field Test Administration*. Smarter Balanced Assessment Consortium.
- American Institutes for Research (2014b). Smarter Balanced Scoring Specification: 2014–2015 Administration.
- Betebenner, D. W. (2011). *New directions in student growth: The Colorado growth model*. Paper Presented at the National Conference on Student Assessment, Orlando, FL, June 19, 2011. Retrieved March 29, 2012, from <http://ccsso.confex.com/ccsso/2011/webprogram/Session2199.html>.
- Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22, 21-29.
- Camara, W. J. (2013). Defining and measuring college and career readiness: A validation framework. *Educational Measurement: Issues and Practice*, 32, 16–27.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Castellano, K. E., & Ho, A. D. (2013). *A Practitioner's Guide to Growth Models*. Council of Chief State School Officers.
- Cohen, J. & Albright, L. (2014). *Smarter Balanced Adaptive Item Selection Algorithm Design Report*. American Institutes for Research.
- Conley, D. T., Drummond, K. V., de Gonzalez, A., Rooseboom, J., & Stout, O. (2011). *Reaching the goal: The applicability and importance of the Common Core State Standards to college and career readiness*. Eugene, OR: Educational Policy Improvement Center.
- Crocker, L. M., Miller, D., and Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, 2, 179-194.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

- Dorey, N (2014). *Smarter Balanced "Tests of the Test" Successful: Field Test Provides Clear Path Forward*
<http://csai-online.org/resource/698>
- Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span.
Psychological Bulletin, 93, 179-197.
- ETS (2015). *Study of the Relationship Between the Early Assessment Program and the Smarter
Balanced Field Tests*. Prepared for the California Department of Education by Educational
Testing Service.
- ETS (2014). *Online Field Test Administration Manual for Spring 2014 Field Tests of English
Language Arts/Literacy and Mathematics*. Smarter Balanced Assessment Consortium.
- ETS (2002). *ETS Standards for Quality and Fairness*. Educational Testing Service.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence.
Educational Measurement: Issues and Practice, 18, 5-9.
- Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Berk (ed.),
Criterion-referenced measurement: the state of the art. Baltimore: Johns Hopkins University
Press.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of
Educational Research*, 64, 425-461.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting
standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and
perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th ed., pp. 17-64).
Washington, DC: American Council on Education/Praeger.
- Linn, R. L. (2006). The Standards for Educational and Psychological Testing: Guidance in test
development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp.
27-38), Mahwah, NJ: Lawrence Erlbaum.
- Loomis, S. C. (2011, April). *Toward a validity framework for reporting preparedness of 12th graders
for college-level course placement and entry to job training programs*. Paper presented at
the annual meeting of the National Council on Measurement in Education, New Orleans.
- Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and
instruction, *Review of Educational Research* 4, 1332-1361.
- McGraw-Hill Education/CTB (2014). *Smarter Balanced Pilot Automated Scoring Research Studies*.
Smarter Balanced Assessment Consortium.
- McGraw-Hill Education/CTB (2014). *Field Test: Automated Scoring Research Studies*. Smarter
Balanced Assessment Consortium.
- Measured Progress/ETS (2012). *Smarter Balanced Assessment Consortium: Annotated
Bibliography: Item and Task Specifications and Guidelines*. Smarter Balanced Assessment
Consortium.
- Measured Progress/ETS (2012). *General Item Specifications*. Smarter Balanced Assessment
Consortium.
- Measured Progress/ETS (2012). *Signing Guidelines*. Smarter Balanced Assessment Consortium.

- Measured Progress/ETS (2012). *Tactile Accessibility Guidelines*. Smarter Balanced Assessment Consortium.
- Measured Progress/ETS (2012). *Performance Tasks Specifications*. Smarter Balanced Assessment Consortium.
- Measurement Incorporated/CTB/McGraw-Hill (2014). *Achievement Level Setting Plan*. Smarter Balanced Assessment Consortium Internal Document.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed.). Washington, DC: American Council on Education.
- Mislevy, R. J. (2009, February). Validity from the perspective of model-based reasoning. *CRESST report 752*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61-90), Mahwah, NJ: Lawrence Erlbaum.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.
- National Center on Educational Outcomes (2015). *Usability, Accessibility, and Accommodations Guidelines*. Smarter Balanced Assessment Consortium.
- National Council on Measurement in Education (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Madison, WI: Author.
- Navigation North Learning (2014). *The Smarter Balanced Technology Strategy Framework and Testing Device Requirements*. Smarter Balanced Assessment Consortium.
- Penfield, R. D., & Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement In Education*, 17, 359-370.
- Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education*, 5, 285-301.
- Porter, A. C., & Smithson, J. L. (2002, April). *Alignment of assessments, standards and instruction using curriculum indicator data*. Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, LA.
- Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). Benchmarking and alignment of standards and testing (Technical Report 566). Washington, DC: Center for the Study of Evaluation.
- Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests*. National Research Council.
- Sireci, S. G. (2012). *Smarter Balanced Assessment Consortium: Comprehensive Research Agenda*. Report Prepared for the Smarter Balanced Assessment Consortium.
- Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299-321.
- Smarter Balanced (2014). *Mathematics Summative Assessment Blueprint*. Smarter Balanced Assessment Consortium.

Smarter Balanced (2014). *ELA/Literacy Summative Assessment Blueprint*. Smarter Balanced Assessment Consortium.

Smarter Balanced (2014). *Statements of Support: Achievement Level Setting for the Smarter Balanced Assessments*. Smarter Balanced Assessment Consortium.

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Scoring Guide for Selected Short-Text Mathematics Items (Field Test 2014)*.

Smarter Balanced Assessment Consortium (2014). *Mathematics Performance Task Specifications (Design, Development, and Scoring Plan)*.

Smarter Balanced Assessment Consortium (2014). *Calculator Availability Information for 2014 Field Test*.

Smarter Balanced (2013). *Career Readiness Frameworks Introduction and Implementation Guide*.

Smarter Balanced (2013). *Initial Achievement Level Descriptors and College Content-Readiness Policy*. Smarter Balanced Assessment Consortium.

Thissen, D. & Wainer, H. (eds.) (2001). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Vasavada, N., Carman, E., Hart, B., & Luisser, D. (2010). Common core state standards alignment: Readiness, PSAT/NMSQT, and SAT. *Research report 2010-5*. New York: The College Board.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*, 7-25.

Chapter 3: Item Development

In order to build a summative assessment that measured the intended claims, the Consortium's test development cycle was iterative, involving experts from various education-related fields, and was based on assessment-related research and best practices.

Item and Task Specifications¹

The item specifications bridge the span from the content specifications and ALDs to the assessment itself. While the content specifications established the Consortium's claims and the types of evidence or targets, that would need to be collected in order to support these claims, more specificity was needed in order to develop items and tasks that measured the claims. Working with three vendors (ETS, Measured Progress, and CTB) with extensive experience in item writing, the Consortium's Item Development Work Group and Performance Task Work Group created item and task specifications for both ELA/Literacy and mathematics.

The first iteration of the item and task specifications were developed in 2011. In early 2012, the Consortium held a series of showcases where the contractors introduced the item and task specifications and collected feedback from member states. Using this feedback, the item and task specifications were revised during the first quarter of 2012.

Using the revised item and task specifications, a small set of items was developed and administered in fall 2012 during a small-scale trial. This provided the Consortium with their first opportunity to administer and score the new item types. During the small-scale trials, the Consortium also conducted cognitive laboratories to better understand how students respond to various types of items. The cognitive laboratories used a think-aloud methodology in which students speak their thoughts while working on a test item. The item and task specifications were again revised based on the findings of the cognitive laboratories and the small-scale trial. These revised specifications were used to develop items for the 2013 pilot test, and they were again revised based on 2013 pilot test results and subsequent review by content experts.

The Consortium's item and task specifications are designed to ensure that the assessment items measure the assessment's claims. Indeed, the purpose of the item and task specifications is to define the characteristics of the items and tasks that will provide the evidence to support one or more claims. To do this, the item and task specifications delineate the types of evidence, or targets, that should be elicited for each claim within a grade level. Then, they provide explicit guidance on how to write items in order to elicit the desired evidence.

In doing this, the item and task specifications provide guidance on how to measure the targets (standards) first found in the content specifications. The item and task specifications provide guidelines on how to create the items that are specific to each assessment target and claim through

¹ <http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/ItemSpecifications/GeneralItemSpecifications.pdf>

the use of task models. In mathematics a task model provides a description of an item/task's key features. These task models describe the knowledge, skills, and processes being measured by each of the item types aligned to particular targets. In addition, the task models sometimes provide examples of plausible distractors. Exemplar items are provided within every task model. In ELA these functions are carried out through item specifications.

These task models were developed for each grade level and target in order to delineate the expectations of knowledge and skill to be included on test questions in each grade. In addition, both the ELA/L and mathematics item and stimulus specifications provide guidance on determining the grade appropriateness of task and stimulus materials (the materials that a student must refer to in working on a test question). The task and stimulus models also provide information on characteristics of stimuli or activities to avoid because they are not important to the knowledge, skill, or process being measured.

This is important because it underscores the Consortium's efforts to develop items that are accessible to the widest range of students possible; in other words, Consortium items are created according to the principle of **universal design**. As the name suggests, the concept of universal design aims to create items that accurately measure the assessment target for all students. At the same time, universal design recognizes that one solution rarely works for all students. Instead, this framework acknowledges "the need for alternatives to suit many different people."²

To facilitate the application of universal design principles, item writers are trained to consider the full range of students who may answer a test question. A simple example of this is the use of vocabulary that is expected to be known by all third-grade students versus only those third-grade students who play basketball. Almost all third-grade students are familiar with activities (e.g., recess) that happen during their school day, while only a subset of these students will be familiar with basketball terms like "double dribble," "layup," "zone defense," or "full-court press."

Classroom Activities

In order to mitigate possible unfamiliarity with context or vocabulary performance tasks are preceded by an unscored activity conducted with the whole classroom. The activity is intended to resolve issues of unfamiliarity so that performance on the task reflects only subject matter knowledge. There are several performance tasks associated with each classroom activity so that the class receives a variety of tasks.

In addition to this, the item specifications discuss accessibility issues that are unique to the creation of items for a particular claim and/or assessment target. The accessibility concerns discuss the different supports that various groups of students may need to access the content of an item. By considering the possible supports that may be needed for each item, item writers are able to create items that will support those adaptations.

² Rose, D. & Meyer, A. (2000). Universal design for learning, associate editor column. *Journal of Special Education Technology* 15(1):66-67

The use of universal design principles allows the Consortium to collect evidence on the widest possible range of students. By writing items that adhere to the item and task specifications, the Consortium is assured that the assessments measure the claims and assessment targets established in the content specifications as well as the knowledge, skills, and processes found in the Common Core State Standards for *all* students for whom the assessment is appropriate.

The Item/task Pool

An **item pool** refers to a collection of test questions (known as items) measuring the same content area (e.g., mathematics) within the same grade. (As explained in the Test Design chapter, the use of off-grade-level items is allowed in some instances.) The quality of the items is a primary concern when building an item pool. The Consortium took multiple steps to ensure the quality of the items in our item pool. Building on the ongoing process of developing item/task specifications and test blueprints described in the previous chapter, the Consortium used an iterative process for creating and revising each item as well as the collection of items. The Consortium tested items and refined its approach to item development through three steps: small-scale tryouts, a large pilot test, and the largest ever field test of a K-12 assessment. Details of the pilot and field tests are found in CH _____. During each phase, the Consortium used cognitive laboratories to understand the strategies that students used to respond to the items. By incorporating this tiered and iterative approach, the item and task specifications that guided the development of the final operational pool were improved based on the lessons learned during these important tryouts.

Using the summative and comprehensive interim test blueprints, the number and distribution of items to be written were specified for item writing teams. Pools of items/tasks were written specifically to support the operational blueprint. Teachers were integrally involved in the creation of the item/task pool from beginning to end. Some participated in the processes described in the flow charts below. Others developed many of our items through a rigorous item writing process, and yet others reviewed the items for accuracy and appropriateness of the content knowledge and skill level required to respond to the items, potential issues of bias in favor of or against any demographic group of students, and accessibility for students with disabilities and English language learners. Teams of educators reviewed items for content, bias, and accessibility prior to administration to any students. Following the pilot and field test administrations, items were again reviewed if pilot or field test data indicated a potential problem. Finally, teachers participated in in range finding and scoring constructed-response items/tasks to ensure that the items/tasks could be properly scored given their scoring rubrics.

In this section, we will examine the primary role that educators played in creating the field-test item pool by writing, reviewing, and scoring items. This section will end by examining the current composition of the item pool.

Item Writing

The job of writing all of the items and performance tasks was no small undertaking, and the Consortium worked with educators throughout the test development cycle to develop items. Prior to the spring 2013 pilot test, the Consortium engaged 136 educators in K-12 and higher education

from 19 member states to write items. Prior to the spring 2014 field test, 184 educators in K-12 and higher education from 16 member states participated in item writing. All K-12 participants:

- Were certified/licensed to teach ELA/L and/or mathematics in a K-12 public school;
- Were currently teaching in a public school within a Smarter Balanced Governing State;
- Had taught ELA and/or mathematics in grades 3 through 8 and/or high school within the past three years (second-grade teachers were also recruited to participate in the development of grade 3 items and/or tasks);
- Had previously reviewed part or all of the CCSS for the content area for which they were writing items and/or performance tasks;
- Submitted a statement of interest that described their interest in developing Smarter Balanced items and/or performance tasks as well as their qualifications for doing so;
- Completed training and achieved qualifications through the certification process.

Qualifications for Higher Education Faculty included:

- Current employment with, or recent retirement from, a college or university located within a Smarter Balanced Governing State;
- Having taught developmental and/or entry-level courses in English, composition, mathematics, statistics or a related discipline within the last 3 years;
- Having previously reviewed part or all of the CCSS for the content area in which they are interested in writing items and/or performance tasks;
- Completing training and achieving qualifications through the certification process.

The selected educators were trained on the Consortium's content specifications, the item and task specifications, and stimulus specifications (ELA/L) as well as the item authoring system in which the items were developed. In addition, professional item writers and the Consortium held regular meetings to provide direction and feedback to the educators. Educators, state partners, and assessment vendors developed the items in the Consortium's item pool.

Training

Educators participated in a series of facilitated, online webinars in order to qualify as item writers. To facilitate participation, the Consortium scheduled multiple sessions in different time zones, including evening sessions. In addition to the facilitated sessions, the Consortium provided training modules that covered background on the Consortium, assessment design principles, and detailed information about item and performance task development. All modules were available in three formats: a PowerPoint presentation with notes, a streaming presentation with narration that could be viewed online, and a downloadable audio/video presentation.

The item writers were specifically trained on the Consortium’s content and item specifications, stimulus specifications,³ sensitivity and bias guidelines,⁴ and general accessibility guidelines.⁵ Training on these specifications and guidelines helped ensure that item writers were trained to write items that allowed the widest possible range of students to demonstrate their knowledge, skills, and cognitive processes in regard to the content. This meant that item writers needed to understand the content for which they were writing items as well as accessibility and sensitivity issues that might hinder students’ ability to answer an item. Item writers were also trained to be aware of issues that might unintentionally bias an item for or against a particular group.

Educator Participation

Consistent with the Consortium process, educators were the primary developers of items. The active involvement of educators was critical to the success of the item writing activities. Educators engage with students on a daily basis, and they understand the ways in which students can demonstrate their knowledge. Their involvement in item writing helped ensure that the assessment system is accurate and efficient, and provides valid evidence of student learning.

State-Managed Item Development

The Consortium invited member states to participate in a separate effort to write items. This voluntary effort, known as State-Managed Item Development, was conducted to build the capacity of states to write items and to support the overall sustainability of the Consortium. To this end, six states (HI, ID, MI, WA, WV, and WY) participated in the state-managed field test item development opportunity. During this opportunity, educators within the six states developed approximately 3,100 items in mathematics and ELA/L across grades 3 through 8 and high school. Many of these items were field tested during the operational test in spring 2015.

Item Reviews

Once items were written, groups of educators reviewed items prior to their pilot test administration in spring 2013 and their field test administration in spring 2014. Items that survived the pilot test were again reviewed prior to their use in the spring 2014 field test.

Accessibility, Bias/Sensitivity, and Content Reviews

³ <http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/EnglishLanguageArtsLiteracy/ELASTimulusSpecifications.pdf>

⁴ <http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/Guidelines/BiasandSensitivity/BiasandSensitivityGuidelines.pdf>

⁵ <http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/Guidelines/AccessibilityandAccommodations/GeneralAccessibilityGuidelines.pdf>

Panels of educators reviewed all items, performance tasks, and item stimuli for accessibility, bias/sensitivity, and content. (Item stimuli refer to the reading passages used on the ELA/L assessments or the figures and graphics used on the mathematics assessments.) Prior to the spring 2013 field test, 122 ELA/L educators and 106 mathematics educators reviewed items and performance tasks for accessibility, bias/sensitivity, or content, and 60 educators reviewed the ELA/L stimuli. Prior to the spring 2014 field test, 107 ELA/L educators and 157 mathematics educators from 14 states reviewed items and performance, and 95 educators from 13 states reviewed the ELA/L stimuli.

The educator qualifications for the accessibility, bias/sensitivity, and content reviews were the same as the educator qualifications for item writing except that participants were not required to submit a statement of interest. In addition, it was preferred (but not required) that educators have previous experience reviewing items, tasks, and/or stimuli.

During the accessibility reviews, panelists identified issues that may negatively affect a student's ability to access stimuli, items, or performance tasks, or to elicit valid evidence about an assessment target. During the bias and sensitivity review, panelists identified content in stimuli, items, or performance tasks that may negatively affect a student's ability to produce a correct response because of their background. The content review focused on developmental appropriateness and alignment of stimuli, items, and tasks to the content specifications and appropriate depths of knowledge. Panelists in the content review also checked the accuracy of the content, answer keys, and scoring materials. Items flagged for accessibility, bias/sensitivity, and/or content concerns were either revised to address the issues identified by the panelists or removed from the item pool.

Data-Related Reviews

The items developed for the item pool were administered during the spring 2013 and spring 2014 pilot tests, and the pilot test data from both administrations were analyzed to examine the statistical quality of the items in the pool. The Consortium established statistical criteria to flag items for possible defects in quality related to content, bias, or accessibility. For example, content-related criteria flagged items for further review if they were extremely difficult or extremely easy. Accessibility-related criteria flagged items that were differentially more difficult for students with disabilities compared to students without disabilities.

Following the spring 2013 pilot, 40 educators participated in the item data review and examined the items for possible content-related issues or accessibility-related issues. Following the spring 2014 pilot, 57 ELA/L educators from 16 states and 30 mathematics educators from 12 states participated in item data review, examining the items for possible content-related issues or accessibility-related issues. At least two educators reviewed each item. These educators were trained via webinars on the flagging criteria and on how to evaluate flagged items. These educators made recommendations on whether to accept the item with no change, revise and re-field test the item, or reject the item from the pool. McGraw-Hill CTB content experts reviewed all items where the reviewers' recommendations disagreed. In addition, McGraw-Hill CTB content experts and psychometricians also reviewed and provided recommendations for all items where both reviewers recommended accepting the item. In each situation, the content expert provided the Consortium with a final recommendation for the item.

The educator qualifications for the item data reviews were the same as the educator qualifications for item writing except that participants were not required to submit a statement of interest.

Item Scoring

For those items that could not be machine scored, the Consortium engaged 102 participants from 20 states in range finding activities for those items requiring human scoring following the spring 2013 pilot. After the spring 2014 pilot, 104 educators participated in range finding. Range finding improves the consistency and validity of scoring for the assessment. During range finding, the educators focused on the performance tasks for mathematics and ELA/L. In mathematics, educators also reviewed constructed response items for grades 7, 8, and high school. During range finding, the participants reviewed student responses against item rubrics, validated the rubrics’ effectiveness, and selected the anchor papers that would be used by professional scorers during the main scoring event.

The educator qualifications for range finding were the same as the educator qualifications for item writing, except that participants were not required to submit a statement of interest. In addition, it was preferred (but not required) that educators had previous range finding experience.

Composition of the Item Pool⁶

The Consortium developed many different types of items beyond the traditional multiple-choice item. This was done to measure the claims and assessment targets with varying degrees of complexity by allowing students to construct their responses rather than simply recognizing a correct response. These different item types are listed in Table 1 below.

Table 1 Item Types found in the Consortium’s Item Pool.

Item Types	ELA/L	Mathematics
Multiple Choice, Single Correct Response	X	X
Multiple Choice, Multiple Correct Response	X	X
Two-part Multiple Choice, with Evidence Responses (EBSR)	X	
Matching Tables	X	X
Hot Text	X	X
Drag and Drop	X	X
Short Text Response	X	
Essay	X	

⁶ Examples of many of the item types may be found at: <http://www.smarterbalanced.org/sample-items-and-performance-tasks/>.

Hot Spot		X
Short Text and Fill-in Tables		X

Each grade’s item pool for the Consortium’s test was necessarily large to support the summative and interim assessments⁷ being delivered via a computer using adaptive test-delivery technology, commonly called a computer adaptive test or CAT. Unlike a traditional paper-and-pencil test where all students take the same items, students taking the Consortium’s CAT will take items and tasks targeted to their ability level. This means that the Consortium needed to develop a very large number of items in order to meet the needs of the student population.

In addition to the items for the CAT, the Consortium also developed performance tasks. All students take performance tasks that are designed to measure a student’s ability to integrate knowledge and skills across multiple assessment targets. These performance tasks may also be delivered via the same online assessment delivery system as the CAT.

Table 2 below shows the total number of CAT items and performance tasks (PT) found in each item pool by grade level and content area. As the table shows, over 1,600 ELA/L CAT items were developed in each of grades 3 – 8, and 5,711 items were developed for high school. In mathematics, approximately 1,500 items were developed in each of grades 3 – 8, and 4,512 items were developed for high school. The items in these pools will support both the summative and interim assessments.

There were approximately 50 PTs per grade developed in each of grades 3 – 8 in both ELA/L and mathematics. In high school, the Consortium created 124 ELA/L PTs and 132 mathematics PTs. Each PT has multiple associated items: four and six items per PT in ELA/L and mathematics, respectively.

⁷ Interim assessments will not be delivered via CAT until the 2015-16 school year. They are fixed-form tests in the 2014-15 school year.

Table 2 Total Number of CAT Items and Performance Tasks (PT) developed by Grade and Content Area

Row Labels	ELA/L		Math	
	CAT	PT	CAT	PT
3	1711	49	1502	50
4	1653	49	1551	49
5	1659	47	1517	49
6	1683	47	1503	49
7	1657	47	1487	49
8	1626	49	1488	49
HS	5711	124	4512	132
Grand Total	15700	412	13560	427

The numbers of items that survived to be field tested is listed in Chapter 7.

Selection of Items for the Operational Item Pool

The statistical quality of the items was again evaluated following the 2014 field test. Items that did not perform well according to established psychometric criteria (for example, item statistics such as difficulty and discrimination) were forwarded to content experts for review. The same psychometric criteria were used to judge items regardless of whether the items were used on the interim assessment or the summative assessment.

For the first operational year (2014-2015), items for both the interim assessment and the summative assessment were drawn from the same item pool. The summative item pool is secure, while the interim pool can be accessed by teachers to aid in planning and interpretation. The long-term plan is that most items will first be administered on the summative assessment before entering the interim item pool. In the first operational year, interim pools supported fixed form tests. Many of the items being field tested in 2015 will move directly from the field test to the interim item pool as necessary to meet the content requirements and support interim adaptive testing where possible.

Table 3. Mathematics Specifications and Archetype Delivery

MATH						
Delivery Number	Number of Specs - Math	Grade Batches for Each Item Spec Delivery		Archetypes		
		Grade	Claim	Number	Grade	Claim
1	6	3, 7, HS		16	3,7,HS	1, 2, 3
2	15	3, 8, HS		18	3,8,HS	1, 2, 3, 4
3	14	7,8,HS		18	7,8,HS	1, 2
4	14	4,5,6		18	4,5,6	1, 2, 3, 4
5	13	4,8,HS		20	4,8,HS	1, 2, 3, 4
6	14	5,7,HS		18	5,7,HS	1, 2, 3, 4
7	12	5,6,HS		16	5,6,HS	1, 2, 3, 4
8	12	3,7,HS		18	3,7,HS	1, 2, 3, 4
9	13	5,6,HS		18	5,6,HS	1, 2, 3, 4
10	14	4,8		18	4,8	1, 2, 3, 4
11	12	3,HS		18	3,HS	1, 2, 3, 4
12	12	3,5,6		18	3,5,6	1, 2, 3, 4
13	10	4,7		19	4,7	1, 2, 3, 4
14	8	3,6		17	3,6	1, 2, 3, 4
	169			250		

Note: Archetypes were assembled to be representative of the entire set of items.

Note: The archetype numbers include any Performance Tasks that are developed as part of the archetype pool.

Table 4. English/Language Arts/Literacy Specifications and Archetype Delivery

ELA							
Delivery Number	Number of Specs - ELA	Grade Batches for Each Item Spec Delivery			Archetypes		
		Number	Grade	Claim	Target	Number	Claim
1	1	Stimulus	1-4				
1	7	3-HS	2	8			
2	7	3-HS	2	9			
3	28	3-HS	2-4	1, 2, 3			
4	28	3-HS	2, 4, PT	3, 4, 6, I/E PT			
5	55	3-HS	1, PT	1-7, N PT			
6	56	3-HS	1, PT	8-14, O/A PT			
7		3-HS			20	2	8, 9
8		3-HS			110	2-4	1-4, 6
9		3-HS			58	1	1-7
10		3-HS			62	1	8-14
Total	182	3-HS	Total		250		

Note: The archetype numbers include any Performance Tasks that were developed as part of the archetype pool.

Use of Systems and Tagging

The CTB Collaborative used DAS and ITS as the item authoring platforms to support production of various item types for the Smarter Balanced Assessment Consortium. Both systems were used for the authoring and review of stimuli and items/tasks for Smarter Balanced Contract 14. Clear procedures for the flow of items through authoring and review steps were developed and communicated to the various review groups and other stakeholders.

To support the implementation of the Smarter Balanced Field Test, the Collaborative ensured the availability of a robust set of item, task, and stimulus metadata to meet several purposes:

- Item tagging to support the computer-adaptive algorithm for test administration
- Content tagging to document full coverage of Smarter Balanced assessment claims and targets and the Common Core State Standards
- Item-bank tagging to support continued use of Smarter Balanced items by states

- Other item tagging for reporting and analysis

In addition to item-attribute tagging, items were associated with annotations to support Smarter Balanced accommodations. The expectation was that all metadata tags will be applied to items in the CTB DAS and the AIR ITS systems. This included annotations for

- Translations (including ASL)
- Braille
- Text-to-speech
- Glossaries (English and second-language)
- Other required accommodations tagging

The Collaborative was prepared to add/edit the list of item attribute tags that were under consideration for Smarter Balanced 16 item development and added tags as needed to meet the needs of the Field Test. This list, once approved, included additional tags to capture Smarter Balanced 16 requirements, such as Task Model, Specification Version Date, and Component Items for performance tasks. Other features required tagging (e.g. use of language complexity rubrics) were accommodated by additional tags.

Targeted training was provided to ensure that all item authors, developers, and reviewers understand the purpose and requirements for item attribute tagging.

Training Activities

Because of the large number of item and performance task writers that were ultimately involved in the development of the Field Test item pool, training involved live virtual sessions and the use of prerecorded modules that could be reviewed and accessed on demand.

In July 2013, the following activities were completed in preparation for training and professional development for educator item writers and editors:

1. Updates and revisions to training modules. While the modules developed for Contract 08 were effective in providing an introduction to the Smarter Balanced assessment system and components of item development, Contract 14 provided additional insight relative to aspects of the training modules that were most successful and what had to be updated and/or revised due to changes made throughout pilot test item and task development. For example, modules that focused on content and item specifications reflected general changes to those specifications as well as included sample items that reflect the current design and approach for Smarter Balanced items and tasks. Similarly, those modules that focused on item types were revised to address response types as indicated through item development conversations.

2. Development of new modules. The contractor collaborative developed new modules that focused on 1) the item authoring system(s), 2) in-depth training for writing items to each claim/target (including the use of item specifications, task models, and CCSS for each content area) and 3) expansion of accessibility considerations such as linguistic complexity. Those modules followed the design of those developed for Pilot Test item development.

3. Selection of educators for item writing. Recruitment began in early May 2013 to ensure parallel development across educator-, state-, and vendor-created items. Item writers were specifically recruited across

a range of content areas, grade levels including higher education, and experience with under-represented student populations.

Virtual Training of Educator Item and Task Writers

Educator training sessions occurred after the school year had ended (mid-July), when participants had more time available. CTB experienced many scheduling conflicts and challenges during Contract 14 that they planned to circumvent via early, regular, and clear communication. For the virtual training workshops, video and/or audio presentations were consistent so that all educators hear the same messages from the Smarter Balanced work groups.

Continued Training during Development

Training and learning throughout item and task development was ongoing. Each educator worked directly with one of the CTB Collaborative assessment specialists. These specialists conducted regularly scheduled meetings in which educators could share and discuss challenges and successes. This approach would be similar to the model that was used during Contract 14. In this model, educators met and worked directly with one of the contractors' assessment specialists to receive appropriate support and guidance. For Contract 16/17, CTB was able to enhance that model through the added value of experienced educator item and task writers, who could also be called upon to provide feedback. Because item/task development was a phased process, retraining was conducted for the appropriate writers immediately prior to the onset of each phase. This was "just-in-time" training that focused on the content and item specifications. If issues arose during item development, The Collaborative conducted ad hoc training for individuals, small groups, or large groups, as needed.

Certification and Management

At the onset of item development, all item development entities (subcontractors, external item development vendors, states, and individual educators) were required to develop the appropriate samples needed for item development certification. CTB provided preliminary training for individual educators recruited as item writers prior to assigning the 20-item certification set. Once potential item developers met the preliminary qualifications, each organization or individual wishing to continue in the certification process created a sample set of 20 items reflective of the anticipated item assignments. This requirement applied to item and task development organizations within our collaborative (CTB, AIR, DRC, MI, and SCALE), any external item and task vendors used for item authoring, groups of educator authors within a state, and individual educator item authors who were recruited directly.

The table below outlines a sample certification set for item authors or organizations seeking approval to move forward with item development. In this example, the organization would create six 20-item sets for approval prior to the start of item development for Claims 1 and 4.

Table 5. Sample Certification Set

Claim	Grade Band		
	3-5	6-8	HS
1	X	X	X
2			
3			
4	X	X	X

Item/Task Set Evaluation

Once the sample item or task set was created, senior content reviewers evaluated the item set based on the approved criteria for certification. The contractor provided initial feedback or additional training to item developers, as needed. Once an item set met all criteria, the item developer was recommended for certification.

All item development entities (subcontractors, external item development vendors, states, and individual educators) were required to develop the appropriate samples needed for item development certification. This 20-item certification set was reflective of the anticipated item assignments. This set of items was reviewed by the collaborative senior reviewers, as well as by Smarter Balanced representatives (SMEs, Accessibility, Sensitivity and Bias experts) for adherence to the item quality review criteria for Smarter item development. For an entity to qualify for Smarter item development, 95% of the items in the certification set and in all sets of items submitted for the Smarter 16 Field Test pool had to meet an acceptance of 95%.

Role of Item Quality Review Panel

The Item Quality Review Panel (IQRP) was recruited in early May 2013 and consisted of seven panelists for ELA and nine members for math. Panel members are content experts or those who inform decisions related to students with disabilities and English Language Learners. The panel gave feedback during reviews of specifications, archetypes, and item and performance task batches.

The IQRP held an initial Face-to-face meeting in late May 2013. The outcomes of this meeting included recommendations to consider for item development and contributions to the item quality-review criteria. The recommendations from this meeting were refined and implemented based on Smarter Balanced confirmation and agreement to the overall vision of the assessment.

Educator Recruitment Activities

Smarter Balanced recruited educators for each of the assessment activities listed in Table 6 below. The inclusion of qualified educators in these assessment activities builds capacity and creates sustainability for the Smarter Balanced assessment system. The contractor, with assistance from the governing states' Teacher Involvement Coordinators (TICs), will sought an educator sample that ensured a balanced representation based on grade levels, content area, and other demographic data across the governing states. For each recruitment activity, TICs used an information packet describing the purpose of the activity and recruitment process. The information packets included the following:

- Documentation for TICs about the recruitment activity, specifically state recruitment and qualifications for educators
- A summary of overall recruitment processes in the coming months
- Specific counts of educators CTB is recruiting from each state for the educator opportunity
- A list of Frequently Asked Questions that may be used by TICs to provide information about educator opportunities
- A sample educator recruitment email that TICs may use in support of state recruitment. State-specific information can be added or clarified in this template.

In addition to requests for specific numbers of educators across content areas and grades, TICs received specific targets for the recruitment of educators with experience in working with under-represented student populations (English language learners and students with disabilities).

The recruitment of educators for each of the activities occurs in phases. The recruiting activities and timelines are provided in Table 6.

Table 6. Recruiting Activities and Timeline

Activity	Number of Educators Required	Information Package(s) to TIC	Timeframe for Recruiting	Notification to TICs of Participants and Alternates	Timeframe for Activity
Phase I					
Pilot Test Range Finding	105	5/13/13	5/14 to 5/28/13	5/30/13	June
Field Test Item and Task Authoring	200	5/15/13	5/16 to 5/30/13	6/4/13	June–October
Phase II					
Pilot Test Item-Data Review	28	5/22/13	5/23 to 6/6/13	6/11/13	August–September
Field Test Stimulus Review	284				June–August
Field Test Item-and-Task Review	296				July–November
Phase III					
Alignment Study	TBD	11/4/13	11/5 to 11/18/13	12/11/13	January–March
Phase IV					
Field Test Range Finding	TBD	2014	2014	2014	TBD

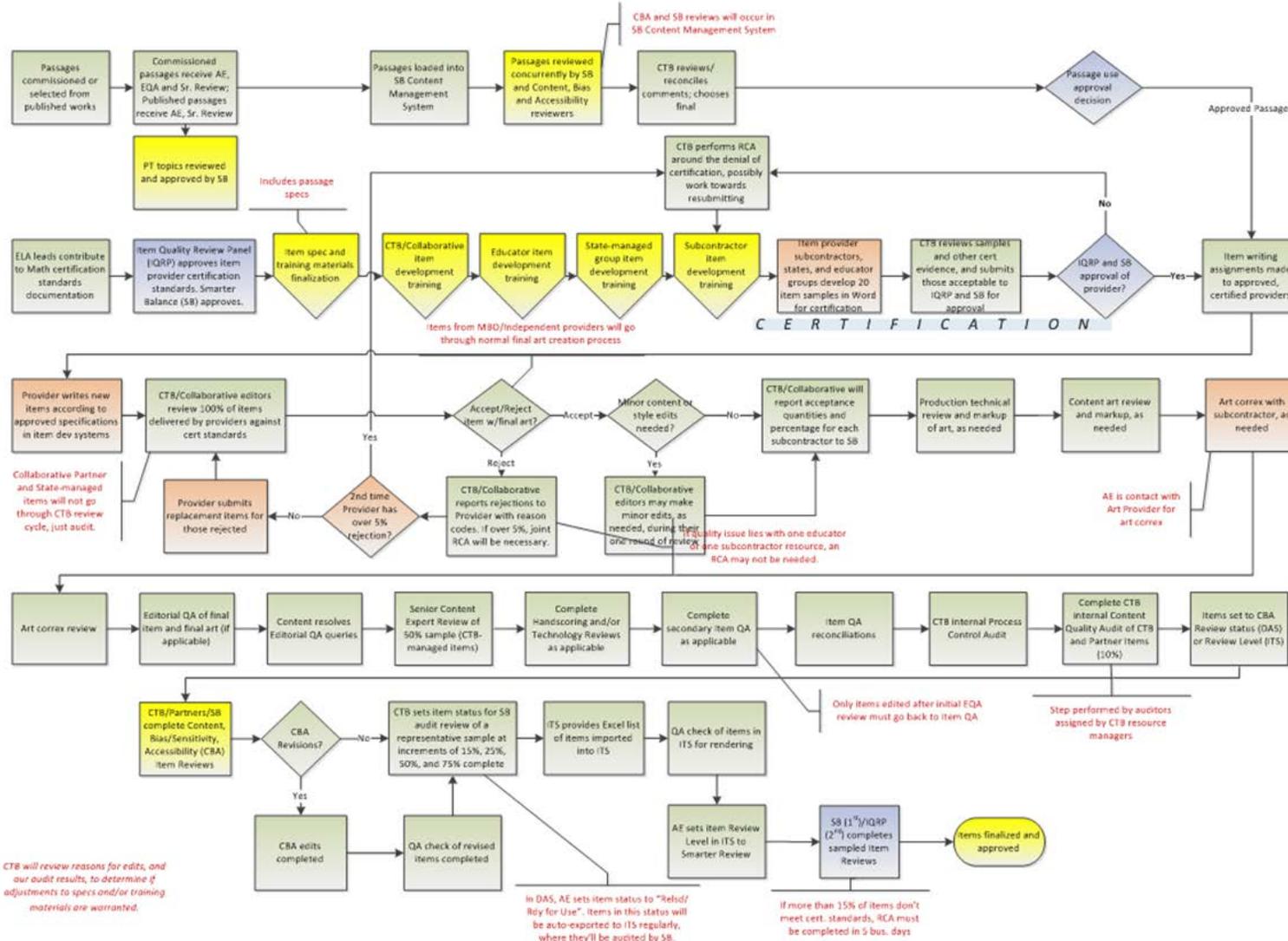
Item development process

The charts below outline the detailed process for stages of item development. They describe the many checks and reviews each item receives before it is approved for field testing. Item content, graphics, artwork, response processes and stimuli get extensive reviews. Items are also subject to reviews for possible cultural bias or material that may distract some test takers because it is in an area of sensitivity. Throughout the process there are checks to assure that items are accessible to as many students as possible.

ELA ITEM DEVELOPMENT PROCESS – SMARTER BALANCED 16

Content Provider	CTB/ Collaborative	Smarter Balanced	Joint Tasks
------------------	--------------------	------------------	-------------

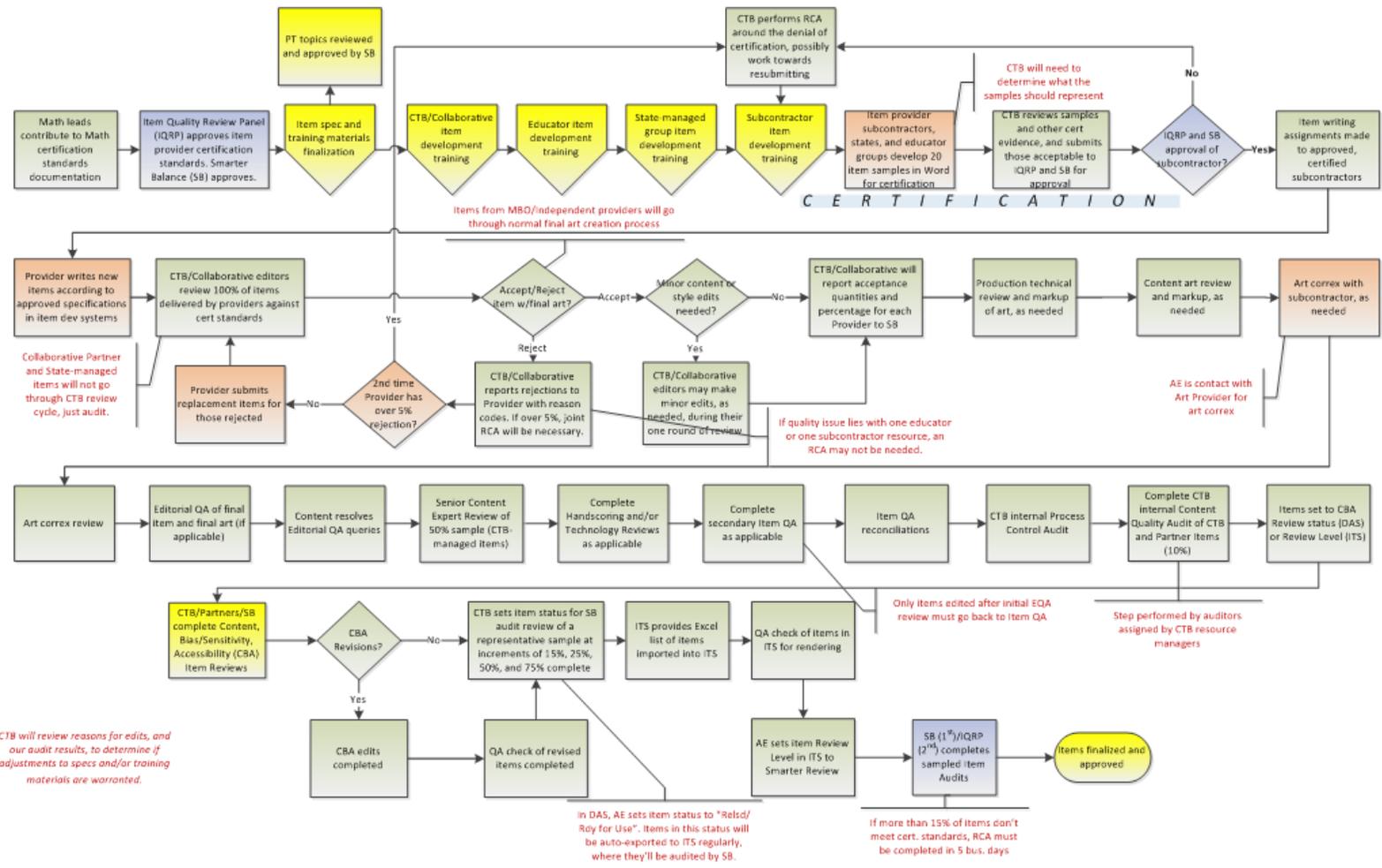
Assumption: All E/IA item development completed in DAS or ITS development system (SB Content Management System)



Assumption: All MATH item development completed in DAS or ITS item development system (SB Content Management System)

MATH ITEM DEVELOPMENT PROCESS – SMARTER BALANCED 16

Content Provider	CTB/ Collaborative	Smarter Balanced	Joint Tasks
------------------	--------------------	------------------	-------------



CTB will review reasons for edits, and our audit results, to determine if adjustments to specs and/or training materials are warranted.

Detailed information about item writing, development, review and scoring can be obtained upon request.

Table 7. Additional item writing, development, review and scoring documentation

Topic	Sub-topic	Document Name
Item Writing	Process Flow	20150512 Item Development Process Description FINAL
		20150512 Smarter process maps FINAL
		Smarter 16 ITS Final Content Approval checklist FINAL
		Smarter 16 Final Web Approval Checklist20150512
	Models-Specifications	20131003 Smarter 16 Item pool specification v12a Math FINALnew
		20131006 Smarter 16 Item pool specification v12d ELA FINALnew
		ELA Archetypes
		Math_Archetype_Metadata
	Review criteria	SB_16_ELA_Quality_Criteria_FINAL
		SB_16_MATH_Quality_Criteria_FINAL
		CBA Item Review Business Rules 9-25
Human Scoring	Process Description	20150512 Smarter Hand Scoring Process FINAL
	Qualifications	20150512 Smarter Hand Scoring Rater Qualifications FINAL
	Quality Monitoring	20150512 Smarter Hand Scoring Quality Monitoring FINAL
	Recruitment-Training	0150512 Smarter Hand Scoring Rater Training FINAL
Data Review		20150512 Smarter 2014 Field Test Data Review Summary Report FINAL
		20150512 Smarter Data Review Results Summary

Chapter 4 Test Design

Test design entails developing a test philosophy (i.e., Theory of Action), identifying test purposes, and determining the targeted examinee populations, test specifications, item pool design, and other features such as test delivery (Schmeiser & Welch, 2006). The Smarter Balanced Theory of Action, test purposes, and the targeted examinee population were outlined in Chapter 1 (Introduction). Other elements of test design are further emphasized here, such as the interim assessments. In developing a system of assessments, the goal of Smarter Balanced was to ensure that its measurement properties reflected the expectations of content, rigor, and performance that comprise the Common Core State Standards (CCSS). The primary mechanism for this was to ensure the alignment of the Smarter Balanced assessments with the CCSS. Figure 1 briefly encapsulates the Smarter Balanced content structure.

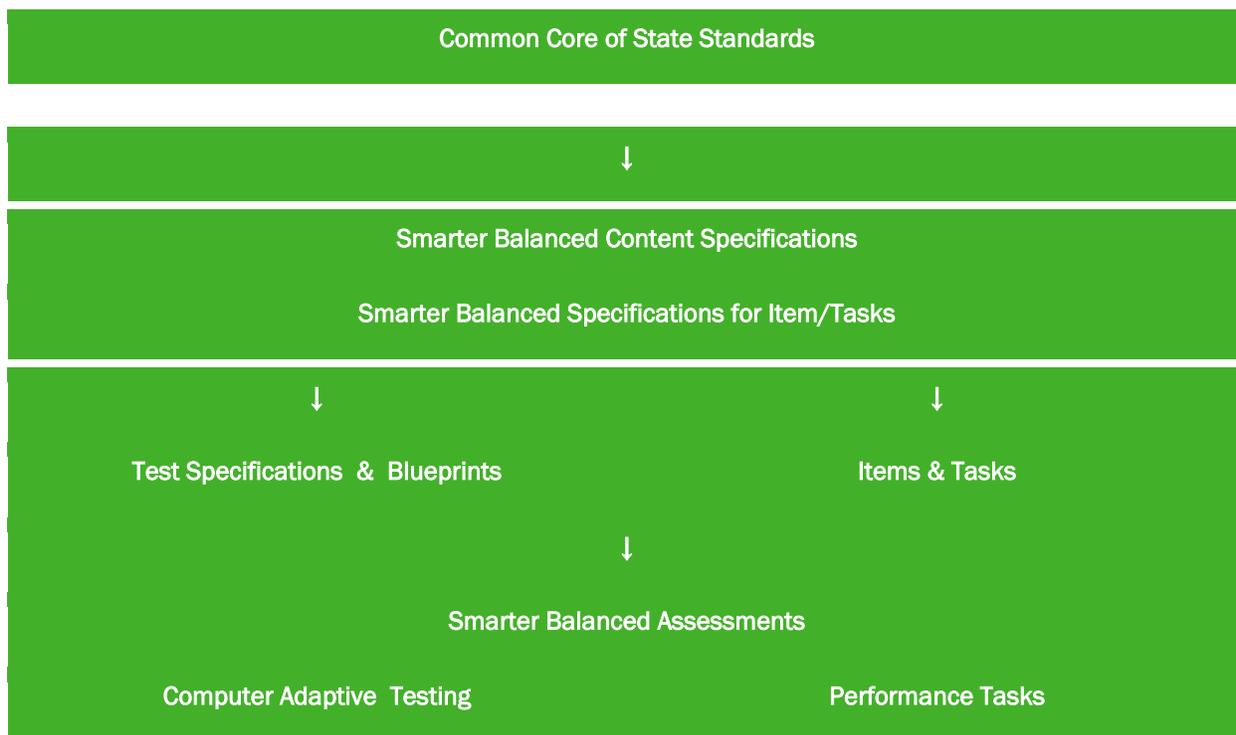


Figure 1. Relationships among Smarter Balanced Content

A Brief Description of Smarter Balanced Content Structure

The Common Core State Standards are the content standards in English language arts/literacy (ELA) and mathematics that many states have adopted. Since the Common Core State Standards were not specifically developed for assessment, they contain extensive rationale and information concerning instruction. Therefore, adopting previous practices used by many state programs, Smarter Balanced content experts produced Content Specifications in ELA and mathematics distilling assessment-focused elements from the Common Core State Standards. The content specifications were expressly created to guide the structure and content of assessment development. Within each of the two subject areas in grades 3 to 8 and high school, there are four broad claims. Within each claim, there are several assessment targets. The claims in ELA and mathematics are given in Table 1.

Table 1. Major Domains Identified for ELA and Mathematics.

Claim	ELA	Mathematics
1	Reading	Concepts and Procedures
2	Writing	Problem Solving
3	Speaking/Listening	Communicating/Reasoning
4	Research	Model and Data Analysis

Currently only the listening part of ELA Claim 3 is assessed. In mathematics, Claims 2 and 4 are reported together, so there are only three reporting categories, but four claims.

Because of the breadth in coverage of the individual claims, the targets within them are needed to define more specific performance expectations within claim statements. The relationship between targets and Common Core State Standards elements is made explicit in the Smarter Balanced content specifications. The Smarter Balanced specifications for items and tasks correspond to targets in the Smarter Balanced content requirements. For every target, a table was produced describing the evidence to be gathered to address the target and several models for items to be developed to measure student performance relative to the target. The item/task specifications and sample items developed from them are intended to guide item and task developers in the future. The item/task types include (but are not limited to) selected-response, constructed-response, technology-enhanced items that capitalize on digital media, and performance tasks. Technology-enhanced items have the same requirements as selected- and constructed-response items, but have specialized types of interaction in which students manipulate information using a defined set of responses. Constructed-response items are intended to address assessment targets and claims that are of greater complexity and require more analytical thinking and reasoning. Most constructed-response items should take between 1 and 5 minutes to complete; some more complex types may take up to 10 minutes for completion. The distinction between constructed-response items given in the computer adaptive test (CAT) and performance tasks is primarily the context in which the items are given. Performance tasks are thematically related that are preceded by an associated classroom activity. The classroom activities are not scored. Smarter Balanced test blueprints/specifications describe the composition of the two assessment components (computer adaptive test and performance assessment) and how their results will be combined for score reporting. For the computer adaptive component, specific items administered to each student are uniquely determined based on an item-selection algorithm and content constraints embedded in the test blueprint. The performance tasks (PTs) act in concert with the computer adaptive test items to fulfill the blueprint.

Synopsis of Assessment System Components

The summative assessment consists of two parts: a CAT and a performance task, which is administered on a computer but is not computer adaptive. **The summative assessment is administered during the last twelve weeks of the school year.** The summative assessment scores will

- accurately describe student achievement and can be used in modeling growth of student learning as part of program evaluation and school, district, and state accountability systems;
- provide valid, reliable, and fair measures of students' progress toward, and attainment of, the knowledge and skills required to be college- and career-ready;

- capitalize on the strengths of computer adaptive testing—efficient and precise measurement across the full range of student achievement; and
- utilize performance tasks to provide a measure of the student’s ability to integrate knowledge and skills across multiple standards.

Optional interim assessments are administered at locally determined intervals in the school calendar. These assessments provide educators with actionable information about student progress throughout the year. Interim Comprehensive Assessments (ICAs) use the same design as the summative assessments. They are designed to include both computer-adaptive and performance tasks to fulfill the test blueprint. The interim system also includes Interim Assessments Blocks (IABs), available in both fixed/linear and adaptive formats. Interim Assessments Blocks focus on more granular aspects of the content standards. In the 2014-15 school year, IABs and ICAs will be available in fixed forms only. The interim assessments will

- assist teachers, students, and parents in understanding the extent to which students are on track toward the goal of college and career readiness and identify strengths and limitations in content domains; and
- be fully accessible for instruction and professional development (non-secure).

Formative assessment practices and strategies are the basis for the Digital Library of professional development materials, resources, and tools aligned to the Common Core State Standards and Smarter Balanced Claims and Assessment Targets. Research-based instructional tools are available to help teachers address learning challenges and differentiate instruction. The Digital Library includes professional development materials related to all components of the assessment system, such as scoring rubrics for performance tasks.

Evidence-Centered Design in Constructing Smarter Balanced Assessments

Evidence-centered design (ECD) is an approach to the creation of educational assessments in terms of reasoning about evidence (arguments) concerning the intended constructs. The ECD begins with identifying the claims, or inferences, that users want to make concerning student achievement to specifying the evidence needed to support those claims, and finally, determining a specification of the items/tasks capable of eliciting that information (Mislevy, Steinberg, & Almond, 2003). Explicit attention is paid to the potential influence of unintended constructs. ECD accomplishes this in two ways. The first is by incorporating an overarching conception of assessment as an argument from imperfect evidence. This argument makes explicit the claims (the inferences that one intends to make based on scores) and the nature of the evidence that supports those claims (Hansen & Mislevy, 2008; Mislevy & Haertel, 2006). The second is by distinguishing the activities and structures involved in the assessment enterprise in order to exemplify an assessment argument in operational processes. By making the underlying evidentiary argument more explicit, the framework makes operational elements more amenable to examination, sharing, and refinement. Making the argument more explicit also helps designers meet diverse assessment needs caused by changing technological, social, and legal environments (Hansen & Mislevy, 2008; Zhang, Haertel, Javitz, Mislevy, Murray, & Wasson, 2009). The ECD process entails five types of activities. The layers focus in turn on the identification of the substantive domain to be assessed; the assessment argument; the structure of assessment elements such as tasks, rubrics, and psychometric models; the implementation of these elements; and the way they function in an operational assessment, as described below. For Smarter Balanced, a subset of the general ECD elements was used.

- **Domain Analysis.** In this first layer, domain analysis involves determining the specific content to be included in the assessment. Smarter Balanced uses the Common Core State Standards as its content domain for mathematics and ELA. Domain analysis was conducted by the

developers of the Common Core State Standards, who first developed college- and career-readiness standards, to address what students are expected to know and understand by the time they graduate from high school, followed by development of K-12 standards, which address expectations for students in elementary through high school.

- **Domain Modeling.** In domain modeling, a high-level description of the overall components of the assessment is created and documented. For Smarter Balanced, the general components of the assessment system were articulated in the proposal to the Race to the Top Assessment Program. At a high level, the components include computer-adaptive summative and interim assessments in mathematics and ELA/literacy. The domain framework was developed by organizing the Common Core State Standards into domain areas that form the structure of test blueprints and reporting categories. This overall structure was created in the course of Smarter Balanced content specification development.
- **The Conceptual Assessment Framework.** Next, the conceptual assessment framework is developed. For Smarter tests, this step was accomplished in developing the Smarter Balanced content specifications, which identify major claim structure, targets within claims, and the relationship of those elements to underlying content of the Common Core State Standards. In this step, the knowledge, skills, and abilities to be assessed (otherwise referred to as the *intended constructs* or the *targets of assessment*), the evidence that needs to be collected, and the features of the tasks that will elicit the evidence are specified in detail. Ancillary constructs that may be required to respond correctly to an assessment task but are not the intended target of the assessment are also specified (e.g., reading skills in a mathematics examination). By identifying any ancillary knowledge, skills, and abilities (KSAs), construct-irrelevant variance can be identified a priori and minimized during item and task development—potential barriers created by the ancillary KSAs can be removed or their effects minimized through the provision of appropriate access features. For Smarter Balanced, the constructs that are the target of assessment defined in blueprints were based on the content specifications. The evidence required to support claims about the Assessment Targets is also defined in the item specification tables. Ancillary constructs are elaborated on in the item specification tables. Details of these processes are described in Chapter 3 on item development.
- **Implementation.** This layer involves the development of the assessment items or tasks using the specifications created in the conceptual assessment framework just described. In addition, scoring rubrics are created and the scoring process is specified. For Smarter Balanced, items, performance tasks, and associated scoring rubrics were developed starting in the spring of 2012. This is also described in Chapter 3, Item Development.
- **Delivery.** In this final layer, the processes for the assessment administration and reporting are created. The delivery system describes the collection of student, evidence, task, assembly, and presentation models required for the assessment and how they function together. The ECD elements chosen lead to the best evaluation of the construct for the intended test purposes. Test delivery and some elements of scoring are discussed below.

Content Alignment in Smarter Balanced Test Design

In developing a system of assessments, Smarter Balanced is committed to ensuring that its measurement reflects the expectations of content, rigor, and performance that correspond to the Common Core State Standards. To that end, Smarter Balanced designed item specifications to demonstrate alignment through methodologies that reflect ECD theory. According to Webb (2002), “Alignment of expectations for student learning and assessments for measuring students’ attainment of these expectations is an essential attribute for an effective standards-based education

system.” DeMauro (2004) states, “Alignment activities . . . should be the guiding principle of test design, and item alignment studies should be sources of validity documentation, as should any studies of test content.” Test content alignment is at the core of content validity and consequential validity (Martone & Sireci, 2009). There is a connection between validity and content alignment, with validity addressing the appropriateness of inferences drawn from test results and alignment concerning “how well all policy elements [e.g., expectations and assessments] guide instruction and, ultimately, impact student learning” (Webb, 1997). The Elementary and Secondary Education Act (ESEA) now requires that state accountability assessments be aligned with state content standards. Since Consortium states have adopted the Common Core State Standards in ELA and mathematics, it was imperative that Smarter Balanced conduct the appropriate alignment studies. Accordingly, the Consortium contracted with the Human Resources Research Organization to conduct an alignment study (HumRRO, 2014).

Webb (1997) identified several categories of criteria for judging content alignment. The Smarter Balanced alignment study describes how well the Smarter Balanced tests address the expectations embodied in the Smarter Balanced content specifications and the CCSS. Test content alignment is at the core of content validity and consequential validity (Martone and Sireci, 2009). Because of the high stakes associated with statewide testing and the need to communicate learning goals during the NCLB era, attention was directed at test alignment in addition to individual item alignment. The emphasis on test content in alignment and validity studies is understandable. After all, a test is a small sampling of items from a much larger universe of possible items/tasks representing a very broad domain. For inferences from test results to be justifiable, that sample of items has to be an adequate representation of the broad domain, providing strong evidence to support claims based on the test results.

Assessment is always constrained to some extent by time and resources. Items and tasks that require extensive time (performance tasks and text responses), items that require expensive scoring, and items that require a lot of computer bandwidth (videos, animations) must be limited and chosen carefully. Smarter Balanced content experts carefully scrutinized each blueprint to assure optimal content coverage and prudent use of time and resources. In general, the Smarter Balanced blueprints represent content sampling proportions that reflect intended emphasis in instruction and mastery at each grade level. Specifications for numbers of items by claim, Assessment Target, depth-of-knowledge, and item type demonstrate the desired proportions within test delivery constraints. The blueprints were subject to state approval through a formal vote.

The alignment study conducted for the Consortium (HumRRO) discusses alignment among elements of content standards, content specifications, item specifications, and blueprints. The study itself extensive, but its overall finding is that Smarter summative tests and supporting item pools exceed levels of DOK representation recommended by Webb. The analysis is done with test blueprint, item and test specifications and item pools. The operational test had not yet been delivered at the time the analysis was completed, so further analysis will be conducted with operationally delivered test forms.

Test Blueprints

Test specifications and blueprints define the knowledge, skills, and abilities intended to be measured on an assessment. A blueprint also specifies how skills are sampled from a set of content standards (i.e., the CCSS). Other important factors such as Depth of Knowledge (DOK) are also specified. Specifically, a test blueprint is a formal document that guides the development and assembly of an assessment by explicating the following types of essential information:

- content (Claims and Assessment Targets) that is included for each assessed subject and grade, across various levels of the system (student, classroom, school, district, state);
- the relative emphasis or weighting of different content strata (e.g., claims) if there is any weighting beyond the proportions of items and points;
- the relative emphasis of content standards generally indicated as the number of items or percentage of points per Claim and Assessment Target;
- item types used or required, which communicate to item developers how to measure each Claim and Assessment Target, and to teachers and students about learning expectations; and
- Depth of Knowledge (DOK), indicating the complexity of item types for each Claim and Assessment Target.

The test blueprint is an essential guide for both assessment developers and for curriculum and instruction. For assessment developers, the blueprint and related test-specification documents define how the test will ensure coverage of the full breadth and depth of content and how it will maintain fidelity to the intent of the Common Core State Standards on which the Smarter Balanced assessment is based. Full content alignment is necessary in order to ensure that educational Stakeholders can make valid, reliable, and unbiased inferences concerning students, classrooms, schools, and state levels. At the instructional level, the test blueprint provides a guide to the relative importance of competing content demands and suggests how the content is demonstrated, as indicated by item type and depth-of-knowledge. In summary, an assessment blueprint provides clear development specifications for test developers and signals to the broader education community both the full complexity of the Common Core State Standards and how performance on these standards are substantiated.

Part of the innovative aspect of the Smarter Balanced assessments is that the test blueprints sample the content domain using both a computer adaptive component (CAT) and a performance task. The test blueprints can be inspected to determine the contribution of the CAT and performance task components in a grade and content area toward the construct intended to be measured. Another aspect of the assessments is the provision of a variety of both machine-scored and human-scored item types. The contribution of these item types is specified in the Smarter Balanced test blueprints.

The Governing States of the Smarter Balanced Assessment Consortium adopted blueprints for the summative assessments of mathematics and ELA/literacy for grades 3 to 8 and high school. Final blueprints for the Smarter Balanced summative assessments will be adopted by Governing States prior to full implementation in the 2014-15 school year. In part, two objectives for the Pilot and Field Tests were to try provisional item types and perform scaling with a representative student sample. Blueprints used for the Field Test were “preliminary” since they used assessment design features that could be refined and revised after Field Test analysis.

Summative Assessment

The summative assessment is composed of the CAT and performance task components, which are described in further detail here. Performance information from both components are combined to sample the test blueprint in a grade and content area and eventually used to produce the overall scale score.

Operational Summative Assessment Blueprints and Specifications. For each designated grade range (3 to 5, 6 to 8, and high school), the blueprint overviews summarize the claim score\reporting category, content category, stimuli used, items by CAT or performance tasks, and total number of

items by claim. Details are given separately for each grade and include Claim, Assessment Target, DOK, item type (CAT/PT), and the total number of items. The Assessment Targets are nested within claims and represent a more detailed specification of content. Note that in addition to the nested hierarchical structure, each blueprint also specifies a number of rules applied at global or claim levels. Most of these specifications are in the footnotes, which constitute important parts of the test designs.

The CAT algorithm selects items necessary to conform to the test blueprint and at the same time meet the IRT target information function. In establishing target requirements for the CAT, designers took advantage of the adaptive pool to allow more variety than would be present in a fixed form test. For example, when the number of targets in a domain area is large, blueprints allow choice within target clusters rather than limiting the number of targets. Since all targets are represented in the pool, any student could potentially get any target while the full set of content constraints is still maintained.

To assist in blueprint interpretation, an overview of the grade 6 mathematics test blueprint is given. Figure 2, for grade six mathematics, presents requirements for each Claim by Assessment Target. It displays the number of items overall by claim and shows the contribution of the CAT and performance task portions to the overall design. Note that some Targets are clustered together. For example, Claim 1 calls for 14 items from targets E, F, A, G, B, and D. Note that six items come from targets E and F, while only two items come from G and B. This represents the appropriate content emphasis, while allowing flexibility in item choice. The detailed blueprint shows how performance tasks and CAT components work in conjunction. Here, the DOK requirements are applied at the claim level, although DOK ranges are listed for each target. Performance tasks are delivered as a fixed set of items within a set of performance tasks common to a class or school.

Target Sampling Mathematics Grade 6						
Claim	Content Category	Assessment Targets	DOK	Items		Total Items
				CAT	P	
1. Concepts and Procedures	Priority Cluster	E. Apply and extend previous understandings of arithmetic to algebraic expressions.	1	5-6	0	16-19
		F. Reason about and solve one-variable equations and inequalities.	1, 2			
		A. Understand ratio concepts and use ratio reasoning to solve problems.	1, 2	3-4		
		G. Represent and analyze quantitative relationships between dependent and independent variables.	2	2		
		B. Apply and extend previous understandings of multiplication and division to divide fractions by fractions.	1, 2			
		D. Apply and extend previous understandings of numbers to the system of rational numbers.	1, 2	2		
	Supporting Cluster	C. Compute fluently with multi-digit numbers and find common factors and multiples.	1, 2	4-5		
		H. Solve real-world and mathematical problems involving area, surface area, and volume.	1, 2			
		I. Develop understanding of statistical variability.	2			
		J. Summarize and describe distributions.	1, 2			

- DOK: Depth of Knowledge, consistent with the Smarter Balanced Content Specifications.
- The CAT algorithm will be configured to ensure the following:
 - For Claim 1, each student will receive at least 7 CAT items at DOK 2 or higher.
 - For Claim 3, each student will receive at least 2 CAT items at DOK 3 or higher.
 - For combined Claims 2 and 4, each student will receive at least 2 CAT items at DOK 3 or higher.

Figure 2. Blueprint for grade 6 showing detailed content structure (Assessment Targets), page 1 of 2

Target Sampling Mathematics Grade 6						
Claim	Content Category	Assessment Targets	DOK	Items		Total Items
				CAT	PT	
2. Problem Solving 4. Modeling and Data Analysis	Problem Solving (drawn across content domains)	A. Apply mathematics to solve well-posed problems arising in everyday life, society, and the workplace.	2, 3	2	1–2	8-10
		B. Select and use appropriate tools strategically. C. Interpret results in the context of a situation. D. Identify important quantities in a practical situation and map their relationships (e.g., using diagrams, two-way tables, graphs, flow charts, or formulas).	1, 2, 3	1		
	Modeling and Data Analysis (drawn across content domains)	A. Apply mathematics to solve problems arising in everyday life, society, and the workplace. D. Interpret results in the context of a situation.	2, 3	1	1–3	
		B. Construct, autonomously, chains of reasoning to justify mathematical models used, interpretations made, and solutions proposed for a complex problem. E. Analyze the adequacy of and make improvements to an existing model or develop a mathematical model of a real phenomenon.	2, 3, 4	1		
		C. State logical assumptions being used. F. Identify important quantities in a practical situation and map their relationships (e.g., using diagrams, two-way tables, graphs, flow charts, or formulas). G. Identify, analyze, and synthesize relevant external resources to pose or solve problems.	1, 2, 3 3, 4	1 0		
3. Communicating Reasoning	Communicating Reasoning (drawn across content domains)	A. Test propositions or conjectures with specific examples. D. Use the technique of breaking an argument into cases.	2, 3	3	0–2	8-10
		B. Construct, autonomously, chains of reasoning that will justify or refute propositions or conjectures. E. Distinguish correct logic or reasoning from that which is flawed, and—if there is a flaw in the argument—explain what it is.	2, 3, 4	3		
		C. State logical assumptions being used. F. Base arguments on concrete referents such as objects, drawings, diagrams, and actions. G. At later grades, determine conditions under which an argument does and does not apply. (For example, area increases with perimeter for squares, but not for all plane figures.)	2, 3	2		

- DOK: Depth of Knowledge, consistent with the Smarter Balanced Content Specifications.
- The CAT algorithm will be configured to ensure the following:
 - For Claim 1, each student will receive at least 7 CAT items at DOK 2 or higher.
 - For Claim 3, each student will receive at least 2 CAT items at DOK 3 or higher.
 - For combined Claims 2 and 4, each student will receive at least 2 CAT items at DOK 3 or higher.

Figure 3. Blueprint for grade 6 showing detailed content structure (Assessment Targets), page 2 of 2

CAT and Performance Task Test Components

Part of the Smarter Balanced Theory of Action is to leverage appropriate technology and innovation. Two primary assessment components are administered for either summative or interim test purposes. These consist of a CAT and a separately administered performance task. Both components can be administered, and associated information can be accessed online. The use of CAT methodologies helps ensure that students across the range of proficiency have an assessment experience that presents them with items that are well targeted at their skill level. The intention is that average-, very low-, and very high-performing students will be more likely to stay engaged in the assessment because they will be responding to items specifically targeted to their skill level. Performance tasks are intended to measure a student's ability to integrate knowledge and skills across multiple standards.

The CAT tests should be more efficient in that fewer items can be administered compared with fixed forms to achieve a comparable level of score precision. For the CAT, there are both content constraints (e.g., a long reading passage in ELA must be administered) as well as psychometric criteria that must be optimized for each student. **Performance tasks are intended to** measure a student's ability to integrate knowledge and skills across multiple standards in a coherent task that requires using integrated skill sets. Performance tasks are used to better measure capacities such as depth of understanding, research skills, and complex analysis, which cannot be completely assessed with individual, discrete items. Some constructed-response items and performance tasks are scored automatically; others are hand-scored by trained raters. Each performance task is preceded by a brief classroom-interaction activity that is grouped into a larger theme for administration.

The Classroom Activity component is an innovative element designed in concert with assessment experts from the Shell Centre, Student Achievement Partners, and Stanford University. The intent of the Classroom Activity is to provide context for the performance tasks. This allows students to demonstrate skills and knowledge without interference from lack of background knowledge or vocabulary. The Classroom Activity does not address the assessed skills but describes the setting and provides related examples or terms. Since performance tasks are often applied using skills in real world settings, the Classroom Activity provides users with external information so that no student is given an advantage or disadvantage based on personal experience.

Operational Adaptive Test Design

Automated test assembly for a CAT depends on a number of factors to produce conformable tests. These depend on the quality of the item bank, reasonableness of the test constraints and precision targets, and the degree to which content or other qualitative attributes of items are salient and can be defined as constraints (Luecht, 1998).

For the operational test, an item-level, fully adaptive test is planned in ELA and mathematics. The adaptive part of summative and interim comprehensive tests is designed to deliver the CAT portion of blueprints in a manner that efficiently minimizes measurement error and maximizes information. Efficiency is interpreted as fewer items being needed compared with non-adaptive (fixed) test forms. The Smarter Balanced Consortium provides a specific CAT delivery engine, but states may choose to use other engines as long as they can deliver a conforming test blueprint with a minimum degree of error, avoid item over- or under-exposure, and provide the design features specified by Smarter Balanced. This section outlines some of the intended design features for the operational adaptive test component.

Early in the development process, Consortium states established a desire to allow students to go back to earlier questions and change their answers. This has implications for test design and delivery. If a student takes a test over the course of two or more days, answers from previous days

cannot be changed. In mathematics, some items permit the use of a calculator, while others forbid calculator use. Mathematics tests are consequently divided into two sections, one for non-calculator items, and one that permits calculator use. Students can change answers within sections but not across different test sections.

Test blueprints display the proportions of items in each area, but not in the order in which students will encounter them. The adaptive algorithm presents passages and items at varying stages. In ELA, the first item can come from either Claim 2 or Claim 4 and must be a machine-scored item. Once the first claim area is administered, the software iterates through claim areas so that the test does not converge on a final score based on a single claim area. In mathematics, the first item can be assigned from any claim.

Expansion of the Item Pool

Under certain conditions, the item pool will expand to include items from adjacent grades that address content in the target test grade. Pool expansion occurs when the following conditions have been met:

- On-Grade content coverage requirements have been met—this is the point at which over 60% of the CAT session has been administered and all claims have been sampled.
- Estimate of performance is clearly far below or far above the proficiency cut score.
- Items in the expanded pool will better satisfy content and measurement requirements.

The algorithm selects items until a defined percentage of the test has been administered, sampling items from all claim areas. A decision point is reached when a substantial proportion of content has been covered. The rules for ELA/literacy and mathematics are the following:

- For the ELA CAT (no human-scored)
 - 1 long info passage (5 items)
 - 1 long lit passage (5 items)
 - 2 listening passages (6 items)
 - 6 writing items (6 items)
 - 5 research items (5 items)
 - Total 27 items out of 44—61%
- For the mathematics CAT (no human-scored)
 - Claim 1 (14 items)
 - Claims 2 & 4 (2 items)
 - Claim 3 (4 items)
 - Total 21 items out of 32—62%

At this point, the distance of the estimated score from the college content readiness cut score is evaluated. This is Level 3 as defined in the Achievement Level Descriptors (see Chapter 10 on standard setting for further details). If there is a determination that the student is in either Level 1 or Level 4 as defined by the Achievement Level Setting Report, the item pool is expanded to include items from no more than two adjacent grades in either direction. In grade 3, the expansion includes items from adjacent upper grades only; in grade 11, only adjacent lower grades are included. Items from adjacent grades have been reviewed for appropriateness by content experts to ensure that they

are instructionally and developmentally appropriate for the targeted grades. For the remainder of the test, both on-grade and off-grade items can be administered. The item with the best content and measurement characteristics is chosen from the pool. Students at or near the cut score when the decision point is reached do not have an expanded pool, but continue with the original pool. For all students, the algorithm delivers the remainder of the blueprint until termination of the test, once all test constraints are met.

Performance Task Design

The Race to the Top Assessment Program Application for the Smarter Balanced Assessment Consortium highlights the importance of performance tasks to “provide a measure of the student’s ability to integrate knowledge and skills across multiple standards—a key component of college and career readiness” (page 42). The development of an assessment system that fulfills this goal necessitates an understanding of how the world is changing and what skills are required to compete in an increasingly global economy. Research suggests that measuring college and career readiness will increasingly require the use of performance-based assessments (Fadel, Honey, & Pasnik, 2007).

A key component of college and career readiness is the ability to integrate knowledge and skills across multiple content standards. Smarter Balanced derives inferences concerning this ability through performance tasks. Performance assessments are intended to represent students’ competencies in applying the requisite knowledge and cognitive skills to solve substantive, meaningful problems. Performance assessments give students opportunities to demonstrate their ability to find, organize, or use information to solve problems, undertake research, frame and conduct investigations, analyze and synthesize data, and apply learning to novel situations.

A Smarter Balanced performance task involves interaction of students with stimulus materials and/or engagement in a problem solution, ultimately leading to an exhibition of the students’ application of knowledge and skills, often in writing. Stimuli include a variety of information forms (e.g., readings, video clips, data), as well as an assignment or problem situation. As shown in the test blueprints, performance tasks are an integral part of the Smarter Balanced test design. When a performance task is assigned and given in its entirety, it fulfills a specific role in the test blueprint for a grade and content area. Performance tasks are intended to challenge students in applying their knowledge and skills to complex, contextually rich problems. These activities are meant to measure capacities such as depth of understanding, writing or research skills, and complex analysis. They consist of collections of questions and activities coherently connected to a single scenario. The performance tasks are administered online via computer (not computer adaptive) and require one to two class periods to complete.

Prior to online administration of the performance task, students engage in non-scored classroom interactions that provide all students an opportunity to gain access to key information in sources before they complete the assigned task. The purpose of these classroom interactions is to create a more level playing field by mitigating the effect of unfamiliar terms or situations. Classroom Activities provide instructional connection, an important part of the Smarter Balanced Theory of Action. When teachers are directly involved in the administration of the task, classroom-based activities have the potential to positively influence teaching. These classroom-based preparatory activities are intended to have positive outcomes for instruction and learning and to provide avenues for teacher professional development by demonstrating good instructional and assessment practice. Tasks are designed to allow for brief context setting and to reduce construct-irrelevant variance. Task models for the scored independent performance task work do not depend on the pre-work classroom activities conducted by the teacher or with classroom peers. In mathematics, the teacher might engage students in an authentic data collection. In ELA, the teacher might tie together key points from a video that students observed. Classroom Activities can help mitigate potential conflation

between reading skills and writing or mathematics results and may increase accessibility to higher scores for students with reading deficiencies.

Performance tasks have a high likelihood of introducing task-specific variance because students have varying levels of knowledge about a particular topic. Classroom activities can reduce this variance by allowing teachers and students to gain familiarity about the context for a problem (see Abedi, 2010). For example, in a mathematics task about designing a merry-go-round, it is important for all students to understand what a merry-go-round is, how it works, and that it comes in many shapes. By involving the teacher in the process of exploring the context (but not the construct), all students enter the task with more similar levels of understanding about the task's primary theme. Engaging teachers in the process of task administration is consistent with the Smarter Balanced commitment to building an assessment system that supports teaching and learning.

Performance tasks were constructed so they can be delivered effectively in the school/classroom environment (Dana and Tippins, 1993). Requirements for task specifications included, but were not limited to, pre-assessment classroom activities, materials and technology needs, and allotted time for assessment. Performance tasks adhered to a framework of specifications used by item writers to develop new tasks that focus on different content but were comparable in contribution to the blueprint.

All Smarter Balanced performance tasks consist of three basic components: stimulus presentation, information processing, and scorable product(s) or performance(s). "Information processing" means student interactions with the stimulus materials and their content. It could include note taking, data generation, and any other activities that increase students' understanding of the stimulus content or the assignment. All activities within a task must have a rationale for inclusion (e.g., to increase understanding, for scaffolding, as early steps in product creation or for product creation). More detail on the possibilities within the three basic process components is presented in the specifications for ELA/literacy and mathematics performance tasks in Chapter 3.

In ELA, each classroom-based performance task comprises a targeted research effort in which students read sources and respond to at least three research items. During this research component, students may take notes to which they may later refer. After the research questions are completed, students write a full essay drawing from source material and research notes. Together, the research items and the composition of full texts using the writing process correspond to the classroom-based performance tasks in the summative assessment, the comprehensive interim assessment, and in the ELA performance task interim blocks. Claim level results in writing and research are based on both CAT and performance task item responses.

In mathematics, each classroom-based performance task comprises a set of stimulus materials and a follow-up item set consisting of six items in Claims 2, 3, and 4 that permit the complete blueprint to be met. Performance tasks address an integrated task in middle and high school and a common theme in grades 3 to 5. Note that results for Claims 2, 3, and 4 are derived from scored responses to both performance tasks and CAT items.

Test Scoring

The method of combining item level scores to produce test scores and subscores is presented in detail in the Test Score Specifications document (AIR, 2014). Scores are calculated using maximum likelihood estimation (MLE) applied at the overall and subscore levels. No special weights for claims, item types or performance tasks are applied. Desired score effects are achieved by content proportions in the blueprints.

Field Test Delivery Modes

For Smarter Balanced operational administrations, a CAT test will be given along with a classroom-based, thematically related performance task where the context and assessment experiences differ from the CAT. The design for the Field Test essentially followed these two test components. For the Field Test, the test delivery modes corresponded to the two separately delivered events, one for the CAT and one for the performance task.

The performance tasks were delivered using computerized fixed forms/linear administrations. For a given performance task, students saw the same items in the same order of presentation and associated test length. Since performance tasks are classroom-based and organized thematically, they were randomly assigned within Classroom Activities assigned at the school and grade level in the Field Test. There was no administration ordering of the two components. Students could take either the CAT or the performance task first.

During the CAT component of the Field Test, linear-on-the-fly testing (LOFT) was used (Gibson & Weiner, 1998; Folk & Smith, 2002). LOFT delivers tests assembled dynamically to obtain a unique test for each student from a defined item pool. Note that a LOFT is similar to a CAT in applying content constraints to fulfill the test blueprint. Each student should obtain a content-conforming unique test form. The major differences between LOFT and item level adaptive testing is that no IRT item statistics are used in the administration and adaptation based on student response/ability is not incorporated into the delivery algorithm. For dynamic real-time LOFT, item exposure control (e.g., Hetter & Sympson, 1997) can be used to ensure that uniform rates of item administration are achieved. That is, it is not desirable to have some items with many observations and others with correspondingly few in comparison. The LOFT administration is closer to the operational CAT so that there are some advantages for IRT scaling. This permits the scaling to reflect the operational CAT deployment. For the test administration, delivering parallel fixed-test forms with potentially thousands of items in a pool in a given grade and content area was not possible. The major advantage of using LOFT was that parallel test forms could be constructed dynamically using the test delivery algorithm. The disadvantage is that some measures of test functioning are not directly available using LOFT. Classical statistics such as observed test reliability cannot be computed since every student essentially takes a unique test form. Even the definition of a criterion score for item-test correlation and for differential item functioning must rely on Item Response Theory (IRT) methods for computing these statistics.

Measurement Models (IRT) Adopted

A unidimensional scale was conceptualized that combines both CAT and performance tasks. The results from the Pilot Test factor analysis study supported the use of a unidimensional scale, both within a grade and across grades in ELA and mathematics, which are presented in detail in the Pilot Test (Chapter 5). Since no pervasive evidence of multidimensionality was shown, the decision was to adopt a unidimensional model for scaling and linking. For the choice of an IRT model, examination of model fit using chi-square showed significant improvement of the two-parameter model over the one-parameter model. Use of the three-parameter logistic model did not significantly improve model fit. Consequently, after discussion with the Smarter Balanced Technical Advisory Committee, a two-parameter unidimensional model was adopted for dichotomous data. The generalized partial credit mode (GPCM, Muraki, 1992) was used in the case of polytomous items (i.e., constructed-response). These models were used in scaling, achievement level setting, and the first years of operational testing. The Consortium plans to revisit the scale and model decisions using a solid base of operational data.

Interim Assessment

The purpose of the Smarter Balanced interim assessment system in mathematics and ELA is to complement the Smarter Balanced summative assessment by

- providing meaningful information on student progress toward mastery of the skills measured by the summative assessment;
- serving as a repository of items and tasks for assessing the Common Core State Standards at strategic points during the school year;
- yielding actionable information on student skills and understanding in instructionally targeted areas of interest; and
- supporting teaching and learning inside and outside of the classroom.

The items on the interim assessments are developed under the same conditions, protocols, and review procedures as those used in the summative assessments and is on the same scale. The items assess the Common Core State Standards, adhere to the same principles of Universal Design to be accessible to all students, and provide evidence to support all Smarter Balanced claims in mathematics and ELA. The application of the same ECD processes and procedures in the development of items and tasks for the interim system ensures that each item or task clearly elicits student responses that support the relevant evidence that is aligned to the associated content standards. The interim assessments are available in grades 3 to 8 and high school. Items for the interim assessments have been administered in the Field Test with all appropriate reviews and scoring applied. The Consortium plans to provide fixed-form Interim Comprehensive Assessments (ICAs) and fixed-form Interim Assessment Blocks (IABs) that include universal tools, designated supports, and accommodations listed in the *Usability, Accessibility, and Accommodations Guidelines*.

The Interim assessments include two distinct types of tests that draw from the same bank of items and performance tasks:

- Interim Comprehensive Assessments (ICAs) use the same blueprints as the summative assessments, assessing the same range of standards, and use the same score-reporting categories. The ICAs include the same item types and formats, including performance tasks, as the summative assessments, and yield results on the same vertical scale. They are administered with the same computer-adaptive algorithm or with the option of a fixed form. The ICAs yield overall scale scores, overall performance level designations, and claim-level information.
- Interim Assessment Blocks (IABs) focus on smaller sets of targets and therefore provide more detailed information targeted at instructional purposes. The blocks are available either as fixed forms or with the use of a computer-adaptive algorithm. The IABs are comprised of several blocks of items and yield overall information for each block. Each block measures a smaller set of targets than does the ICA. These smaller assessments focus on a particular cluster of standards and therefore provide more instructionally relevant types of feedback. They may be computer adaptive or linear, and results are reported on the same scale as the summative assessment with the caveat that the full summative system takes into account a broader range of content.

Fixed-form Interim Comprehensive Assessments (ICAs) and fixed-form Interim Assessment Blocks (IABs) include the universal tools, designated supports, and accommodations listed in the *Usability, Accessibility, and Accommodations Guidelines*. Table 2 gives an overview of interim assessment features. The interim assessments provide results that teachers and administrators can use to track

student progress throughout the year in relation to the Common Core State Standards and to adjust instruction accordingly. The full range of assessment options in the interim system will ultimately depend on the assessment purpose and use of scores, security needs, and the system's technical capabilities, such as secure high school end-of-course assessments to support state-level accountability systems. The ICAs and IABs are available in grades 3 to 8 and high school and can be administered at any time during the school year. The high-school ICAs are constructed to be consistent with the grade 11 summative blueprints. High school IABs are constructed to focus on content that would be appropriate across grade levels. Schools or districts may choose to administer the high school interim assessments in grades 9 to 12. The high school ICA and IAB are constructed to be consistent with the grade 11 blueprint; however, the high school ICA and IAB may still be administered in grades 9 to 12. In addition, the interim assessments are not constrained by grade level; in other words, students may take an off-grade level Interim assessment. For example, a fifth-grade ICA/IAB can be administered to grades above or below fifth grade. The item bank in the initial rollout of the interim assessments will be limited in depth of the available content. Therefore, if ICAs and IABs are administered repeatedly to the same students, individuals may be exposed to the same items on occasion. There are no security expectations for the items in the Interim assessment item bank. The interim assessments are not intended for accountability purposes. Table 3 gives the IABs available for ELA/literacy, and Tables 4 and 5 present them for mathematics.

The scoring of human-scored aspects of constructed-response items and performance tasks for interim is a local/state responsibility. Items can be scored automatically by the Smarter Balanced engine, except for human-scored aspects of performance tasks or selected CAT items, which can be scored locally by teachers or in support of professional development or by professional raters according to established standards for accuracy and fairness.

ELA/Literacy ICA Blueprints

The ELA ICA blueprints summarize coverage of items by grade band (3 to 5, 6 to 8, and 11). Each blueprint specifies the numbers of items by claim (1–4) and content category, item type, and scoring method (machine scored or hand scored). The short-text items (two in Reading and one in Writing) are designed to be hand scored but may eventually be machine scored with an application that yields similar results to hand-scoring.

Like the Summative assessments, the ICAs will report an overall ELA score and scores for four claim-reporting categories for each grade band, each of which will be reported with the overall ELA score. Because the ICAs use the same blueprints as the Summative assessments, the ICA blueprints for both the adaptive and fixed forms begin with the same three-page summary as the ELA/literacy Summative assessment blueprint. The only difference is that the ELA fixed-form summary does not refer to CAT items; instead, it refers to these items as non-PT (non-performance task).

The grade band blueprints for the ICAs mirror the summative blueprints exactly in terms of formatting. Each blueprint specifies the number of items by claim and content category, the number of items within each claim for all Assessment Targets, DOK levels, and numbers of items by type (machine scored, short text, and performance task). The ICA adaptive-form blueprint reflects the same allocation of items (including ranges of items where appropriate) as the Summative blueprint. Where item allocations had been specified as ranges in the ICA adaptive-form blueprint, those ranges were adjusted in the fixed-form blueprint to ensure appropriate levels of coverage of each assessment target relative to the other assessment targets in the ICA fixed form.

Mathematics ICA Blueprints

The blueprint for the mathematics Summative assessment summarizes coverage of items by grade band (3 to 5, 6 to 8, and 11). The numbers of items (including performance tasks and other constructed-response items) by claim (1–4) are specified in the blueprint. In addition, Claim 1 items

are further specified by priority cluster or supporting cluster, with priority and supporting clusters defined in the Smarter Balanced Content Framework for Mathematics. All CAT items in grades 3 to 5 are designed to be machine scored. Claim 2 (problem solving) and Claim 4 (modeling and data analysis) have been combined because of content similarity and to provide flexibility for item development. In grades 6 to 8 and 11, one item per student (from either Claim 3 Target B or Claim 4 Target B) is designated for hand-scoring, which might be machine scored with an application that yields comparable results. There are still four claims, but only three claim scores will be reported with the overall mathematics score. Since the ICAs use the same blueprints as the Summative assessments, the blueprints for both the adaptive and fixed forms of the ICAs for mathematics begin with the same three-page summary as the mathematics summative assessment blueprint.

The ICA blueprints are organized by grade level (3 to 8 and 11). The ICA blueprints mirror the Summative blueprints exactly in terms of formatting. Each blueprint specifies the number of items by claim, and for Claim 1 only, also by priority or supporting cluster. Within each claim, the number of items for all assessment targets associated with the claim is also specified. Finally, within the Assessment-Target-level allocations, possible DOK levels are indicated along with numbers of CAT and performance tasks. The ICA adaptive-form blueprint reflects the same allocation of items (including ranges of items where appropriate) as the summative blueprint. Item allocations that were specified as ranges in the ICA adaptive-form blueprint were adjusted in the fixed-form blueprint to ensure appropriate levels of coverage of each assessment target relative to the other assessment targets in the ICA fixed form.

Interim and Summative Test Administration and Reporting. Both the ICA and IAB components are administered online through the Open Source Test Administration System. Since the purpose of the Smarter Balanced Interim assessment is to provide educators with student-level, CCSS-related results that can be used to adjust instruction, the interim assessments may be administered at multiple points throughout the school year. The administration schedule can be determined by each locale, with some states determining the administration of the interim assessment and others leaving the administration schedule up to schools/districts. There is no system limit on the number of times that the ICA and/or IAB can be administered.

The Summative Assessment will report an overall achievement level designation for a grade and content area and classification at the claim level. The reports will include an overall scale score with error band endpoints and an achievement level per content area as well as claim-level scores. At the claim level, students are assigned to one of three levels of classification (“Below Standard,” “At/Near Standard,” “Above Standard”) related to the overall scale-score at the achievement level 2/3 cut point. The ICA reporting has the same reporting structure as the summative assessment. Likewise for the IAB, students will be classified into one of three levels (“Below Standard,” “At/Near Standard,” “Above Standard”) related to the overall scale-score at the proficient achievement level.

Table 2. Summary of Interim Test Features for ICAs and IABs.

Feature	Interim Comprehensive Assessments (ICAs)	Interim Assessment Blocks (IABs)
Description and Purpose	ICAs meet the blueprint of the summative assessment. They provide teachers with information on a student's <ul style="list-style-type: none"> • general areas of strength or need based on the CCSS and/or • readiness for the end-of-year summative assessment. 	The IABs are short, focused sets of items that measure several assessment targets. Results provide teachers with information about a student's strengths or needs related to the CCSS. The number of blocks varies by grade and subject area. There are between five and seventeen blocks per subject per grade.
Blueprint Characteristics	The ICAs are consistent with the associated Summative blueprint. <ul style="list-style-type: none"> • ICAs will be provided as fixed forms. • ICAs will also be adaptive when the item pool is larger. 	IABs assess the same targets by grade level as specified in the Summative blueprints. <ul style="list-style-type: none"> • IABs will be provided as fixed forms and will be provided as items become available. • IABs will also be adaptive as appropriate when sufficient items are available.
Score Reporting	ICA reporting is the same as for the Summative assessment: <ul style="list-style-type: none"> • Overall scale score with error band endpoints and achievement level per content area/subject. • Claim score reporting is based on three classifications related to the overall scale score cut point between levels 2 and 3. 	Individual student scores are available for each block. Reporting for each block is based on three classifications related to the overall scale score cut point between levels 2 and 3: <ul style="list-style-type: none"> • Below Standard • At/Near Standard, and • Above Standard.

Table 3. Summary of ELA Interim Assessment Blocks.

Grades 3–5	Grades 6–8	High School
Read Literary Texts	Read Literary Texts	Read Literary Texts
Read Informational Texts	Read Informational Texts	Read Informational Texts
Edit/Revise	Edit/Revise	Edit/Revise
Brief Writes	Brief Writes	Brief Writes
Listen/Interpret	Listen/Interpret	Listen/Interpret
Research	Research	Research
Narrative PT	Narrative PT	Explanatory PT
Informational PT	Explanatory PT	Argument PT
Opinion PT	Argument PT	

Table 4. Summary of Mathematics Interim Assessment Blocks.

Grade 3	Grade 4	Grade 5
Operations and Algebraic Thinking	Operations and Algebraic Thinking	Operations and Algebraic Thinking
Numbers and Operations in Base 10	Numbers and Operations in Base 10	Numbers and Operations in Base 10
Fractions	Fractions	Fractions
Measurement and Data	Geometry	Geometry
	Measurement and Data	Measurement and Data
Mathematics PT	Mathematics PT	Mathematics PT
Grade 6	Grade 7	Grade 8
Ratio and Proportional Relationships	Ratio and Proportional Relationships	Expressions & Equations I (and Proportionality)
Number System	Number System	Expressions & Equations II
Expressions and Equations	Expressions and Equations	Functions
Geometry	Geometry	Geometry
Statistics and Probability	Statistics and Probability	
Mathematics PT	Mathematics PT	Mathematics PT

Table 5. High School Mathematics Assessment Blocks.

High School	
Algebra and Functions	Linear Functions Quadratic Functions Exponential Functions Polynomial Functions Radicals Functions Rational Functions Trigonometric Functions
Geometry	Transformations in Geometry Right Triangle Ratios in Geometry Three-Dimensional Geometry Proofs Circles Applications
Interpreting Categorical and Quantitative Data	
Probability	
Making Inferences and Justifying Conclusions	
Mathematics Performance Task	

Pool analysis and adequacy: Background and Recommendations

The quality of a CAT is highly dependent on the quality of the item pool. Quality is primarily related to how well the content constraints and statistical criteria can be met. The content specifications are defined as a combination of item attributes that tests delivered to students should have. There are typically constraints on item content such that they must conform to coverage of a test blueprint. If there are many content constraints and a limited pool, then it will be difficult to meet the CAT specifications. For a given content target, if the available difficulty/item information targeted at a given level ability is not available, then estimation error cannot be reduced efficiently. A third

dimension is that there is usually some need to monitor the exposure of items such that the “best” items are not administered at high rates relative to other ones. Therefore, the quality of the item pools is critical to achieving the benefits that accrue for the CAT over fixed test forms. Quantification of pool adequacy prior to simulation could be accomplished either through the Reckase (2003) “bin” method or the van der Linden (2005) “shadow test” method. Both involve an inventory of items by required blueprint elements and information ranges.

Partitioning the Item Pool. A central question is how many items and what types of items need to be in a pool. Ideally, the more items there are, the better the assessment, because more items allow for greater choice in test assembly and reduced exposure of items. Larger pools typically result in more items that match content criteria, item format, and statistical requirements. For Smarter Balanced, the available summative item pool comprises all items not used in the Interim assessment or in Ordered Item Booklets used in achievement level setting.

For the Summative assessment, a robust pool is necessary to implement the CAT efficiently and maintain exposure control. Since there are a finite number of performance tasks and they are not part of the CAT delivery, these can simply be assigned using some simple decision rules. Once the CAT is partitioned for a grade, the subset of items from adjacent grades can also be evaluated. The preferred method for partitioning the item pools would be to use simulations with the CAT delivery engine to ensure that the constraints could be met reasonably. Barring that, other methods could be used to stratify items by item difficulty and content domain (Claims and Assessment Targets). The problem is to ensure that the summative test has a preponderance of easier and more highly discriminating items since the census pools contain many difficult items.

Evaluating Item Pool Quality Using Simulation. Computer simulation can be employed to evaluate the properties of an item pool after the items have been developed and calibrated. In order to evaluate the delivery system and item pool, the following criteria should be taken into account:

- the fidelity of each test event (Summative and Interim), both real and simulated, to test blueprints and specifications;
- measurement errors for simulated scores for both overall and claim subscores;
- test information functions;
- recovery of simulated examinee ability, including analysis of estimation bias and random error; and
- analysis of summative/interim pool adequacy for scores and claim subscores.

Simulations play an important role in evaluating the operational adaptive algorithm and delivery system. The simulations should be specific to the Smarter Balanced assessments, using item parameter estimates from the Field Test and simulated test taker populations representative of the population of member states. It is suggested that the simulation include 1000 simulees at a given number of theta values (say 20) equally spaced between -4 and +4 and then run each simulee through the adaptive algorithm. The results of those 20,000 test events and resulting ability estimates per item pool can be summarized to examine the degree to which the algorithm and resulting scores meet the criteria outlined below. While simulations are convenient to conduct, they provide only one source of evaluation data. There is always a risk that the simulations may not adequately predict what happens when real students are administered real tests. For that reason, wherever possible, the results from actual students that participated in the Field Test, as well as from simulated cases, need to be examined.

Fidelity of Each Summative/Interim Test Event. Early comparisons of adaptive-testing procedures were made with regard to a narrow set of criteria. Foremost among these was test precision or its

close measure, test efficiency. In a simulation, precision is defined as the degree to which the true underlying proficiencies of simulated test takers are recovered by simulated tests. Efficiency is simply test precision divided by test length. Both precision and efficiency are highly prized because these are the principal “value added” features of adaptive testing. However, when a primary goal of testing is to find out what a student knows about a certain number of content criteria or subscores, consistent content coverage assumes the greatest importance. A conforming test is one that meets all the requirements imposed upon it. Conforming tests, therefore, comply with all content constraints, minimize item over- and under-exposure, and measure to optimal levels of precision. A better test administration algorithm is one capable of delivering conforming tests with the best item exposure rates and lowest measurement errors.

To evaluate the fidelity or conformity of each test event to the test blueprints and specifications, for both simulated data and real test events, information about the content composition of the adaptive tests delivered from each item pool is evaluated. During item selection, the algorithm attempts to meet all the specified criteria. Tables that summarize, for each criterion of the algorithm, both the mean number of items delivered and the proportion of times each criterion is not met are tabulated. These values are reported for both the simulated and real data. The simulated data provide a baseline for how we expect each item pool to perform. Weights can be imposed in the CAT algorithm that reflect the importance of a given test constraint. Violations of constraints with higher weights/importance would be considered more serious than violations of constraints with lower weights.

Measurement Errors for Simulated Scores; Both Overall and Claim Subscores. Test information functions, recovery of simulated examinee ability, and analysis of bias and error are all highly interrelated and can be addressed collectively. The definition of test efficiency hinges on the corresponding definition of test precision. Test precision is loosely defined through the standard error of measurement. All test scores include an error component, the size of which generally varies across test takers. Differences in precision across score ranges are ignored by measures of precision that, like test reliability, are aggregated across score levels. However, IRT provides a related pair of test precision measures that are specific to, or conditional on, score level. Both the test information function and the inversely related conditional standard error trace test-precision level across the score scale. (The conditional standard error function is the inverse of the square root of the test information function.) In a simulation environment, the score bias function measures the extent to which score estimates converge to their true values. The smaller the bias and error, the better the test administration and scoring procedures recover simulated examinee ability. Even if the goal is to measure each student according to some fixed criteria for test information/conditional standard error, test precision can vary not just across proficiency levels but across test takers at the same level of proficiency. However, test administration procedures may differ in the extent to which each test taker is measured on the targeted precision. It should be noted that exceeding the precision target is almost as undesirable as falling short. Measuring some test takers more precisely than necessary wastes resources (in the form of item exposures) that could be used more productively with other test takers.

The evaluation of how well the adaptive algorithm and item pool can recover simulated examinee ability can be presented by summarizing results for the 1000 test events at each theta. For example, summary statistics can be computed for every 1000 simulees with true overall scores and subscores at given intervals. Conditional means, 25th percentiles, 75th percentiles, conditional standard errors of measurement (CSEMs), and difference from target values can be reported for each theta interval. The conditional means and difference from target values will serve as indices of the ability of the algorithm and pool to recover the true abilities across the score range. The CSEM and 25th and 75th percentiles serve as a measure of variability in reported scores for each true score.

Analysis of Summative Pool Adequacy for Scores and Subscores. A number of statistics can be computed to evaluate each of the summative pools in a grade and content area. Any given pool should be a compilation of all relevant item types representing all subscores, with varying levels of difficulty and item information. All pools used for any given test should be randomly equivalent. To investigate this, the composition of each pool should be summarized by reporting the number of various item types separately for each of the subscore levels. In addition, summary statistics of the IRT difficulty and discrimination parameters can be calculated for each pool and each subscore level. These results can be compared across summative pools to see if all pools are similarly composed.

Expected and observed item exposure rates are reported, where item exposure rate is defined as the proportion of the total number of examinees who were administered a particular item. Item exposure is monitored for item and test security purposes to keep the same items from being administered to too many students and to keep pools viable by utilizing as many items as possible. In pools with little or no exposure control, it is possible that 10% of the items account for 70–80% of the items administered. The frequency, percent, and cumulative percentage of items in each pool with various exposure rates can be calculated. Simulated data can be used to obtain the expected rates; actual data can be used to obtain the observed rates. The correlation between expected and observed exposure rates, as well as summary statistics (mean, minimum, maximum, standard deviation) for exposure rates can also be included in this analysis. Overlap between simulated and adaptive test administration should also be examined. There will be less overlap with unconditional samples than samples conditioned on ability, so it is important to control (and monitor) exposure conditionally.

Simulations Studies for 2014-15 operational summative tests

Two sets of simulation studies were conducted for the 2014-15 tests using packaged pools with both the Consortium's proprietary engine and with CRESST's simulation engine, which serves as a baseline for other vendors. These results are published as part of the 2013-14 TECHNICAL REPORT. Simulation is an ongoing effort conducted for each adaptive test. Conventional reliability statistics are produced for the fixed form interim tests.

Test Sustainability. This broad and vitally important criterion is not always considered in adaptive-testing programs. Essentially, sustainability refers to the ease with which a testing program can be operationally maintained over time. At least three factors are important:

- What level of pretesting is needed to maintain summative bank stability? More sustainable testing programs will require less item development and pretesting to maintain summative bank size and quality at stable levels.
- How balanced is summative pool use? More sustainable testing programs will use items effectively by balancing use. With balanced item use, every item appears with roughly equal or uniform frequency. When item use is poorly balanced, a few items appear very often and a large number are rarely used. Unbalanced item use affects sustainability by making a small number of exceptional items carry much of the burden. These items risk becoming known to the test-taker community and so may be removed from use, either temporarily or permanently. However, a large number of new items must be pretested to find the few that are exceptional enough to replace those being released. Under a more balanced test design, items that are more commonplace would be used often enough to reach retirement. Fewer new items would need to be pretested to replace these more typical items.
- How easy are summative pools to develop? Test administration procedures or algorithms that facilitate summative pool development will be more easily sustained over time. Several

factors will influence the ease or difficulty of summative pool development, with some of these factors more easily quantified than other ones. One factor concerns the conditions that the pool must meet in order to be effective. Presumably, summative pools required to meet fewer and weaker conditions will be easier to develop. However, the extent to which pools parallel the structure of the summative bank is also important. Pools broadly representative of the summative bank will likely be easier to develop than pools that sample the bank more selectively. Finally, pools that operate in ways that are more predictable will be easier to develop than pools that function unpredictably. Minor changes to summative pools should result in equally minor changes in the way a pool functions.

Ideally, test sustainability would be evaluated by simulations that predict the effects of several years of operational test administration. This simulation would start with the item banks as they currently stand and then work through several years of operational testing. Summative and interim pools would be built, tests would be administered, item usage would be tracked, frequently administered items would be retired, and new items would be pretested and enter the item bank. Comparing the summative bank at the end of this cycle with that at the outset would reveal whether the test administration procedures and all their assumptions (item development requirements, pretest volumes, pool specifications, pool development, item retirement limits, etc.) are able to keep the item banks stable and the testing program sustainable.

Robustness to Aberrant Responding. Student test takers occasionally respond to test items in unexpected ways. Carelessness, low test-completion rates (speededness), item pre-exposure, unusual educational backgrounds, and a host of other factors are potential causes. Both conventional and adaptive tests are likely to poorly measure test takers who respond idiosyncratically. However, some adaptive administration and scoring procedures may cope better than other ones. A series of simulations can be conducted to evaluate the chosen procedures in this regard. Each simulation will be capable of generating data according to one of several identified nonstandard response models (these would simulate the effects of careless responding, speededness, lucky guessing) and other sources of anomalous responding. The evaluation will determine how successful the test administration and scoring procedures are in recovering true proficiency values despite the presence of unusual or aberrant responding. Although this is less of a concern for interim assessments, it is more visible to users such as teachers.

Test Design Specifications and Outcomes

Major types of assessment design specifications that did not necessarily occur sequentially are summarized below that fall generally under the rubric of test design. These steps primarily relate to content validity of the Smarter Balanced assessments, particularly with respect to nonstandard administrations. Further details can be obtained in Chapter 3 on item and test development. Other test specifications concern the establishment of achievement level descriptors and psychometric specifications that pertain to scaling and implications for scores. In many cases, the results were reviewed by one or more Stakeholder groups.

1) Conducted Initial Analysis of the Content and Structure of the CCSS

An initial analysis of how each standard within the CCSS could be assessed in terms of item/task type and DOK was conducted. This was intended to support content and curriculum specialists and test- and item/task-development experts. Analysis and recommendations were made for all ELA/literacy and mathematics standards in grades 3 to 8 and high school. Multiple levels of review were conducted that included the Smarter Balanced Technical Advisory Committee, Smarter Balanced member states, and Smarter Balanced Executive Committee.

2) Developed Content Specifications for ELA/literacy and Mathematics

Content specifications (e.g., claims, inferences, and evidence), item/task development criteria, and sample item/task sets were developed. This was intended to support the development of test blueprints and test specifications. Key constructs underlying each content area and critical standards/strands were identified in terms of demonstrating evidence of learning. Standards and bundled standards based on “bigger ideas” within the CCSS that require measurement through non-selected-response items (e.g., innovative item types) were identified. Reviews were conducted by CCSS authors, content experts, and assessment specialists.

3) Specified Accessibility and Accommodations Policy Guidelines

Guidelines that describe the accessibility and accommodations framework and related policies for test participation and administration were created that incorporated evidence-based design (ECD) principles and outcomes from small-scale trials. State survey and review of best practices were reviewed as well as recommendations on the use of assessment technology. Input was solicited from the Smarter Balanced English Language Learners Advisory Committee and the Students with Disabilities Advisory Committee.

4) Developed Item and Task Specifications

Smarter Balanced item/task type characteristics were defined as sufficient to ensure that content measured the intent of the CCSS and there was consistency across item/task writers and editors. This included all item types, such as selected-response, constructed-response, technology-enhanced, and performance tasks. In addition, passage/stimulus specifications (e.g., length, complexity, genre) and scoring rubric specifications for each item/task type were included. Specifications for developing items for special forms (e.g., braille) were also included.

5) Developed and Refined Test Specifications and Blueprints

The test form components (e.g., number of items/tasks, breadth and depth of content coverage) necessary to consistently build valid and reliable test forms that reflect emphasized CCSS content were defined. These specifications included purpose, use, and validity claims of each test, item/task, test form, and CAT attribute. These were reviewed and revised based on CAT simulation studies, small-scale trials, Pilot and Field testing, and as other information was made available.

6) Developed Initial Achievement Level Descriptors

Achievement expectations for mathematics and ELA/literacy were written in a manner that students, educators, and parents could understand. Panelists were recruited, and panels consisting of Institutes of Higher Education and a Cross-Consortia Technical Advisory Committee were convened in order to define college and career readiness. A period for public comment and various levels of review was implemented by the Smarter Balanced Technical Advisory Committee and selected focus groups with the approval of Governing States. These activities were coordinated with the PARCC consortium.

7) Developed Item and Task Prototypes

Prototype items and tasks using accessibility and Universal Design principles were produced that maximize fairness and minimize bias by using the principles of evidence-based design. Recommendations were made on how best to measure standards for innovative item types (per content specifications). This included prototypes for scoring guides, selected-response items, constructed-response items, and performance tasks. These prototypes were annotated, describing key features of items/tasks and scoring guides, passage/stimulus specifications (e.g., length, complexity, genre), and scoring rubric guidelines for each item/task type. Reviews, feedback, and revisions were obtained from educator-focus groups and Stakeholders, Smarter

Balanced work groups, the Smarter Balanced English Language Learners Advisory Committee, and the Students with Disabilities Advisory Committee.

8) Wrote Item and Performance Task Style Guide

The style guide specifies item/task formatting sufficient to ensure consistency of item/task formatting and display. The style guide specified the font, treatment of emphasized language/words (e.g., bold, italics), screen-display specifications, constraints on image size, resolution, colors, and passage/stimulus display configuration. Comprehensive guidelines for online and paper style requirements for all item types (e.g., selected-response, constructed-response, technology-enhanced, performance tasks) were specified.

9) Developed Accessibility Guidelines for Item and Task Development

Guidelines were produced for item and task writing/editing that ensure accessibility of test content that addressed all item types. Interoperability standards at the item and test level were determined. Reviews, feedback, and revisions were based on educator-focus groups, Smarter Balanced work groups, the Smarter Balanced English Language Learners Advisory Committee, and the Students with Disabilities Advisory Committee.

10) Developed and Distributed Item/Task Writing Training Materials

Training materials were created that specified consistent use of item/task specifications, style guides, accessibility guidelines, and best practices in item/task development (e.g., Universal Design, bias and sensitivity concerns) that were sufficient to ensure valid and reliable items/tasks that are free from bias and maximize accessibility to content. Training for item/task writing and editing was developed as online modules that enabled writers and editors to receive training remotely. Item writer and editor qualifications were established, and quality control procedures to ensure item writers were adequately trained were implemented.

11) Reviewed State-Submitted Items and Tasks for Inclusion in Smarter Balanced Item Pool

State-submitted items/tasks were reviewed for inclusion in the Pilot and/or Field Test item bank using the item bank/authoring system. This consisted of developing protocols for the submission and collection of state-submitted items/tasks for potential use in Pilot or Field Tests. These items were reviewed for item/task alignment, appropriateness (including access), and bias and sensitivity. Feedback was provided to states on the disposition of submitted items/tasks, and a gap analysis was conducted to determine the item/task procurement needs.

12) Planned and Conducted Small-Scale Trials of New Item and Task Types

Small-scale trials of new item/task types were used to inform potential revision of item/task specifications and style guides. Cognitive labs were conducted for new item/task types. Small-scale trials reflected an iterative development process, such that recommended revisions were evaluated as improvements became available.

13) Developed Automated-Scoring Approaches

The initial automated scoring methodology (e.g., regression, rules-based, or hybrid) was based on information from the content specifications, item/task specifications, item/task prototypes, and response data from the small-scale item/task trials. Reports documenting analysis were created, and independent review of this information with recommendations was made. Consultation, review, and approval of recommendations by the Smarter Balanced Technical Advisory Committee were made.

14) Developed Smarter Balanced Item and Task Writing Participation Policies and Guidelines

Documentation of processes for Smarter Balanced member states and Stakeholders to be involved in Smarter Balanced item/task writing activities (e.g., content and bias/sensitivity, data review, Pilot Testing, Field Testing) was developed. Criteria for selecting committee members (e.g., regional representation, expertise, experience) were also made.

15) Developed Content and Bias/Sensitivity Pilot Item and Task Review Materials

Methods for consistent training for content- and bias-review committees and for meeting logistics guidelines were provided. Review committees were recruited consistent with Smarter Balanced assessment participation policies.

16) Conducted Content and Bias/Sensitivity Reviews of Passages and Stimuli

Feedback from educators and other Stakeholders regarding passage/stimulus accuracy, alignment, appropriateness, accessibility, conformance to passage/stimulus specifications and style guides, and potential bias and sensitivity concerns was obtained. Educator feedback was documented, and procedures for feedback-reconciliation review were made.

17) Conducted Content and Bias/Sensitivity Pilot and Field Item and Task Review Meetings

Feedback from educators and other Stakeholders regarding item/task accuracy, alignment, appropriateness, accessibility, conformance to item/task specifications and style guides, and potential bias and sensitivity concerns was obtained. Reviews included all aspects of items/tasks (stem, answer choices, art, scoring rubrics) and statistical characteristics.

18) Developed Translation Framework and Specifications Languages

Definitions of item/task translation activities that ensure consistent and valid translation processes consistent with Smarter Balanced policy were produced. Review and approval of this process by the ELL Advisory Committee was made.

19) Translated Pilot and Field Test Items and Tasks into Identified Languages

Items/tasks translated into the specified languages were edited in sufficient quantity to support both Pilot- and Field-testing and operational assessments. Items/tasks included a full array of Smarter Balanced item types (selected-response, constructed-response, technology-enhanced, performance tasks). Review for content and bias/sensitivity of item/tasks and passages/stimuli was conducted.

20) Developed Content and Bias/Sensitivity Field Test Item and Task Review Materials

Supporting materials that ensure consistent training for content- and bias-review committees and meeting logistics guidelines were developed.

21) Revised Field Test Items and Tasks Based on Content and Bias/Sensitivity Committee Feedback

Fully revised items/tasks were available to be included on Field Test forms. Review panels were identified and convened, and training of state-level staff to edit and improve items/tasks that included all aspects of items/tasks (e.g., art, scoring rubrics) was conducted.

22) Developed Translation Framework and Specifications Languages

Definitions of item/task translation activities that ensured consistent and valid translation processes consistent with Smarter Balanced policy were created and approved by the ELL Advisory Committee.

23) Translated Pilot and Field Test Items and Tasks into Identified Languages

Translated items/tasks written by vendors, teachers, or provided through state submissions were edited in sufficient quantity to support Pilot and Field Tests and operational assessment.

24) Developed Content and Bias/Sensitivity Field Test Item and Task Review Materials

Review materials that ensure consistent training for content- and bias-review committees and meeting logistics guidelines were created. Feedback from educators and other Stakeholders regarding item/task accuracy, alignment, appropriateness, accessibility, conformance to item/task specifications and style guides, and potential bias and sensitivity concerns was obtained.

25) Produced a Single Composite Score Based on the CAT and Performance Tasks

A dimensionality study was conducted to determine whether a single score and composite score could be produced or if separate scales for the CAT and performance task components should be produced. Based on the Pilot Test, a dimensionality study was conducted and the results presented to the Smarter Balanced Technical Advisory Committee. A unidimensional model was chosen for the Smarter Balanced Field test.

26) Investigated Test Precision for the CAT Administrations

An investigation of targets was conducted for score precision in the case in which tests are constructed dynamically from a pool of items and a set of rules must be established for the adaptive algorithm. A number of supporting simulation studies were conducted. The findings were used to inform subsequent test design for the operational CAT that was presented to the Smarter Balanced Technical Advisory Committee.

27) Selected IRT Models for Scaling

Using the Pilot Test data, the characteristics of various IRT models for selected- and constructed-response items were compared. The results of this study were presented to the Validation and Psychometrics/Test Design Work Group and the Smarter Balanced Technical Advisory Committee for comment. The two-parameter logistic (2-PL) model for selected-response and the Generalized Partial Credit (GPC) Model for constructed-response were chosen as the scaling models.

References

- Abedi, J. (2010). *Performance Assessments for English Language Learners*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.
- American Institutes for Research (2014). *Smarter Balanced Scoring Specification: 2014–2015 Administration*.
- Center for Universal Design (CUD). (1997). About UD: Universal Design Principles. http://www.design.ncsu.edu/cud/about_ud/udprincipleshtmlformat.html (accessed February 13, 2009). Archived at <http://www.webcitation.org/5eZBa9RhJ>.
- Cohen, J. & Albright, L. (2014). *Smarter Balanced Adaptive Item Selection Algorithm Design Report*. May 9, 2014.
- Dana, T. M., & Tippins, D. J. (1993). Considering alternative assessment for middle level learners. *Middle School Journal*, 25, 3-5.
- DeMauro, G. E. (2004). Test Alignment Considerations for the Meaning of Testing. Paper presented at the CCSSO Annual Conference on Large Scale Assessment, Boston, MA.
- Fadel, C., Honey, M., & Pasnik, S. (2007, May). Assessment in the Age of Innovation. *Education Week*. May 18, 2007. Retrieved on July 2, 2012 from: <http://www.edweek.org/ew/articles/2007/05/23/38fadel.h26.html?print=1>.
- Folk, V. G. & Smith, R. L. (2002). Models for Delivery of CBTs. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.). *Computer-Based Testing: Building the Foundation for Future Assessments* (pp. 41-66). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Gibson, W. M. & Weiner, J. A. (1998). Generating Random Parallel Test Forms Using CTT in a Computer-Based Environment. *Journal of Educational Measurement*, 35, 297-310.
- Hetter, R. D. & Simpson, J. B. (1997). Item Exposure Control in CAT-ASVAB. In W.A. Sands, B. K. Waters, & J. R. McBride (Eds.). *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 141-144). Washington, DC: American Psychological Association.
- HumRRO (2014). *Smarter Balanced Assessment Consortium Alignment Study Report*. December 30, 2014.
- Luecht, R. M. (1998). Computer-Assisted Test Assembly Using Optimization Heuristics. *Applied Psychological Measurement*, 22, 224-236.
- Kane, M., & Mitchell, R. (1996). *Implementing Performance Assessment: Promises, Problems, and Challenges*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Martone, A., & Sireci, S. G. (2009). Evaluating Alignment Between Curriculum, Assessment, and Instruction. *Review of Educational Research*, 79, 1-76.
- Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Reckase, M. D. (2003). Item pool design for computerized adaptive tests. Paper presented at the National Council on Measurement in Education, Chicago, IL.
- Rose, D., & Meyer, A. (2000). Universal design for learning, associate editor column. *Journal of Special Education Technology* 15 (1): 66-67.
- Schmeiser, C. B., & Welch, C. J. (2006). Test Development. In R. L. Brennan (Ed.) *Educational Measurement*, 4th Edition (307-353). Washington, DC: American Council on Education.

- van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*. New York: Springer.
- Webb, N. L. (1997a, April). Research Monograph No. 6. *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*. Washington, DC: Council of Chief State School Officers.
- Webb, N. L. (March 28, 2002) *Depth-of-Knowledge Levels for Four Content Areas*, unpublished paper.
- Zhang, T., Haertel, G., Javitz, H., Mislevy, R., Murray, E., & Wasson, J. (2009). *A Design Pattern for a Spelling Bee Assessment for Students with Disabilities*. A paper presented at the annual conference of the American Psychological Association, Montreal, Canada.

Chapter 5 Test Fairness

Introduction

For large-scale programs such as those developed by the Smarter Balanced Assessment Consortium (Smarter Balanced), an essential goal is to ensure that all students have comparable opportunities to demonstrate their achievement level. Smarter Balanced strives to provide every student with a positive and productive assessment experience and results that are a fair and accurate depiction of each student's achievement. Ensuring test fairness is a fundamental part of validity, starting with test design, and is an important feature built into each step of the test development process, such as item writing, test administration, and scoring. The 2014 *Standards for Educational and Psychological Testing* state, "The term fairness has no single technical meaning, and is used in many ways in public discourse." It also suggests that fairness to all individuals in the intended population is an overriding and fundamental validity concern.

The Smarter Balanced system is designed to provide a valid, reliable, and fair measure of student achievement based on the Common Core State Standards (CCSS). The validity and fairness of the measures of student achievement are influenced by a multitude of factors; central among them are:

- a clear definition of the construct—the knowledge, skills, and abilities—that are intended to be measured,
- the development of items and tasks that are explicitly designed to assess the construct that is the target of measurement,
- delivery of items and tasks that enable students to demonstrate their achievement of the construct and,
- capture and scoring of responses to those items and tasks.

Smarter Balanced uses several documents to address reliability, validity, and fairness. The *Common Core State Standards* were originally developed by the National Governors Association (NGA) and the Council Chief State School Officers (CCSSO). The *Smarter Balanced Content Specifications*, developed by the Consortium, articulate the claims and targets of the Smarter Balanced assessments, defining the knowledge, skills, and abilities to be assessed and their relationship to the CCSS. In doing so, these documents describe the major constructs—identified as "Claims"—within English language arts/literacy (ELA/literacy) and mathematics for which evidence of student achievement will be gathered and which will form the basis for reporting student performance. Much of the evidence presented in this chapter pertains to fairness in treatment during the testing process and lack of measurement bias (i.e., DIF). Fairness (minimizing bias) and the design of accessibility supports (i.e., universal tools, designated supports and accommodations in content development) is addressed in Chapter 3.

Definitions for Validity, Bias, Sensitivity, and Fairness. Some key concepts for the ensuing discussion concern validity, bias, and fairness and are described as follows.

Validity. Validity is the extent to which the inferences and actions made based on test scores are appropriate and backed by evidence (Messick, 1989). It constitutes the central notion underlying the development, administration and scoring of a test, as well as the uses and interpretations of test scores. Validation is the process of accumulating evidence to support each proposed score interpretation or use. Evidence in support of validity is extensively discussed in Chapter 2.

Bias and sensitivity. According to the Standards for Educational and Psychological Testing, “Bias in tests and testing refers to construct-irrelevant [i.e., invalid] components that result in systematically lower or higher scores for identifiable groups of examinees” (*Standards*, 1999 (AERA, APA, & NCME, p. 76; *Standards*, (AERA, APA, & NCME, 2014, 51-54). “Sensitivity” is used to refer to an awareness of the need to avoid bias in assessment. In common usage, reviews of tests for bias and sensitivity are reviews to help ensure that the test items and stimuli are fair for various groups of test takers, (*Standards*, 2014 (AERA, APA, & NCME, 2014, p. 64).

The goal of fairness in assessment can be approached by ensuring that test materials are as free as possible of unnecessary barriers to the success of a diverse group of students. Smarter Balanced developed *Bias and Sensitivity Guidelines* to help ensure that the assessments are fair for all groups of test takers, despite differences in characteristics including, but not limited to, disability status, ethnic group, gender, regional background, native language, race, religion, sexual orientation, and socioeconomic status. Unnecessary barriers can be reduced by following some fundamental rules (ETS, 2012):

- not measuring irrelevant knowledge or skills (i.e., construct irrelevant),
- not angering, offending, upsetting, or otherwise distracting test takers, and
- treating all groups of people with appropriate respect in test materials.

These rules help ensure that the test content is fair for test takers as well as acceptable to the many stakeholders and constituent groups within the Smarter Balanced states. The more typical view is that bias and sensitivity guidelines apply primarily to the review of test items. However, fairness must be considered in all phases of test development and use. Smarter Balanced strongly relied on the *Bias and Sensitivity Guidelines* in the development of the Smarter Balanced assessments, particularly in item writing and review. Items had to comply with the *bias and sensitivity Guidelines* in order to be included in the Smarter Balanced assessments. Use of the *Guidelines* will help the Smarter Balanced assessments comply with Chapter 3, Standard 3.2 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p. 82). Standard 3.2 states that “Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests’ being affected by construct-irrelevant characteristics such as linguistic, communicative, cognitive, cultural, physical or other characteristics”. The Smarter Balanced assessments were developed using the principles of evidence-centered design (ECD). Three basic elements of ECD (Mislevy, Steinberg, & Almond, 1999) are stating the claims to be made about test takers, deciding what evidence is required to support these claims, and administering test items that provide the required evidence. ECD provides a chain of evidence-based reasoning that links test performance to the Claims to be made about test takers. Fair assessments are essential to the implementation of ECD. If the items are not fair, then the evidence they provide means different things for the various groups of students. Under those circumstances, the Claims cannot be equally supported for all test takers, which is a threat to validity. Appropriate use of the *Bias and Sensitivity Guidelines* helps to

ensure that the evidence provided by the items allows ECD to function as intended and is equally valid for various groups of test takers.

Fairness. “Fairness” as mentioned previously is a difficult word to define because, as indicated in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, p. 74), “The central idea of fairness in testing is to identify and remove construct-irrelevant barriers to maximal performance for any examinee.” An extensive discussion of the meanings of “fairness” in assessment was also given by Camilli (2006). A useful definition of fairness for the purposes of the *Bias and Sensitivity Guidelines* is the extent to which the test scores are valid for different groups of test takers. For example, a mathematics item may contain difficult language unrelated to mathematics. If the language interfered about equally with all test takers, validity would be negatively impacted for all test takers. If, however, the language were a more significant barrier for students who are not native speakers of English, compared with other students, then the item would be unfair. If items are more difficult for some groups of students than for other groups of students, the items may not necessarily be unfair. For example, if an item were intended to measure the ability to comprehend a reading passage in English, score differences between groups based on real differences in comprehension of English would be valid and, therefore, fair. As Cole and Zieky (2001, p. 375) noted, “If the members of the measurement community currently agree on any aspect of fairness, it is that score differences alone are not proof of bias.” Fairness does not require that all groups have the same average scores. Fairness requires any existing differences in scores to be valid. An item would be unfair if the source of the difficulty were not a valid aspect of the item. For example, an item would be unfair if members of a group of test takers were distracted by an aspect of the item that they found highly offensive. If the difference in difficulty reflected real and relevant differences in the group’s level of mastery of the tested CCSS, the item could be considered as fair.

The Smarter Balanced Accessibility and Accommodations Framework

Smarter Balanced has built a framework of accessibility for all students, including English Language Learners (ELLs), students with disabilities, and ELLs with disabilities, but not limited to those groups. Three additional sources—the Smarter Balanced *Item Specifications*, the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* and the Smarter Balanced *Bias and Sensitivity Guidelines*—are used to guide the development of items and tasks to ensure that they accurately measure the targeted constructs. Recognizing the diverse characteristics and needs of students who participate in the Smarter Balanced assessments, the states worked together through the Smarter Balanced Test Administration and Student Access Work Group to develop an Accessibility and Accommodations Framework that guided the Consortium as it worked to reach agreement on the specific universal tools, designated supports, and accommodations available for the assessments. This work also incorporated research and practical lessons learned through Universal Design, accessibility tools, and accommodations (Thompson, Johnstone, & Thurlow, 2002). Much of the conceptualization for this chapter is a direct reflection of the outcomes from the work of the Test Administration and Student Access Work Group.

In the process of developing its next-generation assessments to measure students’ knowledge and skills as they progress toward college and career readiness, Smarter Balanced recognized that the validity of assessment results depends on each student having appropriate universal tools, designated supports, and/or accommodations when needed, based on the constructs being measured by the assessment. The

Smarter Balanced Assessment System utilizes technology intended to deliver assessments that meet the needs of individual students. Online/electronic delivery of the assessments helps ensure that students are administered a test individualized to meet their needs consistent with their peers. Items and tasks were delivered using a variety of accessibility resources and accommodations that can be administered to students automatically based on their individual profiles. Accessibility resources include but are not limited to foreground and background color flexibility, tactile presentation of content (e.g., Braille), and translated presentation of assessment content in signed form and selected spoken languages.

A principle for Smarter Balanced was to adopt a common set of accessibility resources and accommodations. Moreover, the Notification Inviting Applications (NIA) posted in the Federal Register, April 9, 2010, required “a common set of policies and procedures for accommodations” for any consortia funded by the USED Race to the Top Assessment Program; the following definition was used.

Accommodations means changes in the administration of an assessment, including but not limited to changes in assessment setting, scheduling, timing, presentation format, response mode, and combinations of these changes that do not change the construct intended to be measured by the assessment or the meaning of the resulting scores. Accommodations must be used for equity in assessment and not provide advantage to students eligible to receive them.

The focus is on “equity in assessment” and does not refer to specific student characteristics, a perspective that is consistent with the Accessibility and Accommodations Framework. A fundamental goal was to design an assessment that is accessible for all students, regardless of English language proficiency, disability, or other individual circumstances. The three components of the *Accessibility and Accommodations Framework* are designed to meet that need. The intent was to ensure that the following steps were achieved for Smarter Balanced.

- Design and develop items and tasks to ensure that all students have access to the items and tasks designed to measure the targeted constructs. In addition, deliver items, tasks, and the collection of student responses in a way that maximizes validity for each student.
- Adopt the conceptual model embodied in the Accessibility and Accommodations Framework that describes accessibility resources of digitally delivered items/tasks and acknowledges the need for some adult-monitored accommodations. The model also characterizes accessibility resource as a continuum from those available to all students ranging to ones that are implemented under adult supervision available only to those students with a documented need.
- Implement the use of an individualized and systematic needs profile, or Individual Student Assessment Accessibility Profile (ISAAP), for students that promotes the provision of appropriate access and tools for each student. Smarter created an ISAAP process that helps education teams systematically select the most appropriate accessibility resources for each student and ISAAP tool, which helps teams note the accessibility resources chosen.

The conceptual framework that serves as the basis underlying the usability, accessibility, and accommodations is shown in Figure 1. This figure portrays several aspects of the Smarter Balanced assessment features—universal tools (available for all students), designated supports (available when

indicated by an adult or team), and accommodations as documented in an Individualized Education Program (IEP) or 504 plan. It also displays the additive and sequentially inclusive nature of these three aspects. Universal tools are available to all students, including those receiving designated supports and those receiving accommodations. Designated supports are available only to students who have been identified as needing these accommodations (as well as those students for whom the need is documented). Accommodations are available only to those students with documentation of the need through a formal plan (e.g., IEP). Those students also may access designated supports and universal tools.

A universal tool for a content focus in a specific may be an accommodation for another grade or content focus. Similarly, a designated support may also be an accommodation, depending on the content target and grade. This approach is consistent with the emphasis that Smarter Balanced has placed on the validity of assessment results coupled with access. Universal tools, designated supports, and accommodations are all intended to yield valid scores. Universal tools, designated supports, and accommodations result in scores that count toward participation in statewide assessments. Also shown in Figure 1 are the universal tools, designated supports, and accommodations for each category of accessibility resources. There are both embedded and non-embedded versions of the universal tools, designated supports, or accommodations depending on whether they are provided as digitally delivered components of the test administration or separate from test delivery.

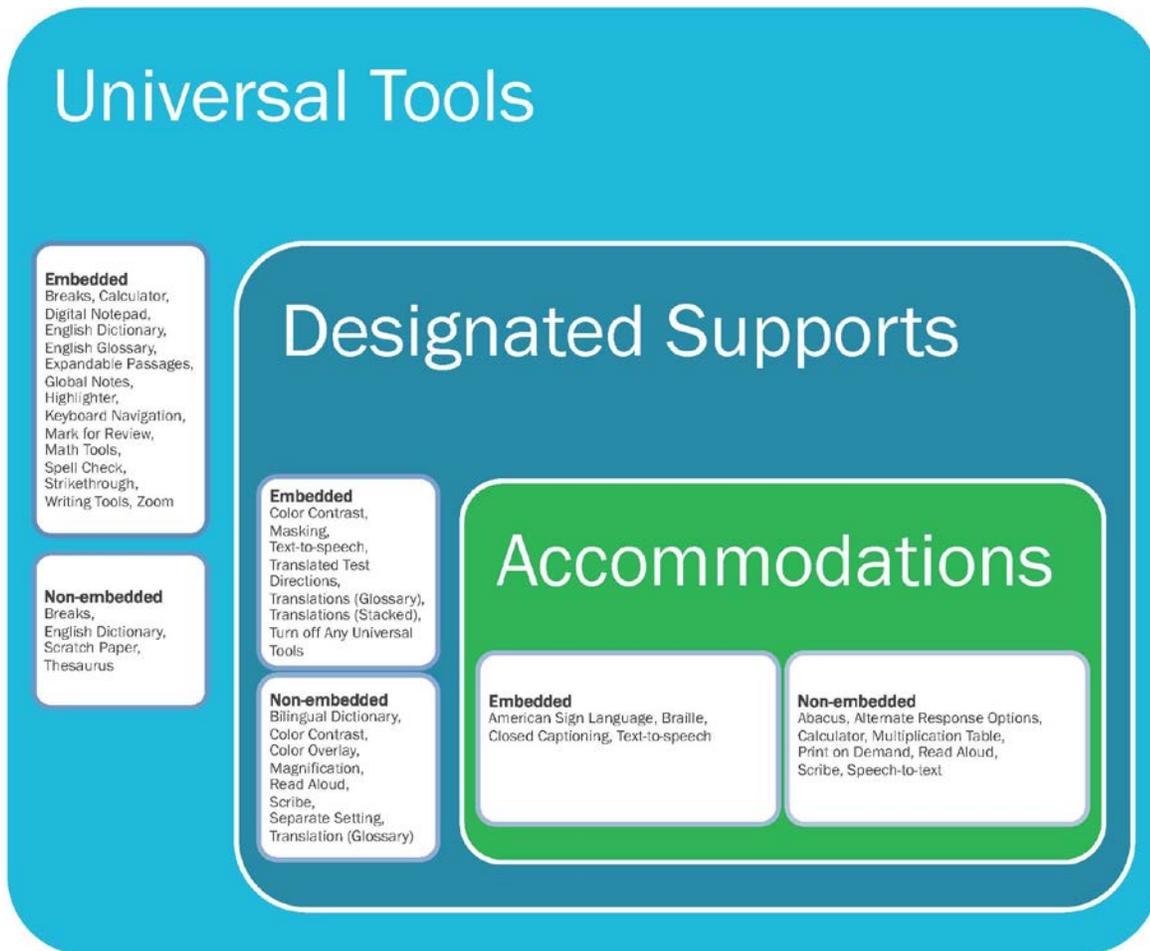


Figure 1. Conceptual Framework for Usability, Accessibility, and Accommodations.

Meeting the Needs of Traditionally Underrepresented Populations. The policy decision was to make accessibility resources available to all students based on need rather than eligibility status or student subgroup categorical designation. This reflects a belief among Consortium states that unnecessarily restricting access to accessibility resources threatens the validity of the assessment results and places students under undue stress and frustration. Additionally, accommodations are available for students who qualify for them. Although the intention of this policy is to ensure a positive and productive assessment experience for all students, elimination of specific eligibility criteria may raise concerns among some educators and advocates who worked to create eligibility criteria that guarantee appropriate assessment supports for their students of interest. Discussion on how a needs-based approach will benefit ELLs, students with disabilities, and ELLs with disabilities is presented here.

How the Framework Meets Needs of Students Who Are ELLs. Students who are ELLs have needs that are unique from those students with disabilities, including language-related disabilities. The needs of ELLs are not the result of a language-related disability, but instead are specific to the student's current level of English language proficiency. The needs of students who are ELLs are diverse and are influenced by the interaction of several factors, including their current level of English language proficiency, their prior exposure to academic content and language in their native language, the languages to which they are exposed outside of school, the length of time they have participated in the U.S. education system, and the language(s) in which academic content is presented in the classroom. Given the unique background and needs of each student, the conceptual framework is designed to focus on students as individuals and to provide several accessibility resources that can be combined in a variety of ways. Some of these digital tools, such as using a highlighter to highlight key information and an audio presentation of test navigation features, are available to all students, including those at various stages of English language development. Other tools, such as the audio presentation of items and glossary definitions in English, may also be assigned to any student, including those at various stages of English language development. Still other tools, such as embedded glossaries that present translation of construct irrelevant terms, are intended for those students whose prior language experiences would allow them to benefit from translations into another spoken language. Collectively, the conceptual framework for usability, accessibility, and accommodations embraces a variety of accessibility resources that have been designed to meet the needs of students at various stages in their English language development.

How the Framework Meets Needs of Students with Disabilities. Federal law requires that students with disabilities who have a documented need receive accommodations that address those needs, and that they participate in assessments. The intent of the law is to ensure that all students have appropriate access to instructional materials and are held to the same high standards. When students are assessed, the law ensures that students receive appropriate accommodations during testing so they can appropriately demonstrate what they know and can do so that their achievement is measured accurately.

The Accessibility and Accommodations Framework addresses the needs of students with disabilities in three ways. First, it provides for the use of digital test items that are purposefully designed to contain multiple forms of the item, each developed to address a specific access need. By allowing the delivery of a given access form of an item to be tailored based on each student's access need, the Framework fulfills the intent of Federal accommodation legislation. Embedding universal accessibility digital tools, however, addresses only a portion of the access needs required by many students with disabilities. Second, by embedding accessibility resources in the digital test delivery system, additional access needs are met. This approach fulfills the intent of the law for many, but not all, students with disabilities, by allowing the accessibility resources to be activated for students based on their needs. Third, by allowing for a wide variety of digital and locally provided accommodations (including physical arrangements), the Framework addresses a spectrum of accessibility resources appropriate for math and ELA assessment. Collectively, the Framework adheres to Federal regulations by allowing a combination of universal design principles, universal tools, designated supports and accommodations to be embedded in a digital delivery system and through local administration assigned and provided based on individual student needs.

The Individual Student Assessment Accessibility Profile (ISAAP). Typical practice frequently required schools and educators to document, a priori, the need for specific student accommodations and then to document

the use of those accommodations after the assessment. For example, most programs require schools to document a student's need for a large-print version of a test for delivery to the school. Following the test administration, the school documented (often by bubbling in information on an answer sheet) which of the accommodations, if any, a given student received, whether the student actually used the large-print form, and whether any other accommodations, such as extended time, were provided. Traditionally, many programs have focused only on those students who have received accommodations and thus may consider an accommodation report as documenting accessibility needs. The documentation of need and use establishes a student's accessibility needs for assessment.

For most students, universal digital tools will be available by default in the Smarter Balanced test delivery system and need not be documented. These tools can be deactivated if they create an unnecessary distraction for the student. Other embedded accessibility resources that are available for any student needing them must be documented prior to assessment. Smarter Balanced intends to obtain information on individual student test administration conditions for students with specific accessibility needs not addressed in the *Usability, Accessibility, and Accommodations Guidelines*. To capture specific student accessibility needs, the Smarter Balanced Assessment System has established an individual student assessment accessibility profile (ISAAP). The ISAAP tool is designed to facilitate selection of the universal tools, designated supports and accommodations that match student access needs for the Smarter Balanced assessments, as supported by the *Smarter Balanced Usability, Accessibility, and Accommodations Guidelines*. The ISAAP Tool should be used in conjunction with the *Smarter Balanced Usability, Accessibility and Accommodations Guidelines* and state regulations and policies related to assessment accessibility as a part of the ISAAP process. For students requiring one or more accessibility resource, schools will be able to document this need prior to test administration. Furthermore, the ISAAP can include information about universal tools that may need to be eliminated for a given student. By documenting need prior to test administration, a digital delivery system will be able to activate the specified options when the student logs in to an assessment. In this way, the profile permits educators and schools to focus on each individual student, documenting the accessibility resources required for valid assessment of that student in a way that is efficient to manage.

The conceptual framework (Figure 1) provides a structure that assists in identifying which accessibility resources should be made available for each students. In addition, the conceptual framework is designed to differentiate between universal tools available to all students and accessibility resources that must be assigned before the administration of the assessment. Consistent with recommendations from Shafer and Rivera (2011), Thurlow, Quenemoen, and Lazarus (2011), Fedorchak (2012), and Russell (2011b), Smarter Balanced is encouraging schools to use a team approach to make decisions concerning each student's ISAAP. Gaining input from individuals with multiple perspectives, including the student, will likely result in appropriate decisions about the assignment of accessibility resources. Consistent with these recommendations avoidance of selecting too many accessibility resources for a student. The use of too many unneeded accessibility resources can decrease student performance.

The team approach encouraged by Smarter Balanced does not require the formation of a new decision-making team, and the structure of teams can vary widely depending on the background and needs of a student. A locally convened student support team can potentially create the ISAAP. For most students who do not require accessibility tools or accommodations, an initial decision by a teacher may be confirmed by a

second person (potentially the student). In contrast, for a student who is an English language learner and has been identified with one or more disabilities, the IEP team should include the English language development specialist who works with the student, along with other required IEP team members and the student, as appropriate. The composition of teams is not being defined by Smarter Balanced; it is under the control of each school and is subject to state and Federal requirements.

Usability, Accessibility, and Accommodations Guidelines: Intended Audience and Recommended Applications. The Smarter Balanced Consortium has developed *Usability, Accessibility, and Accommodations Guidelines* (UUAG) that are intended for school-level personnel and decision-making teams, particularly Individualized Education Program (IEP) teams, as they prepare for and implement the Smarter Balanced assessment. The UUAG provide information for classroom teachers, English development educators, special education teachers, and related services personnel to use in selecting and administering universal tools, designated supports, and accommodations for those students who need them. The UUAG are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment. The Smarter Balanced *Usability*, UUAG emphasize an individualized approach to the implementation of assessment practices for those students who have diverse needs and participate in large-scale content assessments. This document focuses on universal tools, designated supports, and accommodations for the Smarter Balanced content assessments of ELA/literacy and mathematics. At the same time, it supports important instructional decisions about accessibility for students who participate in the Smarter Balanced assessments. It recognizes the critical connection between accessibility in instruction and accessibility during assessment. The UUAG are also incorporated into the Smarter Balanced Test Administration Manual.

All students (including students with disabilities, ELLs, and ELLs with disabilities) are to be held to the same expectations for participation and performance on state assessments. Specifically, all students enrolled in grades 3 to 8 and 11 are required to participate in the Smarter Balanced mathematics except students with the most significant cognitive disabilities who meet the criteria for the mathematics alternate assessment based on alternate achievement standards (approximately 1% or less of the student population).

All students enrolled in grades 3 to 8 and 11 are required to participate in the Smarter Balanced English language/literacy assessment except:

- students with the most significant cognitive disabilities who meet the criteria for the English language/literacy alternate assessment based on alternate achievement standards (approximately 1% or fewer of the student population), and
- ELLs who are enrolled for the first year in a U.S. school. These students will participate in their state's English language proficiency assessment.

Federal laws governing student participation in statewide assessments include the Elementary and Secondary Education Act (ESEA)—reauthorized as the No Child Left Behind Act (NCLB) of 2001, the Individuals with Disabilities Education Improvement Act of 2004 (IDEA), and Section 504 of the Rehabilitation Act of 1973 (reauthorized in 2008).

Since the Smarter Balanced assessment is based on the CCSS, the universal tools, designated supports, and accommodations that are appropriate for the Smarter Balanced assessment may be different from those that state programs utilized previously. For the summative assessments, state participants can only make available to students the universal tools, designated supports, and accommodations consistent with the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines*. When the implementation or use of the universal tool, designated support, or accommodation is in conflict with a member state's law, regulation, or policy, a state may elect not to make it available to students.

The Smarter Balanced universal tools, designated supports, and accommodations currently available for the Smarter Balanced assessments have been prescribed. The specific universal tools, designated supports, and accommodations approved by Smarter Balanced may undergo change if additional tools, supports, or accommodations are identified for the assessment based on state experience or research findings. The Consortium has established a standing committee, including members from Consortium members and staff, that reviews suggested additional universal tools, designated supports, and accommodations to determine if changes are warranted. Proposed changes to the list of universal tools, designated supports, and accommodations are brought to consortium members for review, input, and vote for approval. Furthermore, states may issue temporary approvals (i.e., one summative assessment administration) for individual, unique student accommodations. It is expected that states will evaluate formal requests for unique accommodations and determine whether the request poses a threat to the measurement of the construct. Upon issuing temporary approval, the petitioning state can send documentation of the approval to the Consortium. The Consortium will consider all state-approved temporary accommodations as part of the annual Consortium accommodations review process. The Consortium will provide to member states a list of the temporary accommodations issued by states that are not Consortium-approved accommodations.

Guidelines for Accessibility for English Language Learners. In addition to the use of Universal Design features, Smarter Balanced has built a framework of accessibility for all students, including English Language Learners (ELLs) that were established in the *Smarter Balanced Guidelines for Accessibility for English Language Learners*. ELLs have not yet acquired complete proficiency in English. For ELLs, the most significant accessibility issue concerns the nature of the language used in the assessments. The use of language that is not fully accessible can be regarded as a source of invalidity that affects the resulting test score interpretations by introducing construct-irrelevant variance. Although there are many validity issues related to the assessment of ELLs, the main threat to validity when assessing content knowledge stems from language factors that are not relevant to the construct of interest. The goal of these ELL guidelines was to minimize factors that are thought to contribute to such construct-irrelevant variance. Adherence to these guidelines helped ensure that, to the greatest extent possible, the Smarter Balanced assessments administered to ELLs measure the intended targets. The ELL *Guidelines* were intended primarily to inform Smarter Balanced assessment developers or other educational practitioners, including content specialists and testing coordinators.

For assessments, an important distinction is between content-related language that is the target of instruction versus language that is not content-related. For example, the use of words with specific technical meaning, such as “slope” when used in algebra or “population” when used in biology, should be used to assess content knowledge for all students. In contrast, greater caution should be exercised when including words that are not directly related to the domain. ELLs may have had cultural and social experiences that

differ from those of other students. Caution should be exercised in assuming that ELLs have the same degree of familiarity with concepts or objects occurring in situational contexts. The recommendation was to use contexts or objects based on classroom or school experiences rather than ones that are based outside of school. For example, in constructing mathematics items, it is preferable to use common school objects, such as books and pencils, rather than objects in the home, such as kitchen appliances, to reduce the potential for construct-irrelevant variance associated with a test item. When the construct of interest includes a language component, the decisions regarding the proper use of language becomes more nuanced. If the construct assessed is the ability to explain a mathematical concept, then the decisions depend on how the construct is defined. If the construct includes the use of specific language skills, such as the ability to explain a concept in an innovative context, then it is appropriate to assess these skills. In ELA\literacy, there is greater uncertainty as to item development approaches that faithfully reflect the construct while avoiding language inaccessible for ELLs. The decisions of what best constitutes an item can rely on the content standards, definition of the construct, and the interpretation of the claims and assessment targets. For example, if interpreting the meanings in a literary text is the skill assessed, then using the original source materials is acceptable. However, the test item itself—as distinct from the passage or stimulus—should be written so that the task presented to a student is clearly defined using accessible language. Since ELLs taking Smarter Balanced content assessments likely have a range of English proficiency skills, it is also important to consider the accessibility needs across the entire spectrum of proficiency. Since ELLs by definition have not attained complete proficiency in English, the major consideration in developing items is ensuring that the language used is as accessible as possible. The use of accessible language does not guarantee that construct-irrelevant variance will be eliminated, but it is the best strategy for helping ensure valid scores for ELLs and for other students as well.

Using clear and accessible language is a key strategy that minimizes construct-irrelevant variance in items. Language that is part of the construct being measured should not be simplified. For non-content-specific text, the language of presentation should be as clear and as simple as is practical. The following guidelines for the use of accessible language were proposed as guidance in the development of test items. This guidance was not intended to violate other principles of good item construction. From the *ELL Guidelines*, some general principles for the use of accessible language were proposed as follows.

- Design test directions to maximize clarity and ones that minimize the potential for confusion.
- Use vocabulary widely accessible to all students, and avoid unfamiliar vocabulary not directly related to the construct (August, Carlo, & Snow, 2005; Bailey, Huang, Shin, Farnsworth, & Butler, 2007).
- Avoid the use of syntax or vocabulary that is above the test's target grade level (Borgioli, 2008). The test item should be written at a vocabulary level no higher than the target grade level, and preferably at a slightly lower grade level, to ensure that all students understand the task presented (Young, 2008).
- Keep sentence structures as simple as is possible while expressing the intended meaning. In general, ELLs find a series of simpler, shorter sentences to be more accessible than longer, more complex sentences (Pitoniak, Young, Martiniello, King, Buteux, & Ginsburgh, 2009).
- Consider the impact of cognates (words with a common etymological origin) when developing items and false cognates. These are word pairs or phrases that appear to have the same meaning in two or more languages, but do not. Spanish and English share many cognates, and because the large majority of ELLs speak Spanish as their first language (nationally, more than 75%), the presence of

cognates can inadvertently confuse students and alter the skills being assessed by an item.

Examples of false cognates include: billion (the correct Spanish word is mil millones; not billón, which means *trillion*); deception (engaño; not decepción, which means disappointment); large (grande; not largo, which means long); library (biblioteca; not librería, which means bookstore).

- Do not use cultural references or idiomatic expressions (such as “being on the ball”) that are not equally familiar to all students (Bernhardt, 2005).
- Avoid sentence structures that may be confusing or difficult to follow, such as the use of passive voice or sentences with multiple clauses (Abedi & Lord, 2001; Forster & Olbrei, 1973; Schachter, 1983).
- Do not use syntax that may be confusing or ambiguous, such as using negation or double negatives in constructing test items (Abedi, 2006; Cummins, Kintsch, Reusser, & Weimer, 1988).
- Minimize the use of low-frequency, long, or morphologically complex words and long sentences (Abedi, 2006; Abedi, Lord & Plummer, 1995).
- Teachers can use multiple semiotic representations to convey meaning to students in their classrooms. Assessment developers should also consider ways to create questions using multi-semiotic methods so that students can better understand what is being asked (Kopriva, 2010). This might include greater use of graphical, schematic, or other visual representations to supplement information provided in written form.

Fairness as a Lack of Measurement Bias: Differential Item Functioning Analyses

As part of the validity evidence from internal structure, differential item functioning (DIF) analyses were conducted on the Field Test items. This section presents the evidence to support the frameworks’ claims. Chapters 6 and 8 presents the DIF methodology used and results for the Pilot- and Field Test phases. DIF analyses are used to identify those items for which identifiable groups of students (e.g., males, females) with the same underlying level of ability have different probabilities of answering an item correctly or obtaining a given score level. Students are separated into relevant subgroups based on ethnicity, gender, or other demographic characteristics for DIF analyses. Students in each subgroup are then ranked relative to their total test score (conditioning on ability). Students in the focal group (e.g., females) are then compared to students in the reference group (e.g., males) relative to their performance on individual items. It is part of the Smarter Balanced framework to have ongoing study and review of findings to inform iterative, data-driven decisions. These efforts are to ensure that items are not differentially difficult for any group of students.

Test Fairness and Implications for Ongoing Research

There are many features of the Smarter Balanced assessments that support equitable assessment across all groups of students. The assessments are developed using the principles of evidence-centered design and universal test design. Test accommodations are provided for students with disabilities, and language-tools and supports were developed for ELLs. The Work Group for Accessibility and Accommodations and the Consortium developed a set of guidelines to facilitate accessibility to the assessments. In addition to these general accessibility guidelines embedded in the conceptual framework, procedures for item writing and

reviewing and guidelines for creating audio, sign language, and tactile versions of the items were implemented. Smarter Balanced developed guidelines for item development that aim toward reducing construct-irrelevant language complexities for English language learners (Young, Pitoniak, King, & Ayad, 2012) and comprehensive guidelines for bias and sensitivity (ETS, 2009), and a rubric specifically geared towards scoring language complexity (Cook & MacDonald, 2013). In addition, measurement bias was investigated using DIF methods. This evidence underscores the commitment to fair and equitable assessment for all students, regardless of their gender, cultural heritage, disability status, native language, and other characteristics. Irrespective of these proactive development activities designed to promote equitable assessments, further validity evidence that the assessments are fair for all groups of students should be provided. Many of the equity issues are delineated in the most recent version of the NCLB *Peer Review Guidance* (U.S. Department of Education, 2009). To evaluate the degree to which the Smarter Balanced assessments are fulfilling the purpose of valid, reliable, and fair information that is equitable for all students, several types of additional evidence are recommended based on the relevant types listed in the AERA et al. (2014) *Standards*. Validity studies are described here as well as ones that can be addressed in the ongoing research agenda for Smarter Balanced.

Internal Structure. When evaluating the comparability of different variations of a test, such as different language glossaries or accommodated test administrations, validity evidence based on internal structure is the most common approach (Sireci, Han, & Wells, 2008). These studies most often involve multigroup factor analysis (Ercikan & Koh, 2005) or weighted (multigroup) multidimensional scaling, which has also been used for this purpose (see Chapter 5 Pilot Test; Robin, Sireci, & Hambleton, 2003; Sireci & Wells, 2010). Another important source of validity evidence to support equitable assessment is analysis of differential item functioning across test variations and across subgroups of students using differential bundle functioning (Banks, 2013). DIF studies conducted for Smarter Balanced used several criteria to distinguish statistically significant DIF from substantively meaningful DIF (i.e., reflects construct-irrelevant variance). The presence of DIF does not necessarily indicate bias, and therefore, As part of the data review process described in Chapter 3 DIF studies were followed by qualitative analyses that sought to interpret sources of DIF.

Response Processes. Validity evidence based on the relevant subgroups of students were addressed to examine the amount of time it takes different groups of students to respond to items (i.e., item response time) with and without accommodations. Cognitive interviews or think-aloud protocols should be conducted to evaluate the skills measured by items. In addition, specific studies are needed to evaluate accommodations for ELLs or students with disabilities and should be conducted to determine whether the students are using the accommodations and finding them helpful (Duncan, Parant, Chen, Ferrara, Johnson, Oppler, Shieh, 2005).

Relationships with Other Variables. Two types of evidence based on relations to other variables are relevant for validating that the Smarter Balanced assessments are equitable for all subgroups of students (Dorans, 2004). The first type is differential predictive validity studies that evaluate the consistency of the degree to which the assessments predict external criteria across subgroups of students. Zwick and Schlemmer (2004) provided an example of this approach with respect to the differential predictive validity of the SAT for native English speakers and non-native English speakers. These studies are particularly relevant for the “on track” and “college and career readiness” goals of Smarter Balanced. Observational studies using grouping variables could also be conducted using an expected hypothesis of no difference across groups. For

example, by using changes in students' scale scores over time as the dependent variable, comparisons could be made across students from different ethnic groups, socioeconomic status, gender, and other demographic characteristics.

Test Consequences. The analysis of the assessment results can be used to determine if there are differential consequences for various types of students. In describing validity, studies based on testing consequences investigating the effects on instruction, teacher morale, and on students' emotions and behaviors (e.g., dropouts, course-taking patterns) can be conducted. These types of results could also be broken out by subgroup, but more important, the changes in instructional decisions for students should be investigated at the subgroup level. Some important questions might include: Are minority students dropping out of school at higher rates? Are the success rates for remedial programs higher for different types of students?

References

- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14, 219-234.
- Abedi, J., Lord, C., & Plummer, J. (1995). *Language background as a variable in NAEP mathematics performance* (CSE Technical Report 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disability Research and Practice*, 20(1), 50-57.
- Bailey, A. L., Huang, B. H., Shin, H W., Farnsworth, T., & Butler, F. A., (2007) *Developing academic English language proficiency prototypes for 5th grade reading: Psychometric and linguistic profiles of tasks* (CSE Technical Report 727). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Banks, K. (2013). A Synthesis of the Peer-Reviewed Differential Bundle Functioning Research, *Educational Measurement: Issues and Practice*, 32, 43 -55.
- Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, 25, 133–150.
- Borgioli, G. (2008). Equity for English language learners in mathematics classrooms. *Teaching Children Mathematics*, 15, 185-191.
- Camilli, G. (2006). Test Fairness. In R. L. Brennan (Ed.), *Educational Measurement*, 221-256. Washington, DC: American Council on Education/Praeger.
- Cole, N.S., & Zieky, M. J. (2001). The New Faces of Fairness. *Journal of Educational Measurement*. 38, 4.
- Cook, H.G. & McDonald, R. (2013). Tool to Evaluate Language Complexity of Test Items. Wisconsin Center for Education Research. www.wcer.wisc.edu/publications/.../working_paper_no_2013_05.pdf
- Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, 20, 405-438.

- Dorans, N. J. (2004). Using Subpopulation Invariance to Assess Test Score Equity. *Journal of Educational Measurement*, 41, 43-68.
- Duncan, G. D., del Rio Parant, L., Chen, W-H., Ferrara, S., Johnson, E., Oppler, S., Shieh, Y. (2005). Study of a Dual-language Test Booklet in Eighth-grade Mathematics. *Applied Measurement in Education*, 18, 129-161.
- Ercikan, K. & Koh, K. (2005). Examining the Construct Comparability of the English and French Versions of TIMSS. *International Journal of Testing*, 5(1), 23-35.
- ETS. (2009). *ETS Guidelines for Fairness Review of Assessments*. Princeton, NJ: ETS.
- ETS. (2012). *Smarter Balanced Assessment Consortium: Bias and Sensitivity Guidelines*. Princeton, NJ: ETS.
- Forster, K. I. & Olbrei, I. (1973). Semantic heuristics and syntactic trial. *Cognition*, 2, 319-347.
- Kopriva, R. (2010, September). *Building on student strengths or how to test ELs against challenging math (and science) standards when they don't have the English yet*. Common Core State Standards Implementation Conference.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*. Washington, DC: American Council on Education.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-Centered Assessment Design*. Princeton, NJ: Educational Testing Service.
- Pitoniak, M., Young, J. W., Martiniello, M., King, T., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners*. Princeton, NJ: Educational Testing Service.
- Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the Equivalence of Different Language Versions of a Credentialing Exam. *International Journal of Testing*, 3, 1-20.
- Russell, M. (2011b). *Digital Test Delivery: Empowering Accessible Test Design to Increase Test Validity for All Students*. Paper prepared for Arabella Advisors.
- Schachter, P. (1983). *On syntactic categories*. Bloomington, IN: Indiana University Linguistics Club.
- Shafer W., L., & Rivera, C. (2011). Are EL needs being defined appropriately for the next generation of computer-based tests? *AccELLerate*, 3(2), 12-14.
- Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for Evaluating the Validity of Test Scores for English Language Learners. *Educational Assessment*, 13, 108-131.

- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M. L., Quenemoen, R. F., & Lazarus, S. S. (2011). *Meeting the Needs of Special Education Students: Recommendations for the Race to the Top Consortia and States*. Paper prepared for Arabella Advisors.
- Young, J., Pitoniak, M. J., King, T. C., & Ayad, E. (2012). *Smarter Balanced Assessment Consortium: Guidelines for Accessibility for English Language Learners*. Available from <http://www.smarterbalanced.org/smarter-balanced-assessments/>
- Young, J. W. (2008, December). Ensuring valid content tests for English language learners. *R&D Connections, No. 8*. Princeton, NJ: Educational Testing Service.
- Zwick, R., & Schlemer, L. (2004). SAT Validity for Linguistic Minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice, 23*, 6-16.
- Zwick, R., & Schlemer, L. (2004). SAT Validity for Linguistic Minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice, 23*, 6-16.

Chapter 6 Pilot Test and Special Studies (Dimensionality Analysis and IRT Model Choice)

The Pilot Test administration was designed to collect data on the statistical quality of items and tasks and to implement the basic elements of the program before the Field Test in order to make adjustments accordingly. The Pilot Test also familiarized states, schools, teachers, and students with the kinds of items and tasks that will be part of the Smarter Balanced Summative Assessments to be introduced two years later following the Pilot. Whereas the summative assessment will include a computer adaptive test (CAT) component, the Pilot Tests were not adaptive. They were based on linear (i.e., fixed-form) assessments delivered on computer. Pilot Test forms were intended to resemble the future operational test designs so students and teachers had an additional opportunity to become familiar with the assessment and the types of tasks associated with the Common Core State Standards.

There were two phases of the Smarter Balanced assessment program that preceded the first operational administration in 2014–15. The Pilot Test was conducted in the spring of 2012–13 and the Field Test in the 2013–14 school year. This chapter presents evidence pertaining to the Pilot Test that informed the subsequent Field Test. The goal of the Pilot Test was to gather information for a number of purposes, which included

- performing a “dry run” of the procedures to be used in the Field Test;
- evaluating the performance characteristics of CAT items and performance tasks, including comparing performance of individual, student-based performance tasks with those that have a classroom-based component;
- evaluating item performance to inform subsequent Field Test content development;
- investigating test dimensionality and its implications for Item Response Theory (IRT) scaling;
- selecting IRT scaling models;
- evaluating scoring processes and rater consistency; and
- supporting the eventual operational CAT administration.

A design for the Pilot Test poses considerable challenges given the wide variety of purposes that the data are intended to serve. The variety of these requirements demands a data collection design that is necessarily complicated in its specifications and accompanying details. The Pilot Test design was used to collect data based on a specified sample of students and to obtain the psychometric characteristics of items (e.g., item analyses), tasks, and test forms. It is important to note that all Pilot Test items will be rescaled in later operational phases. Subsequently the Field Test (see Chapter 7) was used to establish the Smarter Balanced scale. To avoid confusion, the test-level scaling results for the Pilot are not presented since they are only informative for a brief period prior to the Field Test. The Pilot Test items were linked onto the final scale (i.e., the raw logistic theta scale) in later operational phases of Smarter Balanced. However, the Pilot Test IRT output was used to inform the IRT model choice used in creating Smarter Balanced scales based on the Field Test data.

A relatively large number of items was necessary to ensure sufficient number item survival to conduct various analyses. This required multiple test forms to be administered for each grade level. These test forms needed to be linked so that all items and tasks from different forms could be placed on a common vertical scale (i.e., linked) across grades. Two methods of linking were used in concert. The first one is called the “common items” method, which requires that the blocks of items overlap across test forms. The second approach was “randomly equivalent groups”, wherein the test content is randomly administered to different student samples. Obtaining random equivalence is greatly facilitated by the assignment of test content online. Both linking approaches have respective strengths and weaknesses. While the common items approach is capable of providing strong linking,

it is both relatively inefficient (due to the overlap or redundancy in test material across groups) and dependent on the common items performing consistently across groups (item position and context effects may prevent this). On the other hand, the randomly equivalent groups method is efficient but vulnerable to sampling errors. Because neither linking method is guaranteed to work completely, the Pilot Test design incorporated both linking types. This was accomplished by assembling partially overlapping blocks of test content and randomly assigning those blocks to students. The result is a design that is both reasonably efficient and robust to many potential sources of error. The resulting data are also well structured for IRT calibration. The designs also incorporated common-item links between grade levels in order to establish a preliminary vertical scale. These links are implemented by administering blocks of test content sampled from the adjacent lower- or upper-grade level at most grade levels. For the Pilot Test, content administered from an upper grade to a lower grade was screened by content experts to minimize concerns regarding students' opportunity to learn.

Pilot Data Collection Design

The data collection designs, a critical component of the Pilot Test design, are primarily configured around the scaling requirements, efficiency, and a careful consideration of the testing time required of participating schools. Any data collection design is necessarily a compromise among cost, practicality, and the expected quality of results. In general, one seeks designs that maximize efficiency given practical constraints without unduly affecting the quality of results. Designs that are robust to common sources of errors but which remain practical to implement are also preferred. The Pilot Test design was intended to best balance these considerations and still meet the purpose of collecting data to perform the linking design that makes use of both common items and equivalent groups. The Pilot Test data collection design maximizes design efficiency (i.e., allows the maximum number of items to be tested within the minimum amount of testing time) while conforming to a number of constraints that were deemed necessary for the horizontal and vertical linking of all the items. These design constraints included the following:

- Each Pilot administration configuration has at least one on-grade CAT component that overlaps with other Pilot forms. Items targeted at the eventual summative and interim CAT item pools are collectively referred to as the CAT component. A CAT component or module is a content-conforming collection of items that are common to a selected sample of students. The on-grade CAT components played a major role in placing on-grade and off-grade CAT, and performance task (PT) collections from different configurations, all onto a common measurement scale.
- Each Pilot form configuration was intended to take approximately the same amount of time to complete within a classroom. This requirement is necessary both in terms of maximizing the number of items administered within the allotted time and providing administrative convenience to schools and classrooms to the extent possible.

The first constraint is important for establishing valid horizontal and vertical scales, and the second is important for spiraling of tests and for maximizing administrative efficiency.

In the Pilot Test data collection design, every student took a CAT component. In the case of English Language Arts/literacy (ELA/literacy), there could be two CAT components assigned to students and three in the case of mathematics. This was intended to balance, in terms of testing time, the other condition where students were assigned a performance task and a single CAT component. The CAT-only component consisted of several on-grade CAT components or an on-grade component(s) plus an off-grade CAT component. The off-grade component contained blueprint-conforming item content for either the adjacent lower- or upper grade. The performance task could have been either an on-grade or an adjacent off-grade performance task. The administration procedures for individually assigned performance tasks and ones with an added classroom activity differed. All performance tasks were

individually based and spiraled together with the CAT components at the student level. The classroom performance tasks were assigned at the school level but different tasks were spiraled within that activity type. Table 1 gives the total numbers of CAT components and performance tasks per grade level for ELA/literacy and mathematics. Also shown in Table 1, five unique performance tasks were developed for each grade and content area, and they were administered to students in both the upper adjacent grade and lower adjacent grades.

Table 1. Total Number of CAT Components and Performance Tasks (PT).

Grade	ELA/literacy		Mathematics	
	CAT	PT	CAT	PT
3	10	5	12	5
4	12	5	12	5
5	12	5	12	5
6	12	5	12	5
7	13	5	15	5
8	13	5	13	5
9	6	5	8	5
10	6	5	8	5
11	12	5	18	5

Pilot Items for the CAT Pool. The CAT consists of both selected-response (SR) and constructed-response (CR) items. CAT components were administered linearly online and were mostly machine scored. Each CAT component reflected the Pilot Test blueprint and was roughly interchangeable in terms of expected testing time with other components, in a given grade/content area. The CAT component test blueprints are presented in Tables 2 to 4 for ELA/literacy and mathematics.

Table 2. ELA/literacy Grades 3 to 10 Pilot Test CAT Component Blueprint.

Claim	Score Reporting Category	Passage	No. Items	Discrete SR	Discrete CR
Reading	Literary	1 short 1 long	5		
	Informational		10		
Writing	Purpose/Focus/Org	N/A	6	1	1
	Evidence/Elaboration			1	1
	Conventions			1	1
Speaking/Listening	Listening	1 passage	8		
Research	Research	N/A	2	1	1
Total No. Of Items			31		
Estimated Average Testing Time			~64 minutes		

Table 3. ELA/literacy Grade 11 Pilot Test CAT Component Blueprint.

Claim	Score Reporting Category	Passage	No. Items	Discrete SR	Discrete CR
Reading	Literary	1 short or 1 long	5 or 10		
	Informational	1 short and 1 long	15		
Writing	Purpose/Focus/Org	N/A	6	1	1
	Evidence/Elaboration			1	1
	Conventions			1	1
Speaking/Listening	Listening	1 passage	8		
Research	Research	N/A	2	1	1
Total No. Of Items			36 or 41		
Estimated Average Testing Time			~75 minutes		

Table 4. Mathematics Grades 3 to 11 Pilot Test CAT Component Blueprint.

Claim	Reporting Category	SR	CR
Concepts and Procedures	Domain Area #1	10	3
	Domain Area #2	2	2
Problem Solving/Modeling & Data Analysis	Prob. Solving	1	2
	Model Data		1
Communicating Reasoning	Comm. Reasoning		2
Total No. Of Items		23	
Estimated Average Testing Time		~45 minutes	

Pilot Performance Tasks. A performance task (PT) is a collection of thematically related items that consists of multiple items/tasks and corresponding scored item responses. Each performance task measured multiple claims and was administered to students in its entirety, due to the thematic nature and the need for reliable information to compare student performance. Each performance task conformed to the test blueprint and was scored using expert raters.

One of the factors addressed by the Pilot design was whether performance tasks should be individually administered or provision made for the addition of a classroom collaboration/activity. An individually based performance task required that students approach the task independently without extensive preparatory activities. A classroom-based performance task entailed classroom activities or student interactions concerning a shared set of performance tasks. Although small-group work may be involved in some part of a Classroom Activity, it was not scored, and preparatory activities were standardized to the extent possible. By definition, all students within a classroom were administered the same Classroom Activity. All performance tasks were developed with a detachable Classroom Activity (i.e., a performance task can be administered with or without the Classroom Activity portion). For the data collection design, both versions of a given performance task (i.e., with and without a Classroom Activity) were administered. The two versions were treated as different performance tasks in the Pilot.

Vertical Linking Item Assignment

Students selected to participate in the Pilot Test took either a mathematics or an ELA/literacy test. Those students taking a combination of a CAT and a PT component covered the full content standards for the Pilot Test. The basic vertical linking design is shown in Figure 1.

- For vertical scaling, the CAT component and PT component assigned to a student in a given grade can be an on-grade or an off-grade from either the adjacent lower grade or the adjacent upper grade. The off-grade content was scrutinized to ensure grade-level appropriateness and representation of the construct and to minimize opportunity-to-learn concerns to the extent possible.
- The item developers identified test content, sampling both on-grade and off-grade content, that best articulated growth across grade levels. In the course of CAT-item and PT

development, items and tasks were given a grade band designation as deemed appropriate and a primary targeted grade. For example, a mathematics item targeted for grade 5 may have a grade band of 4 and 6.

- In each grade, about 60 percent of test content (items, passages, and PTs) were designated as on-grade items; the remaining content was about 20 percent from the adjacent lower grade and 20 percent from the adjacent upper grade. The lowest grade, grade 3, had about 80 percent of the items from grade 3 and about 20 percent of the items from grade 4. Similarly, the highest grade, grade 11, had about 80 percent of the items from grade 11 and about 20 percent of the items off-grade.

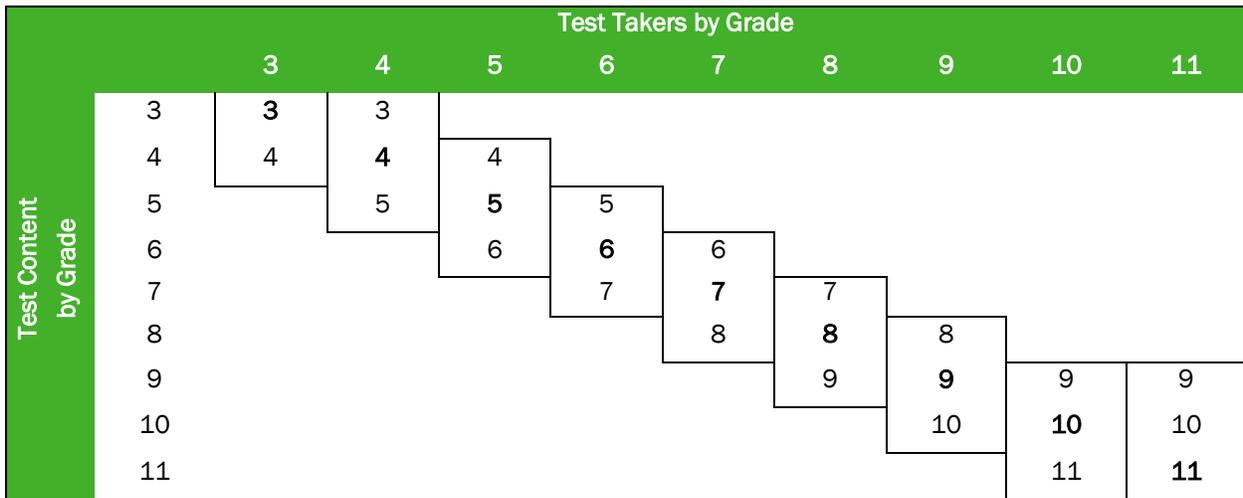


Figure 1. Summary of Vertical Articulation of Test Content by Grade.

Pilot Test Sampling Procedures

Sampling Consideration for the Pilot Test. The characteristics of the Smarter Balanced population provided an operating definition for the composition of the sample and the associated sampling strategies. There were several factors to consider in determining the characteristics of the target population for the Pilot Test, which included state representation, ongoing changes in Consortium membership, transition to the Common Core State Standards (CCSS), and capacity to perform online testing.

The representation of states in the sample is ultimately reflected in the item statistics and the established scales. Two possible state representation models were equal state representation (“Senate”) versus representation proportional to state enrollment population (“House of Representatives”). Equal state representation would place a much greater testing burden on small states and would not represent the larger Smarter Balanced population. On the other hand, if proportional representation were used, a relatively limited number of observations would be obtained for smaller states. Smarter Balanced chose state representation proportional to state enrollment population for the Pilot Test.

Another factor considered in defining the target population was the level of Common Core State Standards implementation. Among Smarter Balanced participants, the extent to which the Common Core State Standards were implemented at the time of the Pilot administration varied considerably. Some states had comparatively high levels of implementation, while others were at the initial stages.

Implementation likely also varied within a state. Since items were written to assess the Common Core State Standards, item performance could be affected by the opportunity to learn these standards. However, since there are no reliable data on levels of implementation across and within states at the time of the Pilot Test, this factor could not be used in sample selection.

The final factor considered in the definition of the target population was the capacity to implement online testing. While some states were already administering online state assessments, other states, districts, and schools had widely varying capacities for online testing. For the purposes of this Pilot, Smarter Balanced decided to target the Pilot Test based on students from schools that had the capacity to implement online testing.

Given the Pilot Test purposes and the nature of the Smarter Balanced assessments, the selected samples were intended to have the following characteristics:

- The selected samples would be representative of the intended/target Smarter Balanced population.
- The selected samples would include students from all Smarter Balanced member states at the time the samples are selected.
- The same sampling procedure would be used to recruit samples for nine grades (3-11) and two content areas (mathematics and ELA/literacy), totaling 18 samples.
- For a given school and grade, a single content area was given in either ELA/literacy or mathematics. Designation of ELA/literacy or mathematics participation occurred through a randomized process in the final step of sampling.
- Due to the need to Pilot both classroom-based and individual-based performance tasks, the smallest sampling units possible were classrooms instead of individual students for these tasks. Performance tasks were spiraled at the classroom level, while the CAT components were assigned at the individual student level.

All schools within the state meeting specifications discussed above were assigned to a stratification cell, and those not initially selected could be used as a replacement when a school declined to participate. As needed, replacement schools were selected from the list of schools/districts that volunteered to participate in the Pilot Test and were not initially selected.

Test Administration and Sample Size Requirements

The Pilot Test is a linear, computer-based administration that was delivered in February and March of 2013. The following were additional characteristics of the sample and test administration conditions:

- The approximate testing time for any configuration of a Pilot form, which consisted of CAT and PT components, would vary somewhat due to the number and types of items a student was administered.
- Some provision was made for multiple test sessions in which test content was administered in defined sections.
- Test content was randomly assigned to individual students to obtain the targeted sample size of 1,500 valid cases for each item. To achieve this target, an oversample of 20 percent (i.e., an additional 300 students) was included. The sample size targeted for each item with oversampling was then 1,800 in total. The sample sizes changed under special circumstances. When an item or a task was deemed appropriate for off-grade administration,

the effective number of observations for the item/task would double so that dependable statistics could be obtained for both grades.

- The number of observations took into account that the three-parameter logistic model was a potential choice of IRT scaling model for selected-response items. More observations were needed to estimate the *c*-parameter accurately than would be the case with models involving fewer parameters.
- Samples were designed so that performance on the Pilot could be compared for designated subgroups or special populations. The Mantel-Haenszel (MH) and standardized mean difference (SMD) procedures were implemented for differential item functioning (DIF) study with estimated IRT ability (θ) as the matching criterion. The minimum sample size for the focal or reference group was 100, and it was 400 for the total (focus plus reference) group.

Note that a sufficient number of cases were scored by raters to permit model building and validation for automated scoring. The cases obtained for the Pilot were designed to be sufficient for this purpose. Table 5 shows the targeted number of students per grade and content area as specified by the Pilot Test Design. In total, approximately one million students were expected to participate in the Pilot Test.

Table 5. Targeted Student Sample Size by Content Area and Grade for the Pilot Test.

Grade	ELA/literacy	Mathematics	Total
3	56,510	51,638	108,148
4	67,227	60,407	127,634
5	67,227	60,407	127,634
6	67,227	60,407	127,634
7	67,227	60,407	127,634
8	67,227	60,407	127,634
9	40,921	35,075	75,996
10	40,921	35,075	75,996
11	64,304	59,433	123,737
Total	538,791	483,256	1,022,047

Pilot Sampling Considerations

In addition to defining the characteristics of the target population, decisions concerning the following sampling issues were evaluated in conducting the Pilot Tests.

Smallest sampling unit. Simple random sampling at the student level cannot be conducted in educational settings because students usually reside within classrooms. In cluster sampling, a population can be composed of separate groups, called clusters. It is not possible to sample

individual students so clusters such as schools or classrooms are used. Whereas stratification generally increases precision when compared with simple random sampling, cluster sampling generally decreases precision. In practice, cluster sampling is often used out of convenience or for other considerations. If clusters have to be used, it is usually desirable to have small clusters instead of large ones. Although cluster sampling normally results in less information per observation than a simple random sample, its inefficiency can usually be mitigated by increasing sample size. One of the purposes of the Pilot Test was to try out both classroom-based and individual-based performance tasks, which required the smallest sampling unit to be no smaller than the classroom. The question is whether the classroom or the school should be the smallest sampling unit. The design effect quantifies the extent to which the expected sampling error departs from the sampling error that might be expected using simple random sampling. Making the classroom the sampling unit certainly has an advantage with regard to sample size requirements and the reduction of design effects. On the other hand, it might be desirable to have the school as the sampling unit in order to facilitate recruiting. In this case, the smallest unit available in educational databases was at the school level. A random sample of schools was selected as clusters within each stratum.

A multiple-stage stratified sampling with nested cluster sampling was used as the primary approach to ensure the representativeness of the selected sample (Frankel, 1983). The states that make up the Smarter Balanced Consortium were used to conduct the first-stage stratification to ensure that each state was adequately represented in the sample. Within each state, additional strata were defined to increase sampling efficiency. Stratification variables (e.g., percentage proficient) were defined as variables that are related to the variable of interest, which is academic achievement on the Common Core State Standards. Out of necessity, stratification variables were limited to those obtained based on school level data. In this complex sampling design, cluster sampling was used within strata due to test administration requirements and cost efficiency. Some variations in the sampling plan permitted flexibility to include all students from selected schools, or to limit the number of students participating. Within each school, one or more grades and content areas were selected. Participating schools were assigned a subject to be administered to each particular grade. Test forms were spiraled within grades. Cluster sampling was also implemented.

Use of sampling weights. Sampling weights can be applied to adjust stratum cells for under- or over-representation (Cochran, 1977; Frankel, 1983). In general, the use of sampling weights, when needed and appropriately assigned, can reduce bias in estimation, but creates complexities in data analyses and increases the chance for errors. One approach is to create a self-weighted sample, in which every observation in the sample gets the same weight. In other words, the probability of selection is the same for every observation unit. To achieve this, the sampling plan needs to be carefully designed. (As an example, it can be noted that self-weighted sampling is not viable for NAEP because it requires oversampling of nonpublic schools and of public schools with moderate or high enrollment of Black or Hispanic students, to increase the reliability of estimates for these groups of students.) In Pilot Test design, a self-weighted sample can be obtained that does not require explicit sample weighting if the following occur:

- consistent state representation in the target population and Pilot sample,
- proportional allocation for the first-stage stratified sampling level,
- under each stratum, cluster sampling with probability proportional to size in the second-stage school sampling and then fixed simple random sampling in that cell.

Nonresponse and replacement. The sampling needs to be designed well to reduce nonresponse errors for schools that decline to participate. A typical procedure to handle nonresponse is to inflate the sampling weights of some of the responding individuals. This nonresponse adjustment acts as if the distributions of characteristics of the nonrespondents within a stratum are the same as those of

the respondents within the same stratum. In the situation where a self-weighted sample is used, two options were suggested to adjust for nonresponses. In both options, replacement schools are selected within the same stratum to ensure that the schools declining to participate are replaced by schools with similar characteristics.

- More schools than required can be selected from each stratum, and schools that decline to participate will be replaced randomly by additional schools selected from the same stratum.
- A single list is created of schools within each stratum in random order. Schools are selected for participation from the list. If school “A” declines to participate, it is replaced using school “B,” which is listed right after school “A” in the original school list. If school “B” has already been selected for participation, it is replaced using school “C,” and so on. The procedure can be repeated as necessary. If school size or other demographic information is available, it is also appropriate to select a replacement school within the same stratum that is most similar in terms of size and demographic features to the school that fails to participate.

Sampling in the context of vertical scaling. Sampling for the Pilot Test considered vertical scaling and some notions with respect to growth. If samples are not consistent across grades, it becomes more difficult to evaluate growth between grades and the quality of the vertical scale may deteriorate.

Kolen (2011, p. 9) states,

Vertical scales can differ across test taker groups, especially when the curriculum differs across groups. For this reason, it is important to use a representative group of students to conduct vertical scaling. In addition, it is important that students in each of the grade groups used to conduct vertical scaling are from the same, or similar, locations. When evaluating vertical scales, it is desirable to compare distributions of vertically scaled scores across grades. Such comparisons are sensible only if the grade groups are from the same, or similar, school districts.

To implement this recommendation, high schools were selected first. Then a middle school was selected, which was intended to be a “feeder” school to the high school selected. In turn, an elementary school was selected, which was a feeder to the middle school. In addition, when a school was identified for participation, tests in the same content area were administered to all grade levels in the school. Under this approach, grade 11 samples would be selected first. Samples for the lower grade levels would first be identified through the feeder approach and then be adjusted to ensure representativeness. This approach, while used for decades by norm-referenced test publishers, is complicated and was highly challenging to execute for this application and was not implemented.

Sampling from voluntary districts/schools. Sampling from voluntary districts/schools is not fundamentally different from recruiting directly from the entire population when districts/schools that do not volunteer are seen as nonresponses. However, the nonresponse rate is expected to be higher under a voluntary approach compared to a “mandatory” recruiting approach. The key question is whether districts/schools that choose not to participate tend to differ from those that volunteer to participate. If systematic differences exist, bias will be introduced from using the subset of volunteering districts/schools.

To minimize bias, it is of critical importance to ensure that the selected samples are representative of the Pilot populations, both in terms of performance on state-level achievement tests and demographic characteristics. To achieve representativeness, pre-Pilot evaluation of volunteering districts/schools was conducted to determine the need for additional recruitment. Districts/schools that volunteered were grouped into different strata. Additional recruiting was needed when the number of students from volunteering districts/schools in each stratum was fewer than required using population characteristics and proportional allocation. After sample selection, sample

representation was checked by comparing state assessment score distributions and demographic summaries of the samples against the state-level population data.

Sampling Procedures

Sample participants were selected from two sampling frameworks. The first sampling framework was from state assessment data for grades 3-8, while grades 9-11 used the QED (QED, 2012). The Quality Education Database (QED, 2012) from the MCH Corporation is a commercially available source used for sampling. There was no indicator for “private” or “public” school in either of these two databases. All schools from the state assessment data and the QED constituted the eligible school population.

Stratification Procedures and Sample Summary. Different stratification variables were necessary for grades 3-8 and grades 9-11, given different sets of variables available from state assessment data and the QED. The percentage proficient on ELA/literacy obtained from a United States Educational Department database was used as the stratification variable for grades 3-8 sample selections. For each grade level, schools were classified into five strata based on the percentage proficient on ELA/literacy such that each stratum constituted about 20 percent of the student population. The percentage of Title I from the QED file was used as the stratification variable to create five equally condensed strata for the grades 9-11 sample selections for most states, except for Hawaii and Nevada. The percentage of Title I information was missing for almost all Nevada schools in the QED data file; therefore, the high school sample from Nevada was selected by using metro/rural information as the stratification variable with four strata being used. Neither the percentage of Title I nor the metro/rural information was available in the QED data file for Hawaii; therefore, all selected high schools or possible replacement schools for high school grades were in a single stratum.

Once the stratification was complete, school demographic variables were used to evaluate the representativeness of the resulting sample. The selected Pilot sample was expected to be representative of the target population at each grade level in the following performance and demographic categories:

- School gender proportions,
- School ethnicity proportions,
- Percentage of students with disabilities,
- Percentage of students classified as having limited English proficiency, and
- Percentage free or reduced-lunch.

Detailed Sampling Procedure. A sample was considered representative of the population when the sample characteristics matched population characteristics in terms of performance as well as demographics. Given the Pilot Test purposes, the sampling involved nine steps:

- Step 1: Determine the number (proportion) of students that should be obtained from each Smarter Balanced member state.
- Step 2: Obtain a list of voluntary districts/schools from each state, if applicable.
- Step 3: Determine the stratification variables that will be used to combine schools into strata within each state.
- Step 4: Determine the number of students that should come from each stratum within each state through proportional allocation.
- Step 5: Select Pilot participants from each stratum using school as the sampling unit.

- Step 6: Evaluate the extent to which the selected sample is representative of the target population.
- Step 7: Designate subjects/content areas within a given grade.
- Step 8: Follow replacement procedures for schools declining to participate.
- Step 9: Check representativeness by evaluating state assessment score distributions and demographic summaries of the samples compared with the state-level population data.

Sample Distribution across States and Demographic Distributions. In total, approximately 1,044,744 students from 6,444 schools were targeted for pilot participation. Among the 6,444 schools, 4,480 schools had two grade levels selected for participation, and 1,964 schools had one grade level selected for participation. The numbers of targeted and obtained students by content area and grade level are shown in Table 6. It also summarizes the overall numbers of targeted and obtained students by content area for each state. Tables 7 and 8 show the resulting demographic characteristics after the Pilot Test administration for ELA/literacy and mathematics.

Table 6. Approximate Sample Sizes by Content Area and State, the Sample Target and the Number Obtained for the Pilot Test.

State	ELA/literacy		Mathematics		Total	
	Target	Obtained	Target	Obtained	Target	Obtained
California	199,122	199,052	178,598	179,195	377,719	378,247
Connecticut	17,632	18,018	15,815	15,625	33,448	33,643
Delaware	4,054	8,777	3,636	7,470	7,689	16,247
Hawaii	5,745	5,948	5,153	6,439	10,898	12,387
Idaho	8,801	9,174	7,893	9,233	16,694	18,407
Iowa	15,116	14,341	13,558	13,943	28,674	28,284
Kansas	14,921	15,504	13,383	12,835	28,305	28,339
Maine	5,964	6,116	5,349	5,611	11,313	11,727
Michigan	52,074	52,536	46,707	46,467	98,781	99,003
Missouri	28,699	29,043	25,741	25,930	54,440	54,973
Montana	4,522	3,867	4,056	4,421	8,577	8,288
Nevada	13,668	14,565	12,259	12,516	25,927	27,081
New Hampshire	6,244	7,602	5,600	7,825	11,844	15,427
North Carolina	46,847	47,466	42,019	41,610	88,866	89,076
Oregon	18,064	18,374	16,202	16,368	34,266	34,742
South Carolina	22,471	22,242	20,154	20,749	42,625	42,991
South Dakota	3,935	3,758	3,529	4,315	7,464	8,073
Vermont	2,785	3,454	2,498	2,909	5,284	6,363
Washington	32,942	33,738	29,546	29,453	62,488	63,191
West Virginia	8,644	9,261	7,753	8,084	16,396	17,345
Wisconsin	26,544	27,045	23,808	23,865	50,352	50,910
Total	538,793	549,881	483,257	494,863	1,022,050	1,044,744

Table 7. ELA/literacy Student Population and Sample Characteristics (Percentages).

Demographic Groups	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10		Grade 11	
	Pop.	Sample	Pop.	Sample	Pop.	Sample												
Female	48.77	49.52	48.80	49.27	48.75	49.73	48.69	49.45	48.76	49.23	48.86	49.67	NA	49.71	NA	48.84	NA	48.80
Male	51.19	50.48	51.17	50.73	51.24	50.27	51.31	50.55	51.25	50.77	51.16	50.33	NA	50.29	NA	51.16	NA	51.20
White	52.31	43.23	52.65	42.77	52.93	40.65	52.97	37.66	52.65	39.77	53.09	41.02	58.79	40.70	60.05	39.25	60.05	32.45
Asian	8.00	7.08	7.89	7.51	8.04	6.62	7.22	6.68	7.06	6.73	7.35	5.55	7.14	7.96	6.73	5.83	6.73	4.75
Black	13.60	7.33	13.67	6.77	13.70	7.37	13.08	6.53	12.57	8.39	12.46	8.57	11.36	8.33	11.91	8.21	11.91	8.95
Hispanic	28.97	8.47	28.50	10.64	28.03	11.70	27.22	13.29	26.79	11.24	26.53	9.38	21.57	11.50	20.17	13.96	20.17	14.86
Native American	2.59	0.83	2.56	0.72	2.51	0.67	1.92	0.85	1.66	0.82	1.62	1.00	1.05	0.77	1.04	0.80	1.04	0.63
Pacific Islander	NA	0.85	NA	0.87	NA	0.90	NA	0.74	NA	0.80	NA	0.55	NA	1.46	NA	0.31	NA	1.72
Multi-Race	4.20	16.26	4.15	15.50	3.93	17.04	3.52	15.60	3.21	15.26	3.10	13.43	NA	14.01	NA	8.25	NA	14.16
Unknown	NA	15.96	NA	15.23	NA	15.05	NA	18.64	NA	16.99	NA	20.50	NA	15.27	NA	23.38	NA	22.49
No IEP	NA	58.31	NA	59.72	NA	61.27	NA	60.55	NA	57.08	NA	58.05	NA	64.56	NA	57.58	NA	60.45
IEP	NA	8.41	NA	8.51	NA	8.36	NA	7.65	NA	7.27	NA	6.56	NA	6.46	NA	6.34	NA	6.00
Unknown	NA	33.28	NA	31.77	NA	30.38	NA	31.80	NA	35.65	NA	35.39	NA	28.98	NA	36.08	NA	33.55
Not LEP	NA	50.44	NA	51.70	NA	54.29	NA	53.31	NA	53.67	NA	57.11	NA	62.25	NA	54.47	NA	54.64
LEP	20.27	15.78	17.59	16.04	14.75	16.51	11.58	14.91	10.06	10.79	9.62	9.73	NA	8.81	NA	7.58	NA	10.35
Unknown	NA	33.78	NA	32.27	NA	29.20	NA	31.78	NA	35.54	NA	33.16	NA	28.94	NA	37.95	NA	35.01
Not Title 1	NA	43.18	NA	46.74	NA	47.81	NA	47.88	NA	49.32	NA	45.64	NA	56.97	NA	45.82	NA	50.44
Title 1	NA	21.34	NA	21.82	NA	21.14	NA	19.35	NA	16.03	NA	16.23	34.36	5.89	32.91	12.74	32.91	13.34
Unknown	NA	35.48	NA	31.44	NA	31.05	NA	32.77	NA	34.65	NA	38.14	NA	37.14	NA	41.43	NA	36.22
Stratum 1		11.05		12.25		12.19		12.80		15.24		15.23		35.79		35.90		19.58
Stratum 2		20.01		18.79		19.95		20.36		18.44		20.65		24.01		22.28		26.52
Stratum 3		21.23		23.06		25.16		23.24		24.14		25.70		17.66		21.36		31.44
Stratum 4		26.59		24.44		24.64		22.85		25.84		21.86		10.52		14.48		14.82
Stratum 5		20.59		21.18		17.76		20.62		16.09		16.44		9.72		5.06		6.50
Unknown		0.53		0.27		0.30		0.13		0.25		0.12		2.31		0.93		1.14

Table 8. Mathematics Student Population and Sample Characteristics (Percentages)

Demographic Groups	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10		Grade 11	
	Pop.	Sample	Pop.	Sample	Pop.	Sample												
Female	48.77	49.11	48.80	49.16	48.75	49.39	48.69	49.89	48.76	49.40	48.86	49.28	NA	50.95	NA	49.63	NA	49.84
Male	51.19	50.89	51.17	50.84	51.24	50.61	51.31	50.11	51.25	50.60	51.16	50.72	NA	49.05	NA	50.37	NA	50.16
White	52.31	41.06	52.65	40.81	52.93	40.27	52.97	42.49	52.65	36.71	53.09	37.18	58.79	35.43	60.05	39.47	60.05	43.42
Asian	8.00	6.91	7.89	6.58	8.04	6.97	7.22	5.96	7.06	4.62	7.35	6.20	7.14	4.94	6.73	6.57	6.73	7.83
Black	13.60	7.11	13.67	8.14	13.70	6.62	13.08	9.30	12.57	8.13	12.46	8.15	11.36	8.62	11.91	7.83	11.91	8.49
Hispanic	28.97	11.30	28.50	11.46	28.03	13.19	27.22	10.75	26.79	11.15	26.53	10.45	21.57	14.75	20.17	13.60	20.17	13.51
Native American	2.59	0.60	2.56	0.60	2.51	0.97	1.92	0.78	1.66	0.63	1.62	0.74	1.05	0.62	1.04	0.80	1.04	0.61
Pacific Islander	NA	1.04	NA	0.95	NA	0.81	NA	0.86	NA	0.68	NA	1.20	NA	2.52	NA	2.16	NA	1.15
Multi-Race	4.20	15.93	4.15	16.28	3.93	14.81	3.52	15.85	3.21	17.49	3.10	14.67	NA	15.03	NA	17.33	NA	12.96
Unknown	NA	16.06	NA	15.18	NA	16.36	NA	14.01	NA	20.60	NA	21.39	NA	18.10	NA	12.23	NA	12.03
No IEP	NA	60.39	NA	60.48	NA	59.79	NA	62.45	NA	60.71	NA	59.18	NA	54.51	NA	70.52	NA	65.45
IEP	NA	8.20	NA	8.42	NA	8.20	NA	8.01	NA	7.54	NA	6.77	NA	6.42	NA	7.09	NA	6.19
Unknown	NA	31.41	NA	31.10	NA	32.01	NA	29.53	NA	31.75	NA	34.06	NA	39.08	NA	22.39	NA	28.35
Not LEP	NA	53.00	NA	53.32	NA	52.60	NA	55.54	NA	57.13	NA	54.54	NA	49.74	NA	59.29	NA	60.60
LEP	20.27	16.58	17.59	16.74	14.75	16.16	11.58	13.05	10.06	13.65	9.62	9.15	NA	11.92	NA	12.83	NA	9.53
Unknown	NA	30.42	NA	29.94	NA	31.24	NA	31.41	NA	29.22	NA	36.31	NA	38.34	NA	27.87	NA	29.87
Not Title 1	NA	44.62	NA	42.53	NA	47.65	NA	48.79	NA	43.33	NA	47.16	NA	36.70	NA	57.76	NA	57.41
Title 1	NA	23.68	NA	24.72	NA	22.23	NA	19.00	NA	19.63	NA	15.09	34.36	18.34	32.91	15.70	32.91	11.51
Unknown	NA	31.70	NA	32.75	NA	30.12	NA	32.22	NA	37.04	NA	37.75	NA	44.96	NA	26.54	NA	31.08
Stratum 1		10.96		10.93		11.75		13.74		17.51		15.80		31.22		18.05		31.27
Stratum 2		16.53		20.26		19.18		18.89		21.47		15.68		22.61		33.76		25.46
Stratum 3		25.38		24.52		24.85		25.80		23.73		23.20		26.30		31.17		23.97
Stratum 4		25.94		25.47		22.66		24.11		21.99		22.83		9.79		8.49		10.37
Stratum 5		20.54		18.09		21.45		17.36		14.24		21.79		9.53		7.42		7.51
Unknown		0.64		0.73		0.12		0.10		1.06		0.70		0.54		1.11		1.42

Although the samples were intended to be representative of their respective populations in characteristics such as their 2012 state test performance, gender, ethnicity, and special programs, the Pilot Test administration resulted in a convenience sample due to administration constraints. Due to the lack of sample representativeness, any comparisons of results over grades and generalizations to larger student populations should be made cautiously. In the context of a pilot, sample size was generally sufficient for item calibration and estimating item difficulty.

Pilot Classical Test Results

This section contains the statistical analysis summary of results pertaining to the Smarter Balanced Pilot Test. This section focuses on and summarizes the data inclusion/exclusion rules, classical statistics, differential item functioning (DIF) analysis, and other relevant factors such as test duration. The Pilot Test provided additional insight into many factors and areas in which modification to the program and content might be necessary for the Field Test. The following interpretive cautions for the Pilot Test administration are given:

- The Pilot Test administration used a preliminary version of the Smarter Balanced test blueprints.
- While Pilot tests were being delivered or scored, some items and item types were eliminated.
- Although the initial design was intended to have representative student samples, the student samples that were obtained largely resulted in convenience samples.
- The performance task component underwent significant revision after the Pilot Test so that the Classroom Activity would be a required component of the Field Test administration.
- The number of scorable performance tasks was very small for some tests, and there were no surviving performance tasks for the mathematics tests.
- In the case of constructed-response, scoring using raters was performed that targeted a maximum of 1,800 scored responses for each item. However, some item types were well below the targeted number of observations.
- Based on the preliminary data review, recommendations were implemented concerning which items to include or exclude from the item bank. Items were included if they were not rejected by data review and if they had item-total correlations greater than 0.15.
- Items meeting all acceptance criteria will be recalibrated onto the Smarter Balanced scale in an operational phase.

Major Pilot Test activities were item and DIF analyses for CAT items used as an input into data review (completed in October 2013). Two additional studies were performed using the Pilot Test data to inform test design for the Field Test. A dimensionality study was used to explore grade-level and adjacent-grade dimensional structure. A comparison of IRT models was conducted to provide a basis for the selection of an IRT model. IRT model choice results were reviewed on May 1, 2014 by the Smarter Balanced Technical Advisory Committee in concert with the Smarter Test Validation and Psychometrics Work Group. After considering their comments and recommendations, the consortium adopted the two-parameter (2-PL) and generalized partial credit model (GPCM) for the program.

In the Pilot, students took either a CAT component/modules or a combined CAT and performance task (PT) configuration. Students taking only CAT components took two ELA/literacy or three mathematics content representative item collections as stated previously. Each mathematics component had a total of 23 selected-response (SR) and constructed-response (CR) items and was expected to require approximately 45 minutes in testing time along with time for administrative instructions. An ELA/literacy component had about 29 items at lower grade levels and 33 items at

high school grades, and each component was expected to take about 60–75 minutes to complete. All single-selection SR items had four choices and multiple-selection selected-response (MSR) items had five to eight choices. The performance task items had maximum scores ranging from one to four points. In accordance with the test design, other groups of students were administered a single CAT component and a performance task. A performance task was expected to have approximately five scorable units yielding approximately 20 score points in total. Overall, 1,602 ELA/literacy CAT items, 49 ELA/literacy performance tasks (which included 318 items), and 1,883 mathematics CAT items were evaluated. No mathematics performance tasks were scored and used for subsequent analyses. These items collections, in aggregate, represented ELA/literacy and mathematics in all claims. The majority of the Pilot Tests (CAT components and PTs) were administered to students at the grade for which the items/tasks were developed (i.e., the on-grade administration of items/tasks). Selected Pilot CAT components and performance tasks were also administered to students at the adjacent upper or lower grade intended to facilitate vertical linking (i.e., the off-grade administration of items/tasks).

Pilot Classical Item Flagging Criteria

In this section, the item analysis and differential item functioning (DIF) flagging criteria for the Smarter Balanced 2013 spring Pilot Test administration is summarized. Statistics from the item analysis and DIF procedures were used to determine the disposition of Pilot Test items in the context of a data review conducted by content experts. Three possible outcomes based on item review resulted for these items.

- An item could be directly deposited into the Field Test item bank without modification except if further scaling was still required.
- If an item was not functioning as expected, it was modified accordingly (i.e., replaced with a new edited item) before being deposited in the item bank and rescaled as necessary.
- The item (or item type) was eliminated from the pool.

Very poor-functioning items that were not initially eliminated could affect the criterion score used in computing the item-test correlation.

Criteria based on Classical Item Analyses. A high-level description of the flagging criteria is given below.

- Observed Percentage of Maximum (p -value): Items with average item difficulty < 0.10 or > 0.95 .
- Omits or Not Responding: Items with omits/no response greater than 20 percent.
- Point Biserial Correlation and Item-test Correlations: Items with point-biserial correlation less than 0.30. This was under the assumption that the within-grade data structure is essentially unidimensional. Items with a very low point-biserial ($< .05$) have the answer keys verified.
- Other Criteria for Selected-response Items
 - Items with proportionally more higher-ability students selecting a distractor over the key.
 - Items with higher total score mean for students that choose a distractor rather than the keyed response.
- Other Criteria for Polytomous Items (i.e., items with more than two score categories): Items with percentages obtaining any score category less than 3 percent.

Criteria based on Differential Item Functioning Analyses. DIF analyses are used to identify items in which defined subgroups (e.g., males, females) with the same ability level have different probabilities of obtaining a given score point. Items are classified into three DIF categories of “A”, “B”, or “C.” Category A items contain negligible DIF, category B items exhibit slight or moderate DIF, and category C items have moderate to large values of DIF. Negative values (B- or C-) imply that, conditional on the matching variable, the focal group (female, Asian, African-American, Hispanic, Native-American, etc.) has a lower mean item score than the reference group (male, white). In contrast, a positive value (B+ or C+) implies that, conditional on total test score, the reference group has a lower mean item score than the focal group. DIF was not conducted if the sample size for either the reference- or focal group was less than 400 or 100, respectively.

Description of Pilot Classical Statistics Evaluated

Item Difficulty. The observed proportion of maximum or p -value is computed for each item as an indicator of item difficulty with a range of 0 to 1. The higher the p -value value is, the easier the item is. A p -value of 1.0 for an item indicates that all students received a perfect score on the item. Likewise, p -values of 0.0 for an item indicate that no students got the item correct or even received partial credit for a constructed-response item. For a dichotomous item, the p -value is equivalent to the proportion of students who answered the item correctly. For a polytomous item, the p -value refers to the observed mean score as a proportion of the maximum possible total score. For instance, for a polytomous item with scores ranging from 0 to 3 and an observed mean score of 2.1, the observed proportion of maximum is calculated as $2.1/3 = 0.7$.

Items covering a wide difficulty level range are needed to support future operational CAT and performance tasks. Very easy and very difficult items, however, will need to be reviewed to ensure that the items are valid for assessing grade-appropriate content standards. Note that some items serve as anchor items in vertical scaling. These items are administered across multiple grade levels and therefore can have several sets of grade-level specific item statistics. The p -values from different grade levels are assessed to evaluate if students in a higher-grade level perform better on these items than students in a lower grade level.

Item Discrimination. Item discrimination analysis evaluates how well an item distinguishes between students of high and low ability. This is typically done by calculating the correlation coefficient between item score and criterion score (usually total score or IRT ability estimate), generally referred to as “item-total correlation.” A large item-total correlation coefficient value is desired, as it indicates that students with higher scores on the overall test tend to perform better on this item. In general, item-total correlation can range from -1 (for a perfect negative relationship) to 1 (for a perfect positive relationship). However, a negative item-total correlation signifies a problem with the item, as the higher-ability students are getting the item wrong and the lower-ability students are getting the item right.

Typical coefficients used in computing item-total correlations are the polyserial correlation coefficient (used for polytomous items) and the Pearson correlation coefficient (with the point-biserial correlation coefficient being a special case of the Pearson correlation coefficient used for dichotomous items). Point-biserial correlations are computed as

$$r_{ptbis} = \frac{(\bar{X}_+ - \bar{X}_-)}{s_{tot}} \sqrt{pq}$$

where \bar{X}_+ is the mean criterion score of test takers answering the item correctly; \bar{X}_- is the mean criterion score of the examinees answering the item incorrectly; s_{tot} is the standard deviation of the criterion score of students answering the item; p is the proportion of test takers answering the item correctly, and q equals $(1 - p)$.

The polyserial correlation measures the relationship between a polytomous item and the criterion score. A polytomous item is an item that is scored with more than two ordered categories, such as the ELA/literacy long-write essay. Polyserial correlations are based on a polyserial regression model (Drasgow, 1988; Lewis & Thayer, 1996), which assumes that performance on an item is determined by the students' location on an underlying latent variable that is normally distributed at a given criterion score level. Based on this model, the polyserial correlation can be estimated using

$$r_{polyreg} = \frac{bs_{tot}}{\sqrt{b^2s_{tot}^2 + 1}}$$

where b is estimated from the data using maximum likelihood and s_{tot} is the standard deviation of the criterion score.

Item Response Distribution. For each selected-response item, distractor analyses are conducted. The quality of distractors is an important component of an item's overall quality. Distractors should be clearly incorrect, but at the same time plausible and attractive to the less able students. The following distractor analyses are conducted to evaluate the quality of item distractors:

- Percentage of students at each response option is calculated. For the answer key, this percentage is equal to the p -value. If the percentage of students who selected a distractor is greater than the percentage of students who selected the answer key, the item is then examined for errors or double-keys. On the other hand, if there are no students or very few students who selected a particular distractor, then this distractor might be implausible or too easy and is not contributing to the performance of the item. An implausible distractor in a multiple-choice item can make the item easier than intended.
- A point-biserial correlation is calculated for each response option. While the key should have positive biserial correlations with the criterion score, the distractors should exhibit negative point-biserial correlations (i.e., lower-ability students would likely choose the distractors, while the higher-ability students would not).
- The average estimated ability level is calculated for students at each response option. Students choosing the answer key should be of higher ability levels than students choosing distractors.
- The percentage of high-ability students at each response option is calculated. High-ability students are defined to be the top 20 percent of all students in the ability distribution. If the percentage of high-ability students who selected a distractor is greater than the percentage of high-ability students who selected the key, the item should be examined further.

For each constructed-response item, the following analyses are conducted to examine the score distribution.

- The percentage of students at each score level is calculated. If there are very few students at certain score levels, this might suggest that some score categories need to be collapsed and the scoring rubric needs to be adjusted or the item eliminated.
- The average ability level is calculated for students at each score level. Students at a higher score level on this item should be of higher ability levels (i.e., having higher average ability estimates) than students at a lower score level on this item. The observed percent of the maximum possible raw score was used as the student ability estimate here.

Pilot Results

The response data for the items were collected from student samples ranging in size from approximately 12,000 students in some high school grades to more than 40,000 in grades 3 to 8. Table 9 summarizes the obtained item pool sizes and associated student samples for all 18 tests.

Table 9. Summary of Number of Pilot Test Items and Students Obtained.

Grade	ELA/literacy		Mathematics	
	Number of Items	Number of Students	Number of Items	Number of Students
3	241	41,450	212	41,502
4	236	49,797	214	43,722
5	184	49,522	210	46,406
6	227	49,670	213	42,051
7	210	44,430	230	41,408
8	232	41,132	224	44,650
9	146	25,690	135	19,298
10	157	16,079	139	12,438
11	287	18,904	306	24,405

The CAT component statistical characteristics such as mean scores and reliability are summarized by grade and content area. Students performed noticeably worse on the components from upper-adjacent grade, which is somewhat expected. Tables 10 and 11 provide an overview of test-level statistics for ELA and Mathematics. Mathematics components were considerably more difficult than ELA ones, especially at Grades 7 and up. At these grade levels, the test-level difficulties range from .10 to .30. There are a larger number of items flagged for low average score and low item-total correlation. It was unclear whether this is due to student low motivation or if the items are simply too difficult.

Table 10. Overview of ELA/literacy CAT Component Statistics.

Grade	No. of Components	Sample <i>N</i>		Reliability			Percent of Maximum		
		Min	Max	Min	Max	Median	Min	Max	Median
3	13	1,369	9,539	0.75	0.86	0.81	33.95	54.80	45.62
4	18	1,092	7,426	0.70	0.83	0.77	34.78	54.38	44.75
5	18	1,177	9,976	0.64	0.80	0.72	37.00	53.27	45.26
6	18	1,278	4,915	0.60	0.80	0.72	37.34	48.26	42.96
7	19	1,060	4,534	0.55	0.84	0.72	34.21	50.05	41.33
8	19	491	4331	0.53	0.79	0.69	35.07	46.36	42.38
9	12	1,139	4,858	0.50	0.84	0.70	33.39	50.73	42.43
10	12	507	2,838	0.64	0.81	0.72	31.40	47.07	36.76
11	24	249	1,772	0.59	0.83	0.74	27.12	42.36	33.16

Table 11. Overview of Mathematics Component Statistics.

Grade	No. of Components	Sample <i>N</i>		Reliability			Percent of Maximum		
		Min	Max	Min	Max	Median	Min	Max	Median
3	15	1,743	6,199	0.67	0.87	0.79	26.01	51.82	36.77
4	18	1,917	4,763	0.67	0.87	0.81	15.80	48.42	35.96
5	16	2,062	5,116	0.74	0.86	0.83	23.72	42.50	35.57
6	18	1,801	4,498	0.65	0.88	0.79	22.13	45.01	32.46
7	21	893	3,642	0.62	0.84	0.79	15.57	35.95	26.08
8	20	1,416	5,166	0.59	0.84	0.75	11.59	34.38	25.02
9	15	705	3,527	0.58	0.76	0.63	9.90	26.55	20.88
10	16	631	2,106	0.54	0.79	0.69	14.70	33.22	20.89
11	30	536	2,272	0.52	0.83	0.72	10.03	28.33	18.92

Test Duration. A key characteristic of a test is the time required per student to complete the assessment. Test duration is defined as the time span from when the student entered the

assessment until the submit button was used to end the testing. Individual student response time for the items administered corresponds to total testing time when summed. This can be averaged across students to obtain an estimate of the testing time for a grade and content area. This was not possible due to a number of complicating factors. For instance, students could stop (i.e., pause) the test as many times they wanted. While the number of pauses was captured, the total pause time was not collected. In addition, multiple items can be presented on a single page, which further complicates the estimation of item response time. Table 12 provides the available information regarding the student testing duration.

Table 12. Student Testing Durations in Days (Percentage Completion).

Grade	Subject	Percentages of Students Completing a Test			Maximum Days Used
		Within a Day	More than One Day	Duration Unknown	
3	ELA/literacy	29.65	55.77	14.57	28
	Mathematics	34.15	55.12	10.73	28
4	ELA/literacy	31.05	57.58	11.36	30
	Mathematics	38.81	49.83	11.36	16
5	ELA/literacy	32.93	55.12	11.95	30
	Mathematics	39.44	48.90	11.66	17
6	ELA/literacy	39.37	48.32	12.31	29
	Mathematics	39.24	46.19	14.57	26
7	ELA/literacy	37.97	48.88	13.15	26
	Mathematics	38.46	42.15	19.39	24
8	ELA/literacy	39.17	47.23	13.60	24
	Mathematics	42.74	41.17	16.09	19
9	ELA/literacy	52.00	34.85	13.15	16
	Mathematics	52.07	33.58	14.35	14
10	ELA/literacy	53.96	31.53	14.52	20
	Mathematics	64.09	28.36	7.56	11
11	ELA/literacy	56.10	30.55	13.34	20
	Mathematics	58.62	28.62	12.76	14

CAT Summary Statistics. After receipt of the scored student response data, analyses were conducted to gain information about basic item characteristics and quality of tasks. These analyses included the review of item difficulty and discrimination, item response frequency distribution, and differential item functioning (DIF). In Table 13, statistics for the CAT pool are presented, including the number of students obtained, test reliabilities (i.e., coefficient alpha), and observed score distributions as percentages of the maximum possible scores of the item collections. In general, the on-grade administration of Pilot CAT components received more student responses than the off-grade administration. The median component score as a percentage of the component's maximum score shows that the items, when appearing as a collection, were on average difficult for Pilot administration participants. Most items had item difficulty below 0.5. For DIF, only a relatively small number of items demonstrated some performance differences between student groups.

Table 13. Summary of Reliability and Difficulty for CAT Administrations.

Grade	No. of Students		Reliability			Percentage of Maximum			Item Discrimination		
	Min	Max	Min	Max	Median	Min	Max	Median	Min	Max	Median
ELA/literacy											
3	1,369	9,539	0.75	0.86	0.81	34.0	54.8	45.6	-0.20	0.70	0.48
4	1,092	7,426	0.70	0.83	0.77	34.8	54.4	44.8	-0.04	0.74	0.44
5	1,177	9,976	0.64	0.80	0.72	37.0	53.3	45.3	-0.01	0.60	0.43
6	1,278	4,915	0.60	0.80	0.72	37.3	48.3	43.0	-0.22	0.66	0.39
7	1,060	4,534	0.55	0.84	0.72	34.2	50.1	41.3	-0.25	0.74	0.41
8	491	4,331	0.53	0.79	0.69	35.1	46.4	42.4	-0.40	0.64	0.34
9	1,139	4,858	0.50	0.84	0.70	33.4	50.7	42.4	-0.42	0.71	0.38
10	507	2,838	0.64	0.81	0.72	31.4	47.1	36.8	-0.37	0.65	0.40
11	249	1,772	0.59	0.83	0.74	27.1	42.4	33.2	-0.37	0.74	0.39
Mathematics											
3	1,743	6,199	0.67	0.87	0.79	26.0	51.8	36.8	-0.09	0.73	0.53
4	1,917	4,763	0.67	0.87	0.81	15.8	48.4	36.0	0.04	0.84	0.54
5	2,062	5,116	0.74	0.86	0.83	23.7	42.5	35.6	-0.02	0.74	0.54
6	1,801	4,498	0.65	0.88	0.79	22.1	45.0	32.5	-0.24	0.76	0.50
7	893	3,642	0.62	0.84	0.79	15.6	36.0	26.1	-0.20	0.77	0.47
8	1,416	5,166	0.59	0.84	0.75	11.6	34.4	25.0	-0.25	1.00	0.44
9	705	3,527	0.58	0.76	0.63	9.9	26.6	20.9	-0.15	0.68	0.34
10	631	2,106	0.54	0.79	0.69	14.7	33.2	20.9	-0.09	0.74	0.42
11	536	2,272	0.52	0.83	0.72	10.0	28.3	18.9	-0.19	1.00	0.43

Item Flagging. After completion of the classical item analysis for the Pilot Test, poorly functioning items were flagged. The flag definition and recommendations are given in Tables 14 and 15 for selected- and constructed-response items. This information was used by content experts in reviewing the items. Tables 16 and 17 present the number of items flagged using these designations for ELA/literacy and mathematics, respectively. Many items had low item-test correlations. Prior to conducting the dimensionality study and IRT analyses, the items were reviewed by content experts in light of these statistics. After the data review, more than 75 ELA/literacy items and more than 83 mathematics items were deemed appropriate for inclusion in the dimensionality study and IRT analyses (except in grade 9, where fewer than 70 ELA/literacy items and fewer than 75 mathematics items were included).

Table 14. Description of Item Flagging for Selected-response Items.

Flag	Flag Definition	Flag Interpretation and Recommended Follow-up Actions
A	Low average item difficulty (less than 0.10).	Item is difficult. Check if the answer key is the only correct choice, if item is assessing the required content standards, and check grade-level appropriateness.
D	Proportionally more high ability students select a distractor over the answer key.	High ability students tend to choose a distractor rather than the answer key. Check if all distractors are incorrect, especially distractors with positive point-biserial correlation values.
F	Higher criterion score mean for students choosing a distractor than the mean for those choosing the answer key.	Students choosing a distractor have higher ability than students choosing the answer key. Check if all distractors are wrong, especially distractors with positive point-biserial correlation values.
H	High average item difficulty (greater than 0.95).	The item is very easy; check the grade-level appropriateness.
P	Positive distractor item point-biserial correlation.	A student with high ability level is more likely to choose this distractor than a student with low ability level. Check if the distractors with positive point biserial correlations are clearly incorrect.
R	Low item-total correlation (point-biserial correlation less than 0.30).	Item is not capable of separating high ability students from low ability students. Check if key is the only correct choice, and if item is assessing required content standards at an appropriate grade level.
V	Item more difficult in a higher-grade level.	The item is more difficult for students at a higher grade-level than compared with ones in a lower grade level. Investigate the reason for the reverse growth pattern and grade level appropriateness.
Z	Flagged by statisticians as an additional item requiring content review.	Check if key is the only correct choice, item is assessing the required content standards and check grade-level appropriateness.

Table 15. Description of Item Flagging for Constructed-response Items.

Flag	Flag Definition	Flag Interpretation and Recommended Follow-up Actions
A	Low average item difficulty (less than 0.10).	Item is difficult. Check if item is assessing the required content standards and check grade level appropriateness.
B	Percentage obtaining any score category < 3%.	Check if this score category is reasonably defined. Evaluate the need to collapse score categories and improve the scoring rubric.
C	Higher criterion score mean for students in a lower score-point category.	Higher ability students tend to get a lower score on this item than the lower ability students. Confirm reasonableness of scoring rubric.
H	High average item score (greater than 0.95).	Item is easy. Check grade level appropriateness.
R	Low item-total correlation (polyserial correlation less than 0.30).	Item is not capable of separating high ability students from low ability students. Check if item is assessing required content standards at the appropriate grade level and check the reasonableness of scoring rubric.
V	Smaller average item score at a higher-grade level.	Item more difficult for a student at a higher-grade level than for a student at a lower grade level. Investigate the reason for the reverse scoring pattern. Check grade level appropriateness.
Z	Flagged by statisticians as an additional item that needs content review.	Check if item is assessing the required content standards and check grade-level appropriateness.

Table 16. Number of Items Flagged for ELA/literacy by Selected- and Constructed-response.

Item Flags	Grade																	
	3		4		5		6		7		8		9		10		11	
	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR
A	2	2	3	2	4	3	6	2	5		2	1	2	1	5	3	10	10
B		5		3		2	1	3	3	3	1	9		3	3	7	2	18
C		2						2				8			1	1		6
D	4		4		1		2		2		9		6		4		7	
F	1								2		5		4		2		2	
H																		
N																		
O																		
P	14		17		13		28		16		33		18		20		30	
R	21	6	39	8	26	2	46	10	38	16	39	27	29	9	21	12	41	21
V	8	2	10	3	6	3	18	9	22	6	19	10	67	25	73	47	5	6
Z																		

Table 17. Number of Items Flagged for Mathematics by Selected- and Constructed-response.

Item Flags	Grade																	
	3		4		5		6		7		8		9		10		11	
	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR
A	11	9	10	9	7	13	17	18	20	25	15	17	23	13	11	22	71	89
B	4	8	5	3	4	6	8	12	10	16	9	11	21	9	4	11	37	51
C		3						1	1				1					
D	2		1				4		9		5		4		5		9	
F					1		2		2		2						4	
H	4																	
N																		
O																		
P	5		6		5		11		23		16		6		7		25	
R	5	3	8	2	8	2	21	6	33	9	37	11	39	10	19	9	67	21
V	1		4	2	14	19	21	15	17	14	12	18	51	25	45	40	5	1
Z						2												

Differential Item Functioning. In addition to classical item analyses, differential item functioning (DIF) analyses were conducted on the Pilot items. DIF analyses are used to identify those items that defined groups of students (e.g., males, females) with the same underlying level of ability that have different probabilities of answering an item correctly. Test takers are separated into relevant subgroups based on ethnicity, gender, or other demographic characteristics for DIF analyses. Then test takers in each subgroup are ranked relative to their total test score (conditioned on ability). Test takers in the focal group (e.g., females) are compared to students in the reference group (e.g., males) relative to their performance on individual items.

The following procedure is followed for DIF analysis. First, students are assigned to subgroups based on ethnicity, gender, or other demographic characteristics. It is possible to perform a DIF analysis for any two groups of students, but the “focal groups” are commonly female students or students from specified ethnic groups. For each focal group, there is a corresponding “reference group” of students who are not members of the focal group. Then students in each subgroup are ranked relative to their ability level. Students in the focal group (e.g., females) are compared to students of the same ability level in the reference group (e.g., males) relative to their performance on a designated item. A DIF analysis asks, “If we compare focal-group and reference-group students of comparable ability (as indicated by their performance on the full test), are any test questions significantly harder for one group than for the other?” If the item is differentially more difficult for an identifiable subgroup when conditioned on ability, it may be measuring something different from the intended construct to be measured. DIF statistics are used to identify items that are *potentially* functioning differentially. However, DIF-flagged items might be related to actual differences in relevant knowledge or skills (item impact) or statistical Type I errors. As a result, DIF statistics are used to identify items that are potentially functioning differentially. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences. The DIF analysis definitions are presented in Tables 18, 19, and 20.

Table 18. Definition of Focal and Reference Groups.

Category	Focal Groups	Reference Groups
Gender	Female	Male
Ethnicity	African American	White
	Asian/Pacific Islander	
	Native American/Alaska Native	
	Hispanic	
	Multiple	
Special Populations	English Learner	English Proficient
	Disability (TBD)	No disability

Table 19. DIF Categories for Selected-Response Items.

DIF Category	Flag Definition
A (negligible)	Absolute value of the MH D-DIF is not significantly different from zero, or is less than one.
B (slight to moderate)	1. Absolute value of the MH D-DIF is significantly different from zero but not from one, and is at least one; OR 2. Absolute value of the MH D-DIF is significantly different from one, but is less than 1.5. Positive values are classified as “B+” and negative values as “B-”.
C (moderate to large)	Absolute value of the MH D-DIF is significantly different from one, and is at least 1.5. Positive values are classified as “C+” and negative values as “C-”.

Table 20. DIF Categories for Constructed-Response Items.

DIF Category	Flag Definition
A (negligible)	Mantel Chi-square p -value > 0.05 or $ SMD/SD \leq 0.17$
B (slight to moderate)	Mantel Chi-square p -value < 0.05 and $ SMD/SD > 0.17$, but ≤ 0.25
C (moderate to large)	Mantel Chi-square p -value < 0.05 and $ SMD/SD > 0.25$

Statistics from two DIF detection methods were computed. The Mantel-Haenszel procedure (Mantel & Haenszel, 1959) and the standardization procedure (Dorans & Kulick, 1983, 1986). As part of the Mantel-Haenszel procedure, the statistic described by Holland & Thayer (1988), known as MH D-DIF, was used. This statistic is expressed as the difference between members of the focal group (e.g., female, Asian, African American, Hispanic, and Native American) and members of the reference group (e.g., males and White) after conditioning on ability (e.g., total test score). This statistic is reported on the delta scale, which is a normalized transformation of item difficulty (p -value) with a mean of 13 and a standard deviation of four. Negative MH D-DIF statistics favor the reference group, and positive values favor the focal group. The classification logic used for flagging items is based on a combination of absolute differences and significance testing. Items that are not statistically significantly different based on the MH D-DIF ($p > 0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. For items where the statistical test indicates significant differences ($p < 0.05$), the effect size is used to determine the direction and severity of the DIF. Selected-response items were classified into DIF categories of A, B, and C, as described in Table 19.

For polytomous items (i.e., constructed-response), the Mantel-Haenszel procedure was executed where item categories are treated as integer scores and a chi-square test was carried out with one degree of freedom. The standardized mean difference (SMD) (Zwick, Donoghue, & Grima, 1993) was used in conjunction with the Mantel chi-square statistic. The standardized mean difference compares item performance of two subpopulations adjusting for differences in the distributions of the two subpopulations (Dorans & Kulick, 1986; Dorans & Schmitt, 1991/1993). A positive value for

SMD reflects DIF in favor of the focal group. The SMD can be divided by the total standard deviation to obtain a measure of the effect size. A negative SMD value shows that the question is more difficult for the focal group, whereas a positive value indicates that it is more difficult for the reference group. The classification logic for polytomous items is based on a combination of absolute differences and significance testing, as shown in Table 20.

Tables 21 to 24 show the number of items flagged for C DIF for ELA/literacy and mathematics in grades 3 to 11. Note that the items flagging also reflect items that were administered off-grade for vertical linking. A relatively small number of items were flagged for significant level of DIF in the Pilot Test.

Table 21. Number of DIF Items Flagged by Item Type and Subgroup (ELA/literacy, Grades 3 to 7).

DIF Comparison		Grade 3		Grade 4			Grade 5			Grade 6			Grade 7							
		3		4		5	4		5	6	5		6	7	6		7	8		
		SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	
Female vs. Male	C+						1							1					4	
	C-													1					1	
Asian vs. White	C+	4	2	1	1		5	2			2	1			2	1				1
	C-	4		3			3	1							2				1	1
Black vs. White	C+	1																		
	C-	1					1				1	1							1	1
Hispanic vs. White	C+				1					1					1			1	1	
	C-	2					1				2				1	1		1		1
Native American vs. White	C+																			
	C-																			

Table 22. Number of DIF Items Flagged by Item Type and Subgroup (ELA/literacy, Grades 8 to 11).

DIF Comparison		Grade 8						Grade 9						Grade 10						Grade 11					
		7		8		9		8		9		10		9		10		11		9		10		11	
		SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR
Female vs. Male	C+			1	1			1							1										
	C-			1									1												
Asian vs. White	C+		1						1	1						2									
	C-		1												1	1									
Black vs. White	C+																						1		
	C-	2		1											1								1	2	
Hispanic vs. White	C+																								
	C-	1		2	1											2								1	
Native American vs. White	C+																								
	C-																								

Table 23. Number of C DIF Items Flagged by Item Type and Subgroup (Mathematics, Grades 3 to 7).

DIF Comparison		Grade 3			Grade 4			Grade 5			Grade 6			Grade 7 Math														
		3	4		3	4	5	4	5	6	5	6	7	6	7	8												
		SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR											
Female vs. Male	C+																1											
	C-		1			1				1				1				1										
Asian vs. White	C+	8		2	1		3	4			1			3	1			1	5			1		3	1			
	C-		1	1	2		1							1	1				1	1		2					1	
Black vs. White	C+			1										1	2													
	C-	1													1							3						
Hispanic vs. White	C+	2	1				1	5						6					1	1						2		1
	C-						1		2	1	1			4	1				1	1						1		1
Native American vs. White	C+																											
	C-																											

Table 24. Number of C DIF Items Flagged by Item Type and Subgroup (Mathematics, Grades 8 to 11).

DIF Comparison	Grade 8						Grade 9						Grade 10						Grade 11						
	7		8		9		8		9		10		9		10		11		9		10		11		
	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	
Female vs. Male	C+		1																						
	C-														1							1			
Asian vs. White	C+		3	3					4	6					6	3			1	1	1		14	6	
	C-			1						5					1	4							3	3	
Black vs. White	C+																						3		
	C-		2			1									3										
Hispanic vs. White	C+		2								2				1	1	1		1						
	C-		1						2						4	2	1							1	
Native American vs. White	C+																								
	C-																								

Dimensionality Study

Before undertaking the Pilot calibration and scaling, Smarter Balanced sought insight concerning test dimensionality that will affect the IRT scaling design and ultimately the composite score that denotes overall student proficiency. This section describes the procedures used and outcomes pertaining to the dimensionality study based on the Pilot Test administration.

Rationale and Approach

As a factor analytic approach, multidimensional IRT (MIRT) was used to examine the dimensional structure. The first component to evaluate pertains to assessing the degree to which essential unidimensionality is met within a single grade and content area. The second aspect concerns the degree of invariance in the construct across two adjacent grades. Both criteria can be met or violated. A multidimensional composite of scores can be identified, but it should be consistent across grades in order to best support unidimensional scoring (Reckase, Ackerman, & Carlson, 1988).

The MIRT approach has a number of advantages. First, MIRT is very close to the more familiar unidimensional IRT scaling techniques. This approach can utilize familiar unidimensional models as a starting point for model comparison. The baseline model is the unidimensional case with which other candidate models can be compared. Second, from a practical perspective the sparse data matrix used for unidimensional scaling can be leveraged without the need to create other types of data structures (i.e., covariance matrices). In addition, further insight can be obtained with respect to the vertical scaling. Using exploratory approaches, the shift in the nature of the construct across levels can be inspected across adjacent grade levels. The factor analysis approach is both exploratory and confirmatory in nature. Simple structure refers to items loading on a single specified factor in a confirmatory approach. Complex structure refers to freeing individual items to load on multiple factors using an exploratory approach. By using an exploratory approach, the dimensional structure can be evaluated graphically using item vectors. Global fit comparisons were undertaken to arrive at a preferred model to determine the scaling approach and the resulting score reporting. Both the overall model test fit (e.g., Bayesian Information Criterion) and graphical depictions using item vectors can be utilized in evaluating the factor structure. Another focus of investigation is the claim structure for ELA/literacy and mathematics.

Factor Models

ELA/literacy and mathematics are scaled using multidimensional IRT using grades 3 to 11. Due to the mixed format data for the Smarter Balanced assessments containing selected- and constructed-response items, both unidimensional and multidimensional versions of the 2PL (M-2PL) and 2PPC (M-2PPC) IRT scaling models were used. Unidimensional and multidimensional models were compared using a number of model fit measures and graphical methods.

The analysis consisted of two phases. In the first phase, we examined each grade and content area separately (i.e., dimensionality within grade). In the second phase, we investigated the dimensionality of two adjacent grade levels that contained unique grade specific items and common “vertical” linking items. The first step is a within-grade scaling. The results of the within-grade analysis were evaluated before proceeding to the across grades vertical linking. In the second phase, all items across two grades were estimated concurrently where a multigroup model was implemented (Bock & Zimowski, 1997). The adjacent-grade levels have vertical linking items in common across grade groups. In both types of analysis, the choice in a candidate model can be assessed using the Akaike Inference Criterion (AIC) measures of global fit, the difference chi-square, and by vector based methods (i.e., graphical) as well as item cluster techniques.

Unidimensional Models

The baseline model for comparison is the unidimensional version. Since unidimensional models are constrained versions of multidimensional ones, MIRT software can be used to estimate them as well. The unidimensional versions were implemented with the same calibration software to afford a similar basis of comparison with other multidimensional models. Comparisons of model fit were with the unidimensional model, which is the most parsimonious one.

Multidimensional Models

Exploratory Models (Complex Structure). The exploratory models “let the data speak” by adopting a complex structure in which items are permitted to load freely on multiple factors. Consistent with the approach outlined for unidimensional models, in the first phase we examined each grade and content area separately (within-grade configuration). The next step was to concurrently scale two adjacent-grade test levels and examine the resulting structure. Using a two-dimensional exploratory model, item vectors can be evaluated graphically. An important aspect was to note the direction of measurement of items and the overall composite vector (Reckase, 1985). If the same composite of factors is consistently present across grade levels, this supports the use of unidimensional IRT scaling approaches and the construction of the vertical scale. By contrast, if distinct groups of items exist and are inconsistent across grades, this would argue against the adoption of a vertical scale.

Confirmatory Models (Simple Structure). Confirmatory models specify the loading of items on the factors, referred to as simple structure here, according to defined criteria. Two types of confirmatory models were investigated.

- A. *Claims.* This model evaluates factors corresponding to the claims for each content area according to the Pilot Test blueprints (see Tables 2, 3 and 4). For example, four claims for mathematics are Concepts & Procedures using Domains 1 and 2, Problem Solving/Modeling, and Communicating & Reasoning. A four-factor model also results in ELA/literacy consisting of Reading, Writing, Speaking/Listening, and Research.
- B. *Bifactor Model.* A bifactor model is used in which an overall factor is proposed along with two or more minor ones. The minor factors will correspond to the claim structure at each grade. A depiction of the bifactor model is given in Figure 2, consisting of a major factor and minor ones shown as claims.

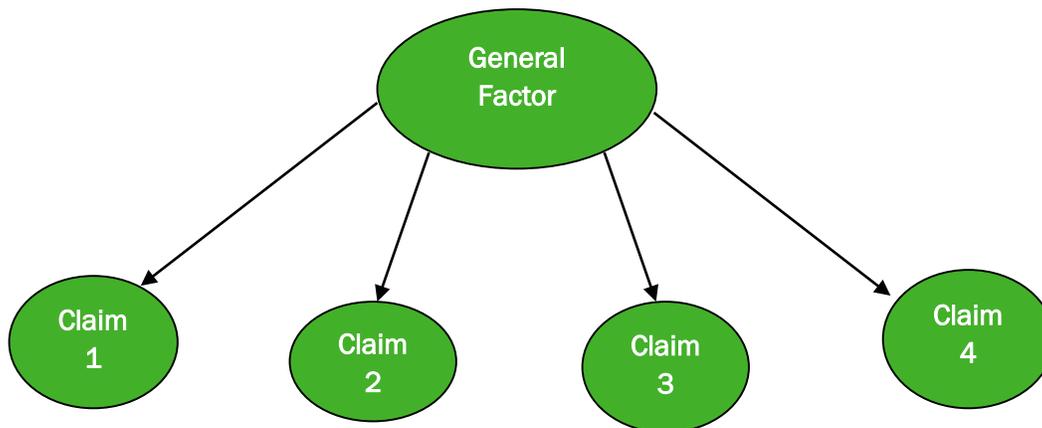


Figure 2. An Example of the Bifactor Model with Four Minor Factors Corresponding to Claims.

In total, four different models were evaluated for each content area, both within and across grades. The model and analysis configuration is summarized in Table 25 for the within-grade analysis and the across-grade configurations that show the number of MIRT models implemented across grades and content areas.

Table 25. Summary of MIRT Analysis Configuration Showing Number of Content, Grades and MIRT Models.

Model	Configuration	Content Areas	Grades	Total
Unidimensional				
	Within grade	2	9	18
	Across grades	2	8	16
Multidimensional				
Exploratory	Within grade	2	9	18
	Across grades	2	8	16
Claim Structure	Within grade	2	9	18
	Across grades	2	8	16
Bifactor	Within grade	2	9	18
	Across grades	2	8	16
Total MIRT Models (Runs)				170

MIRT Scaling Models

With mixed data present in the Pilot Test, different types of IRT scaling models must be chosen. For SR items, the two-parameter logistic (2PL) model was used or the M-2PL (McKinley & Reckase, 1983a) in the case of the multidimensional version. For CR items that include all polytomous data, the two-parameter partial-credit model (2PPC) was used. Likewise, for the dimensionality analysis, the multidimensional two-parameter partial-credit model (M-2PPC) was used (Yao & Schwarz, 2006). The multidimensional models used are compensatory in nature since high values for one theta (factor) can balance or help compensate for low values in computing the probability of a response to an item for a student. The MIRT models chosen for the dimensionality analysis correspond to unidimensional models used for horizontal and vertical scaling of the Pilot Test. The M-2PL model for selected response is

$$P_{ij} = 1 - \frac{1}{1 + e^{\bar{\beta}_{2j} \bar{\theta}_i + \beta_{\delta j}}} = \frac{1}{1 + e^{-\bar{\beta}_{2j} \bar{\theta}_i - \beta_{\delta j}}}$$

where $\vec{\beta}_{2j} = (\vec{\beta}_{2j1} \dots \vec{\beta}_{2jD})$ is a vector of dimension D corresponding to items' discrimination parameters, $\beta_{\delta j}$ is a scale difficulty parameter, and $\vec{\beta}_{2j} \square \vec{\theta}_i = \sum_{l=1}^D \beta_{2jl} \theta_{il}$. For polytomously scored items, the probability of a response $k-1$ for a test taker with ability $\vec{\theta}_i$ is given by the multidimensional version of the 2-PPC model (Yao & Schwarz, 2006):

$$P_{ijk} = P(x_{ij} = k - 1 | \vec{\theta}_i \vec{\beta}_j) = \frac{e^{(k-1)\vec{\beta}_{2j} \square \vec{\theta}_i - \sum_{t=1}^k \beta_{\delta jt}}}{\sum_{m=1}^{K_j} e^{(m-1)\vec{\beta}_{2j} \square \vec{\theta}_i - \sum_{t=1}^m \beta_{\delta jt}}},$$

where $X_{ij=0, \dots, K_j-1}$ is the response of test taker i to item j , $\beta_{\delta jk}$ for $k = 1, 2, \dots, K_j$ are threshold parameters, $\beta_{\delta j1} = 0$, and K_j is the number of response categories for the j^{th} item.

Software and System Requirements

A wide variety of scaling models are implemented by BMIRT necessary for scaling mixed item types. The program also produces model fit and multigroup (i.e., across-grade) analysis. The BMIRT program (Yao, 2003) implements a full Bayesian approach to parameter estimation that uses the Metropolis-Hastings algorithm. Using a batch file approach to implement the program permits the analysis of many grades and content areas efficiently. The R package (Weeks, 2010) **PLINK** performs multidimensional linking and other types of functions such as plotting of item characteristic curves. A scaling approach is needed that can implement models associated with mixed item types and one that makes provisions for missing data “not presented” by design. This “not-presented” or “not-reached” option is necessary here since any student by design only took a very small subset of the total available items. To be practical, the factor analysis needed to use the same data structures used for the traditional unidimensional IRT modeling.

For parameter estimation, 1,000 Markov chain Monte Carlo (MCMC) iterations were used with 250 discarded for the MCMC burn-in. The resulting item parameters were then used as start values for another 1,000 MCMC cycles; 250 were discarded from these iterations as well. These second sets of iterations were used to compute the final parameter estimates. Note that 0.4 was used for the covariance for the prior ability functions (abilityPriorCov). Values of 0.0 corresponding to no relationship between factors and 0.8 indicating high correlations between factors were also evaluated. The difference in fit using these two other values was very small compared with the covariance of 0.4. BMIRT program defaults that were used for other priors or proposal functions.

Evaluation of the Number and Types of Dimensions and MIRT Item Statistics

A primary method for evaluating models is to use overall test fit indices. The Bayesian and Akaike Information Criterion (Akaike, 1973; Schwarz, 1978) provided by **BMIRT** was used where

$$BIC_k = G_k^2 + 2 \log(N) df_k$$

$$AIC_k = G_k^2 + 2df_k$$

where G_k^2 is the likelihood and $2 \log(N) df_k$ and $2df_k$ are penalties imposed for adding extra parameters to the model. These fit statistics can be used to compare either nested or non-nested models. Lower values of AIC and BIC indicate a better fitting candidate model. A preferred factor structure results when it demonstrates the minimum fit value among several competing models. This permits comparison of model fit between unidimensional and multidimensional versions. For the comparison of model fit within a grade, the difference chi-square, and the ratio of the chi-square to the difference in degrees of freedom are also presented for ELA/literacy and mathematics. For the

difference chi-square, all comparisons were with the unidimensional case as the base model. Since MCMC methods are used, BMIRT considers both items and student ability in the likelihood. Graphical evaluation of the item vectors and clustering of angle measures were also performed.

Despite considerable advances in the estimation of a variety of complex models, no clear criteria exist for model acceptance. Several criteria were evaluated to determine if the expected inferences are supported. This process of model choice is somewhat judgmental. To warrant the expense and operational complications involved in implementing a multidimensional scaling model, the preponderance of information would need to demonstrate the data are strongly multidimensional and that this multidimensionality varies over grades.

In Tables 26 and 27, AIC, BIC, the likelihood, degrees of freedom (*df*), difference chi-square, its degrees of freedom, and the ratio of the difference chi-square to its degrees of freedom are given. The degrees of freedom reflect both items and students (i.e., θ). The difference chi-square compares the unidimensional case with the other models. These tables show the overall fit by grade configuration (within grade). They show the fit measures for the unidimensional, exploratory, claim scores, and bifactor models. The second set of global fit measures in Tables 28 and 29 show the across (adjacent) grade analysis. The measures for overall fit (across adjacent grades) are given for each grade separately. Based on the comparatively low values of AIC, the unidimensional model is consistently the preferred model.

For example, using grade 3 ELA/literacy, the value of AIC for the unidimensional model was 1,580,927, which is lower than the values for the exploratory, claim scores, and bifactor models. No model fit the data particularly well, possibly due to student sample size. The difference chi-square suggests that no model improved over the unidimensional one. For the across-grade fit that contained vertical linking items, the unidimensional model was also substantiated. The comparative fit across-grade models followed the same pattern as the within-grade analysis.

Table 26. Models and Fit Measures for ELA/literacy Within Grade.

Grade	Model	AIC	Likelihood	<i>df</i>	Difference Chi-square	Difference <i>df</i>	Ratio X^2/df
3	Unidimensional	1,580,927	-748,655	41,809			
	Exploratory	1,637,492	-735,518	83,228	26,274	41,419	0.634
	Claim Scores	1,736,151	-702,006	166,069	93,298	124,260	0.751
	Bifactor	1,847,184	-716,101	207,491	65,108	165,682	0.393
4	Unidimensional	1,671,889	-785,799	50,145			
	Exploratory	1,743,604	-771,915	99,887	27,768	49,742	0.558
	Claim Scores	1,874,179	-737,715	199,374	96,168	149,229	0.644
	Bifactor	2,003,755	-752,758	249,119	66,082	198,974	0.332
5	Unidimensional	1,269,024	-584,728	49,784			
	Exploratory	1,338,209	-569,872	99,233	29,712	49,449	0.601
	Claim Scores	1,471,467	-537,599	198,134	94,258	148,350	0.635
	Bifactor	1,600,465	-552,647	247,586	64,162	197,802	0.324
6	Unidimensional	1,422,993	-661,524	49,972			
	Exploratory	1,500,371	-650,603	99,583	21,842	49,611	0.440
	Claim Scores	1,639,063	-620,724	198,808	81,600	148,836	0.548
	Bifactor	1,763,784	-633,470	248,422	56,108	198,450	0.283
7	Unidimensional	1,310,456	-610,484	44,744			
	Exploratory	1,372,121	-596,958	89,102	27,052	44,358	0.610
	Claim Scores	1,488,947	-566,652	177,821	87,664	133,077	0.659
	Bifactor	1,605,914	-580,775	222,182	59,418	177,438	0.335

Grade	Model	AIC	Likelihood	<i>df</i>	Difference Chi-square	Difference <i>df</i>	Ratio χ^2/df
8	Unidimensional	1,282,613	-599,857	41,450			
	Exploratory	1,344,545	-589,766	82,506	20,182	41,056	0.492
	Claim Scores	1,457,239	-563,999	164,621	71,716	123,171	0.582
	Bifactor	1,561,028	-574,834	205,680	50,046	164,230	0.305
9	Unidimensional	723,096	-335,611	25,937			
	Exploratory	760,617	-328,737	51,572	13,748	25,635	0.536
	Claim Scores	835,337	-314,823	102,845	41,576	76,908	0.541
	Bifactor	898,965	-320,999	128,483	29,224	102,546	0.285
10	Unidimensional	486,630	-226,999	16,316			
	Exploratory	511,248	-223,314	32,310	7,370	15,994	0.461
	Claim Scores	552,276	-211,837	64,301	30,324	47,985	0.632
	Bifactor	597,408	-218,406	80,298	17,186	63,982	0.269
11	Unidimensional	724,846	-342,958	19,465			
	Exploratory	745,309	-334,360	38,294	17,196	18,829	0.913
	Claim Scores	795,682	-321,886	75,955	42,144	56,490	0.746
	Bifactor	837,513	-323,969	94,787	37,978	75,322	0.504

Table 27. Models and Fit Measures for Mathematics Within Grade.

Grade	Model	AIC	Likelihood	<i>df</i>	Chi-square	<i>df</i>	χ^2/df
3	Unidimensional	1,243,707	-581,019	40,835			
	Exploratory	1,293,666	-565,528	81,305	30,982	40,470	0.766
	Claim Scores	1,415,106	-545,305	162,248	71,428	121,413	0.588
	Bifactor	1,521,203	-557,881	202,721	46,276	161,886	0.286
4	Unidimensional	1,361,780	-636,775	44,115			
	Exploratory	1,420,052	-622,197	87,829	29,156	43,714	0.667
	Claim Scores	1,560,890	-605,185	175,260	63,180	131,145	0.482
	Bifactor	1,671,350	-616,698	218,977	40,154	174,862	0.230
5	Unidimensional	1,614,121	-760,281	46,780			
	Exploratory	1,664,992	-739,327	93,169	41,908	46,389	0.903
	Claim Scores	1,818,934	-723,517	185,950	73,528	139,170	0.528
	Bifactor	1,919,462	-727,389	232,342	65,784	185,562	0.355
6	Unidimensional	1,245,624	-580,395	42,417			
	Exploratory	1,301,437	-566,257	84,462	28,276	42,045	0.673
	Claim Scores	1,444,817	-553,853	168,555	53,084	126,138	0.421
	Bifactor	1,540,013	-559,403	210,603	41,984	168,186	0.250
7	Unidimensional	1,123,242	-520,561	41,060			
	Exploratory	1,186,090	-511,323	81,722	18,476	40,662	0.454
	Claim Scores	1,318,147	-496,025	163,049	49,072	121,989	0.402
	Bifactor	1,419,308	-505,940	203,714	29,242	162,654	0.180
8	Unidimensional	1,182,794	-546,363	45,034			

Grade	Model	AIC	Likelihood	<i>df</i>	Chi-square	<i>df</i>	χ^2/df
	Exploratory	1,243,004	-531,827	89,675	29,072	44,641	0.651
	Claim Scores	1,398,606	-520,343	178,960	52,040	133,926	0.389
	Bifactor	1,496,807	-524,800	223,604	43,126	178,570	0.242
9	Unidimensional	516,180	-238,530	19,560			
	Exploratory	536,809	-229,557	38,848	17,946	19,288	0.930
	Claim Scores	612,138	-228,642	77,427	19,776	57,867	0.342
	Bifactor	648,848	-227,706	96,718	21,648	77,158	0.281
10	Unidimensional	367,643	-171,071	12,750			
	Exploratory	382,795	-166,223	25,175	9,696	12,425	0.780
	Claim Scores	425,940	-162,942	50,028	16,258	37,278	0.436
	Bifactor	454,729	-164,909	62,456	12,324	49,706	0.248
11	Unidimensional	505,284	-228,087	24,555			
	Exploratory	543,836	-223,388	48,530	9,398	23,975	0.392
	Claim Scores	630,439	-218,736	96,483	18,702	71,928	0.260
	Bifactor	683,748	-221,413	120,461	13,348	95,906	0.139

Table 28. Models and Fit Measures for ELA/literacy Across Adjacent Grades.

Grades	Model	Group	AIC	BIC	Likelihood	df
3 to 4	Unidimensional	Overall	3,255,135	4,123,262	-1,535,423	92,145
		3	1,582,366	1,944,195	-749,267	41,916
		4	1,672,770	2,115,573	-786,156	50,229
	Exploratory	Overall	3,381,393	5,108,951	-1,507,330	183,367
		3	1,637,806	2,357,468	-735,534	83,369
		4	1,743,587	2,625,139	-771,796	99,998
	Claim Scores	Overall	3,703,214	7,149,559	-1,485,804	365,803
		3	1,734,620	3,169,972	-701,032	166,278
		4	1,968,594	3,727,544	-784,772	199,525
	Bifactor	Overall	4,057,828	8,363,605	-1,571,889	457,025
		3	1,850,234	3,643,444	-717,383	207,734
		4	2,207,595	4,405,267	-854,506	249,291
4 to 5	Unidimensional	Overall	2,942,383	3,894,243	-1,371,059	100,132
		4	1,672,823	2,115,732	-786,170	50,241
		5	1,269,560	1,709,105	-584,889	49,891
	Exploratory	Overall	3,084,751	4,980,134	-1,342,989	199,387
		4	1,742,772	2,624,456	-771,373	100,013
		5	1,341,979	2,217,475	-571,616	99,374
	Claim Scores	Overall	3,446,338	7,228,691	-1,325,280	397,889
		4	1,870,656	3,629,915	-735,768	199,560
		5	1,575,682	3,322,982	-589,512	198,329
	Bifactor	Overall	3,837,936	8,563,813	-1,421,824	497,144
		4	2,004,632	4,202,692	-752,981	249,335
		5	1,833,305	4,016,530	-668,843	247,809
5 to 6	Unidimensional	Overall	2,693,333	3,643,487	-1,246,701	99,966

Grades	Model	Group	AIC	BIC	Likelihood	df
		5	1,269,703	1,709,283	-584,956	49,895
		6	1,423,631	1,864,913	-661,744	50,071
	Exploratory	Overall	2,842,088	4,734,451	-1,221,948	199,096
		5	1,342,161	2,217,736	-571,698	99,383
		6	1,499,927	2,378,712	-650,251	99,713
	Claim Scores	Overall	3,202,642	6,979,344	-1,203,973	397,348
		5	1,468,207	3,215,798	-535,741	198,362
		6	1,734,435	3,488,126	-668,231	198,986
	Bifactor	Overall	3,594,141	8,313,051	-1,300,592	496,478
		5	1,603,426	3,787,039	-553,860	247,853
		6	1,990,715	4,181,881	-746,732	248,625
	6 to 7	Unidimensional	Overall	2,734,953	3,632,171	-1,272,554
6			1,423,768	1,865,024	-661,816	50,068
7			1,311,185	1,701,494	-610,737	44,855
Exploratory		Overall	2,869,962	4,656,033	-1,246,020	188,961
		6	1,498,621	2,377,370	-649,602	99,709
		7	1,371,341	2,147,975	-596,419	89,252
Claim Scores		Overall	3,228,796	6,792,497	-1,237,369	377,029
		6	1,635,272	3,389,033	-618,642	198,994
		7	1,593,524	3,142,710	-618,727	178,035
Bifactor		Overall	3,580,506	8,033,060	-1,319,186	471,067
		6	1,766,238	3,957,518	-634,481	248,638
		7	1,814,268	3,749,752	-684,705	222,429
7 to 8	Unidimensional	Overall	2,595,184	3,403,519	-1,211,203	86,389
		7	1,311,172	1,701,351	-610,746	44,840

Grades	Model	Group	AIC	BIC	Likelihood	df	
	Exploratory	8	1,284,012	1,642,351	-600,457	41,549	
		Overall	2,712,594	4,320,731	-1,184,431	171,866	
		7	1,368,710	2,145,152	-595,125	89,230	
	Claim Scores	8	1,343,883	2,056,577	-589,306	82,636	
		Overall	3,037,992	6,245,658	-1,176,184	342,812	
		7	1,488,031	3,037,025	-566,002	178,013	
	Bifactor	8	1,549,961	2,971,268	-610,181	164,799	
		Overall	3,357,227	7,364,695	-1,250,324	428,289	
		7	1,608,923	3,544,206	-582,055	222,406	
	8 to 9	Unidimensional	8	1,748,304	3,523,941	-668,269	205,883
			Overall	2,007,224	2,623,188	-935,996	67,616
			9	723,748	936,000	-335,843	26,031
Exploratory		8	1,283,475	1,642,125	-600,153	41,585	
		Overall	2,106,595	3,330,799	-918,914	134,384	
		9	760,054	1,181,582	-328,330	51,697	
Claim Scores		8	1,346,541	2,059,674	-590,583	82,687	
		Overall	2,355,982	4,796,591	-910,079	267,912	
		9	901,573	1,741,563	-347,769	103,018	
Bifactor		8	1,454,408	2,876,536	-562,310	164,894	
		Overall	2,592,106	5,640,955	-961,373	334,680	
		9	1,027,475	2,076,717	-385,057	128,681	
9 to 10	Unidimensional	8	1,283,475	1,642,125	-600,153	41,585	
		Overall	1,211,766	1,578,699	-563,413	42,470	
		9	723,694	935,849	-335,828	26,019	
		10	488,071	614,499	-227,585	16,451	

Grades	Model	Group	AIC	BIC	Likelihood	<i>df</i>
	Exploratory	Overall	1,274,759	2,001,981	-553,209	84,171
		9	761,797	1,183,186	-329,218	51,680
		10	512,962	762,658	-223,990	32,491
	Claim Scores	Overall	1,417,427	2,865,158	-541,149	167,565
		9	833,651	1,673,535	-313,821	103,005
		10	583,776	1,079,925	-227,328	64,560
	Bifactor	Overall	1,561,259	3,369,279	-571,364	209,266
		9	899,379	1,948,523	-321,021	128,669
		10	661,880	1,281,275	-250,343	80,597
10 to 11	Unidimensional	Overall	1,213,870	1,518,346	-570,955	35,980
		9	487,682	613,971	-227,408	16,433
		10	726,188	879,570	-343,547	19,547
	Exploratory	Overall	1,261,019	1,860,730	-559,642	70,868
		9	513,973	763,477	-224,520	32,466
		10	747,047	1,048,380	-335,121	38,402
	Claim Scores	Overall	1,375,980	2,566,093	-547,354	140,636
		9	552,391	1,048,348	-211,660	64,535
		10	823,589	1,420,740	-335,694	76,101
	Bifactor	Overall	1,485,638	2,970,985	-567,295	175,524
		9	598,001	1,217,196	-218,430	80,571
		10	887,637	1,632,715	-348,865	94,953

Table 29. Models and Fit Measures for Mathematics Across Adjacent Grades.

Grades	Model	Group	AIC	BIC	Likelihood	<i>df</i>
3 to 4	Unidimensional	Overall	2,609,055	3,402,805	-1,219,552	84,976
		3	1,245,590	1,597,234	-581,946	40,849
		4	1,363,465	1,746,733	-637,606	44,127
	Exploratory	Overall	2,724,905	4,305,109	-1,193,282	169,171
		3	1,299,575	1,999,652	-568,463	81,325
		4	1,425,330	2,188,322	-624,819	87,846
	Claim Scores	Overall	3,024,199	6,177,237	-1,174,546	337,553
		3	1,417,002	2,813,971	-546,221	162,280
		4	1,607,197	3,129,542	-628,326	175,273
	Bifactor	Overall	3,226,816	7,166,308	-1,191,660	421,748
		3	1,521,641	3,267,069	-558,061	202,759
		4	1,705,175	3,607,218	-633,599	218,989
4 to 5	Unidimensional	Overall	2,981,009	3,836,472	-1,399,584	90,921
		4	1,364,880	1,748,122	-638,316	44,124
		5	1,616,129	2,025,368	-761,268	46,797
	Exploratory	Overall	3,086,050	4,789,382	-1,361,990	181,035
		4	1,427,225	2,190,182	-625,770	87,842
		5	1,658,825	2,473,795	-736,220	93,193
	Claim Scores	Overall	3,436,284	6,835,279	-1,356,887	361,255
		4	1,564,470	3,086,883	-606,954	175,281
		5	1,871,814	3,498,151	-749,933	185,974
	Bifactor	Overall	3,637,368	7,884,233	-1,367,315	451,369
		4	1,673,654	3,575,809	-617,825	219,002
		5	1,963,715	3,995,757	-749,490	232,367
5 to 6	Unidimensional	Overall	2,867,813	3,705,554	-1,344,691	89,215

Grades	Model	Group	AIC	BIC	Likelihood	df
		5	1,617,910	2,027,052	-762,169	46,786
		6	1,249,902	1,616,768	-582,522	42,429
	Exploratory	Overall	2,975,243	4,643,466	-1,309,964	177,657
		5	1,669,399	2,484,237	-741,521	93,178
		6	1,305,844	2,036,300	-568,443	84,479
		Claim Scores	Overall	3,309,818	6,638,931	-1,300,376
	5		1,823,344	3,449,602	-725,707	185,965
		6	1,486,474	2,944,012	-574,669	168,568
		Bifactor	Overall	3,497,384	7,656,979	-1,305,717
	5		1,920,776	3,952,756	-728,028	232,360
		6	1,576,608	3,397,710	-577,689	210,615
		6 to 7	Unidimensional	Overall	2,373,141	3,151,522
6	1,247,380			1,614,177	-581,269	42,421
7	1,125,761			1,479,486	-521,812	41,068
Exploratory	Overall		2,494,563	4,044,090	-1,081,079	166,202
	6		1,305,116	2,035,476	-568,090	84,468
	7		1,189,447	1,893,434	-512,990	81,734
Claim Scores	Overall		2,803,345	5,895,090	-1,070,052	331,620
	6		1,448,121	2,905,634	-555,496	168,565
	7		1,355,223	2,759,640	-514,557	163,055
Bifactor	Overall		2,985,167	6,848,058	-1,078,251	414,333
	6		1,546,176	3,367,277	-562,473	210,615
	7		1,438,991	3,193,645	-515,778	203,718
7 to 8	Unidimensional	Overall	2,310,404	3,115,824	-1,069,098	86,104
		7	1,125,531	1,479,238	-521,699	41,066

Grades	Model	Group	AIC	BIC	Likelihood	df	
	Exploratory	8	1,184,873	1,576,994	-547,399	45,038	
		Overall	2,432,489	4,035,883	-1,044,833	171,412	
		7	1,189,292	1,893,253	-512,915	81,731	
	Claim Scores	8	1,243,198	2,024,001	-531,918	89,681	
		Overall	2,758,373	5,957,640	-1,037,167	342,020	
		7	1,322,333	2,726,827	-498,103	163,064	
	Bifactor	8	1,436,040	2,994,112	-539,064	178,956	
		Overall	2,946,770	6,944,012	-1,046,057	427,328	
		7	1,424,973	3,179,747	-508,754	203,732	
	8 to 9	Unidimensional	8	1,521,798	3,468,526	-537,303	223,596
			Overall	1,702,770	2,288,505	-786,775	64,610
			9	518,613	672,584	-239,736	19,570
Exploratory		8	1,184,158	1,576,296	-547,039	45,040	
		Overall	1,785,655	2,951,024	-764,280	128,547	
		9	540,168	845,931	-231,221	38,863	
Claim Scores		8	1,245,487	2,026,316	-533,059	89,684	
		Overall	2,027,808	4,352,371	-757,491	256,413	
		9	626,486	1,235,746	-235,805	77,438	
Bifactor		8	1,401,321	2,959,559	-521,686	178,975	
		Overall	2,160,865	5,065,062	-760,082	320,350	
		9	656,270	1,417,297	-231,407	96,728	
9 to 10	Unidimensional	8	886,249	1,156,660	-410,798	32,326	
		9	516,989	670,991	-238,920	19,574	
		10	369,260	463,987	-171,878	12,752	

Grades	Model	Group	AIC	BIC	Likelihood	<i>df</i>
	Exploratory	Overall	922,509	1,458,271	-397,207	64,047
		9	537,339	843,148	-229,800	38,869
		10	385,170	572,203	-167,407	25,178
	Claim Scores	Overall	1,052,414	2,118,811	-398,726	127,481
		9	614,092	1,223,540	-229,584	77,462
		10	438,322	809,885	-169,142	50,019
	Bifactor	Overall	1,110,102	2,441,850	-395,849	159,202
		9	649,857	1,411,136	-228,168	96,760
		10	460,246	924,092	-167,681	62,442
10 to 11	Unidimensional	Overall	876,674	1,194,819	-400,926	37,411
		10	369,223	464,151	-171,832	12,779
		11	507,452	706,648	-229,094	24,632
	Exploratory	Overall	933,765	1,561,840	-393,026	73,856
		10	388,458	575,789	-169,011	25,218
		11	545,306	938,637	-224,015	48,638
	Claim Scores	Overall	1,072,896	2,320,763	-389,710	146,738
		10	428,775	800,932	-164,288	50,099
		11	644,121	1,425,630	-225,421	96,639
	Bifactor	Overall	1,141,252	2,699,050	-387,443	183,183
		10	454,125	918,706	-164,521	62,541
		11	687,127	1,662,746	-222,922	120,642

MIRT Item Statistics and Graphs

The Reckase, Martineau, & Kim (2000) item vector approach was used to evaluate the characteristics of exploratory models using complex structure. Three primary MIRT item characteristics were computed corresponding to discrimination, direction, and difficulty; they are presented graphically (Reckase, 1985). The magnitude given by the length of the vector corresponds to its discriminating power

$$\sqrt{\mathbf{a}'\mathbf{a}}.$$

The angle measure of the vector with each axis is

$$\alpha_{ij} = \arccos \frac{a_{ij}}{\sqrt{\mathbf{a}'\mathbf{a}}},$$

where a_{ij} is the j -th element of the vector of item discriminations for item i . In order to obtain degrees, the angle measure in radians is multiplied by $180/\pi$. If an item measured only the primary trait, the angle measure α would be 0; whereas if the item measured the primary factor and secondary factor equally, the α would be 45° . The quadrant of the plot in which an item resides roughly corresponds to its difficulty. The multidimensional difficulty is

$$\frac{-b_i}{\sqrt{\mathbf{a}'\mathbf{a}}},$$

where b_i is the location or scalar item parameter related to item difficulty.

A composite directional vector can be computed using the matrix of discriminations \mathbf{a} and then computing the eigenvalues for $\mathbf{a}'\mathbf{a}$. Each diagonal value in the matrix is the sum of the squared \mathbf{a} -elements for each ability dimension of the matrix. The off-diagonal values are the sums of the cross products of the \mathbf{a} -elements from different dimensions. The eigenvector that corresponds to the largest eigenvalue is eigenvector one. The sum of the squared elements of the eigenvector is equal to one, and these elements have the properties of direction cosines. The direction cosines give the orientation of the reference composite with respect to the coordinate axes of the ability space. The angle between the reference composite and the coordinate axes can be determined by taking the arccosine of the elements of the eigenvector.

The graphs showing the item vectors used the exploratory model with two dimensions. The development of these measures is conducted in a polar coordinate system so that direction can be specified as an angle from a particular axis. Using the MIRT item discrimination, the directions of maximum discrimination and MIRT item difficulty can all be depicted in the same graph. The origin of the item vectors is the MIRT difficulty. Item vectors that point in the same essential direction measure essentially the same dimension. Note that by definition, graphs of simple structure are not useful since all items are assigned to a defined axis corresponding to a factor. The reference composite vector composed of all items is also shown as a large red arrow.

The exploratory model is presented for diagnostic purposes to lend further insight into item functioning across dimensions. The resulting item vector plots are presented using the two-dimensional exploratory model. Plots are presented for ELA/literacy and mathematics within grade, across two adjacent grades, and for the subset of common, vertical linking items. The graphs of directional measures are presented in Figures 3 to 11 for ELA/literacy. Figures 12 to 19 show item vectors for ELA/literacy across adjacent grades while Figures 20 to 27 show them for the subset of vertical linking items. The graphs of directional measures are presented in Figures 28 to 36 for

mathematics. Figures 37 to 44 show item vectors for mathematics across adjacent grades and Figures 45 to 52 display ones for the subset of vertical linking items. The plots using the two-dimensional exploratory model suggest that most items are primarily influenced by a composite of both factors. The item vector plot for mathematics for the vertical linking items for grades 8 and 9 shows the composite vector more closely associated with the first factor (θ_1). This closer association may indicate the transition to high school course-specific content. In addition, for the vertical linking set for ELA/literacy grades 9 and 10, some highly discriminating items are associated with the first factor. To understand these item vectors further, clustering was performed on the angle measures. Items were clustered based on having item angles with either 20 or 30 degrees. If an item measured only the primary trait, the angle measure α would be 0° ; whereas if the item measured the primary factor and secondary factor equally, the angle would be 45° . The angle of 20 would correspond to item clusters being more closely associated with a given factor than an angle of 30. Barplots of these are shown in Figures 53 to 70. Like the vector plots, the barplots are given within a grade, across adjacent grades, and for the subset of vertical linking items for both ELA/literacy and mathematics. The number of items associated with a cluster is plotted in the barplots. Using Grade 3 ELA shown in Figure 53 as an example, the item clusters for angle 20 shows two distinct groups with slightly over a hundred items in each one. Items with highly similar loading are demonstrated by the height of the barplot. When the clustering uses an angle measure of 30° then a single cluster is clearly distinct that includes a preponderance of the items.

Discussion and Conclusion

The evidence based on these analyses suggests that no consistent and pervasive multidimensionality was demonstrated. However, no model fit the data particularly well. The outcome based on the global fit measures suggested that the unidimensional model was consistently the preferred model. This was generally indicated by lower fit values for the unidimensional model relative to other ones. The difference chi-square need not indicate significant improvement over the unidimensional case; that would have been indicated by a ratio of the chi-square to the degrees of freedom in the 3 to 5 range. Using the two dimensional exploratory model, item vector plots evaluated how the items were associated based on the respective traits. The vector plots indicated most items were a composite of the two factors falling along the 45° diagonal as indicated by the composite item vector. No clear pattern in the item vectors was exhibited that might have permitted factor rotation that would have further facilitated interpretation. In the final step, the exploratory model based on clustering of the item angle measures into groups with similar factor loadings, shown in the bar plots, was examined. The clusters were investigated within grade, across adjacent grades, and for vertical linking items. Clusters of 20 degrees usually showed two distinct clusters being formed with one of them often being more prominent. This pattern is generally consistent with the definition of essential unidimensionality where there is a primary dimension and some minor ones. When the clustering criterion was 30, a single distinct measure was usually present in which the vast majority of items were grouped.

Although a unidimensional model was preferred, differences in dimensionality were most evident in mathematics in the transition from grade 8 to grade 9. This difference is expected since this delimits the transition into the course-specific content characterized by high school.

Based on results of the dimensionality analyses, no changes are warranted to the scaling design, and all items for a grade and content area were calibrated together simultaneously using unidimensional techniques. The approach adopted here was to use the best available information from the Pilot Test to inform decision making regarding future development phases. Mathematics performance-task items were not available for inclusion in the analysis. At a minimum, the test dimensionality study based on the Pilot Test can only be viewed as preliminary and will need to be readdressed in the future. This is partly reflected in the changes that occurred in the item types,

content configurations, and test design used in the Pilot Test compared with those employed for the Field Test. An overall concern is the degree of implementation of the Common Core State Standards across the Consortium at the time of this study. This may affect the results of this dimensionality study in ways that cannot currently be anticipated. The Field Test and future operational administration will better reflect student performance while schools are more evenly implementing the Common Core State Standards.

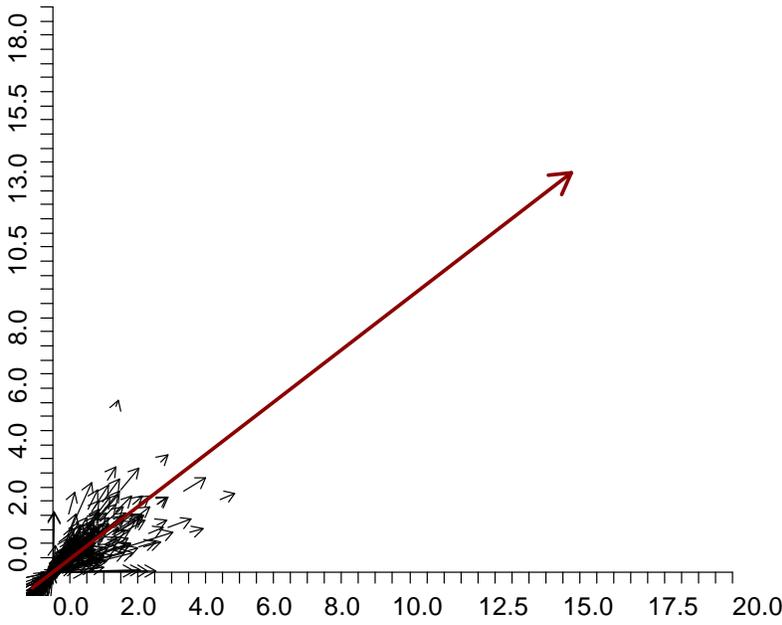


Figure 3. Item Vector Plot for ELA/literacy Grade 3 (Within Grade)

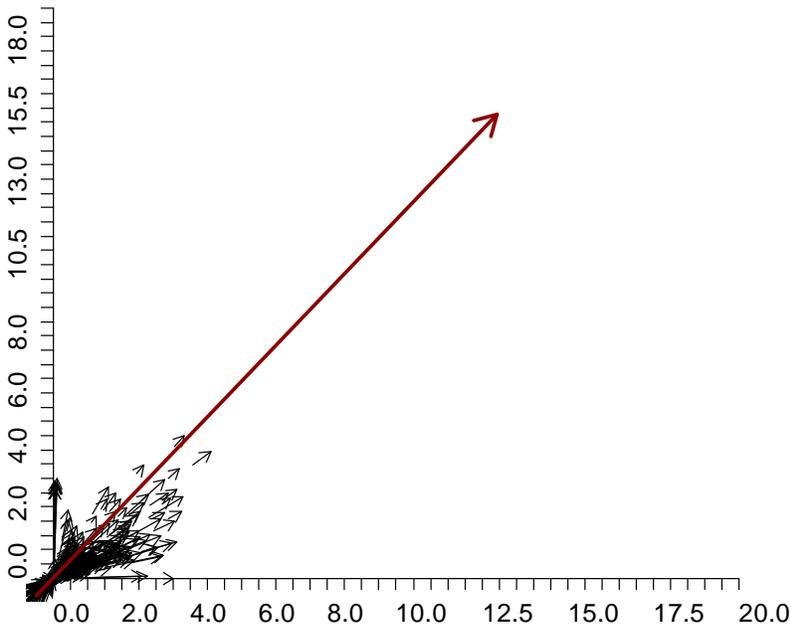


Figure 4. Item Vector Plot for ELA/literacy Grade 4 (Within Grade)

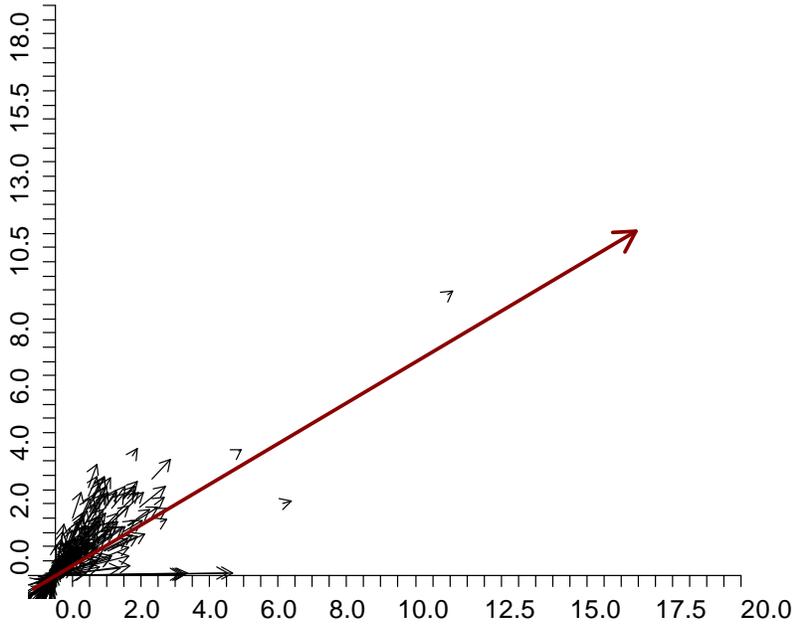


Figure 5. Item Vector Plot for ELA/literacy Grade 5 (Within Grade)

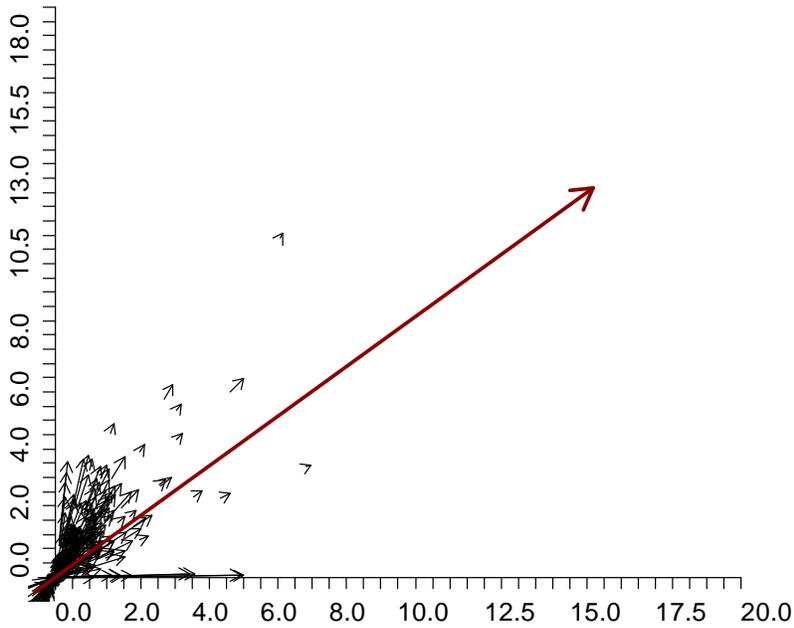


Figure 6. Item Vector Plot for ELA/literacy Grade 6 (Within Grade)

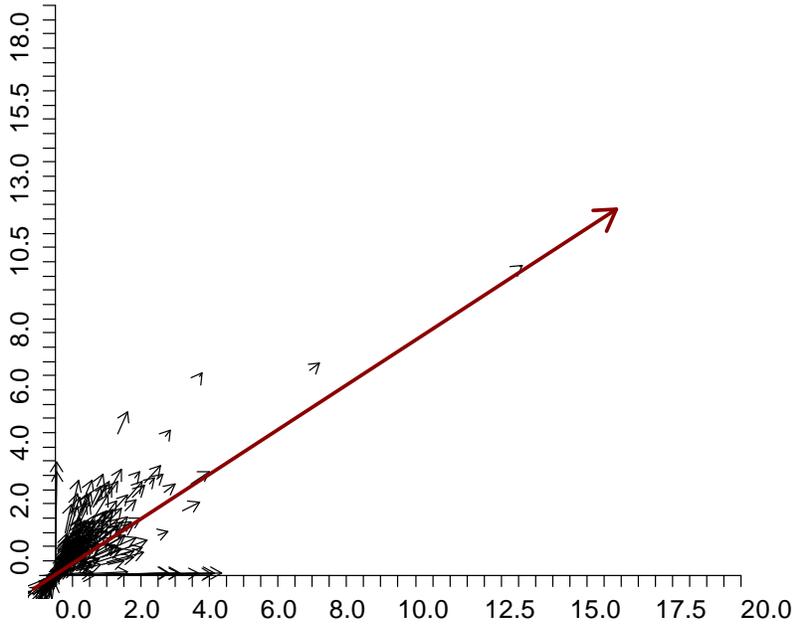


Figure 7. Item Vector Plot for ELA/literacy Grade 7 (Within Grade)

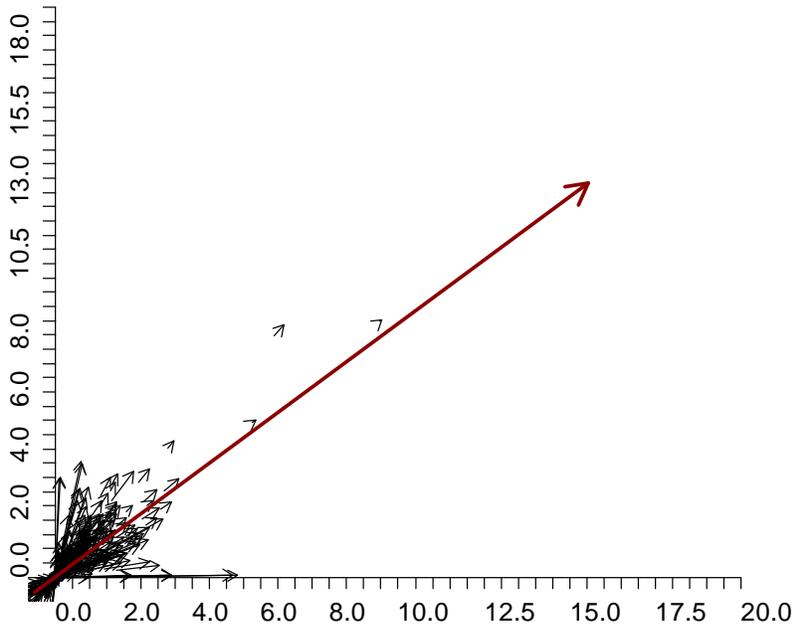


Figure 8. Item Vector Plot for ELA/literacy Grade 8 (Within Grade)

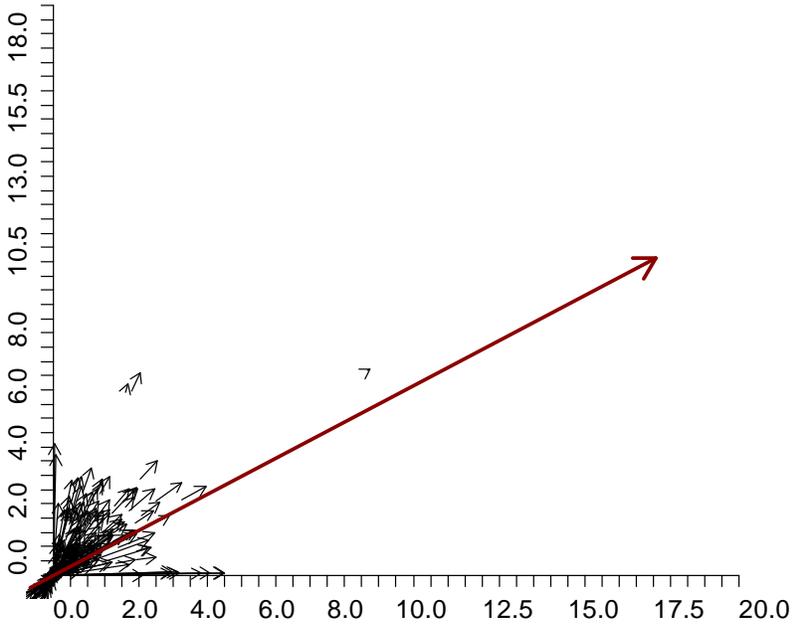


Figure 9. Item Vector Plot for ELA/literacy Grade 9 (Within Grade)

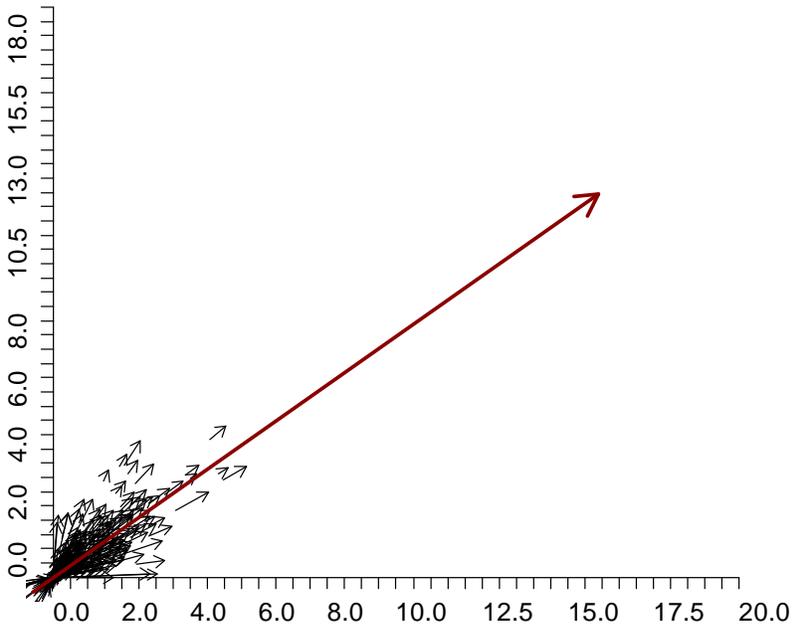


Figure 10. Item Vector Plot for ELA/literacy Grade 10 (Within Grade)

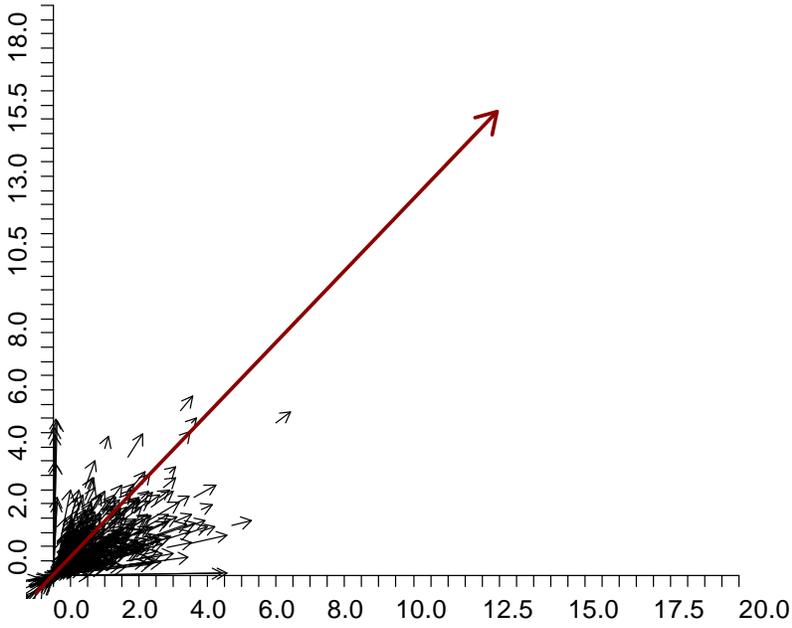


Figure 11. Item Vector Plot for ELA/literacy Grade 11 (Within Grade)

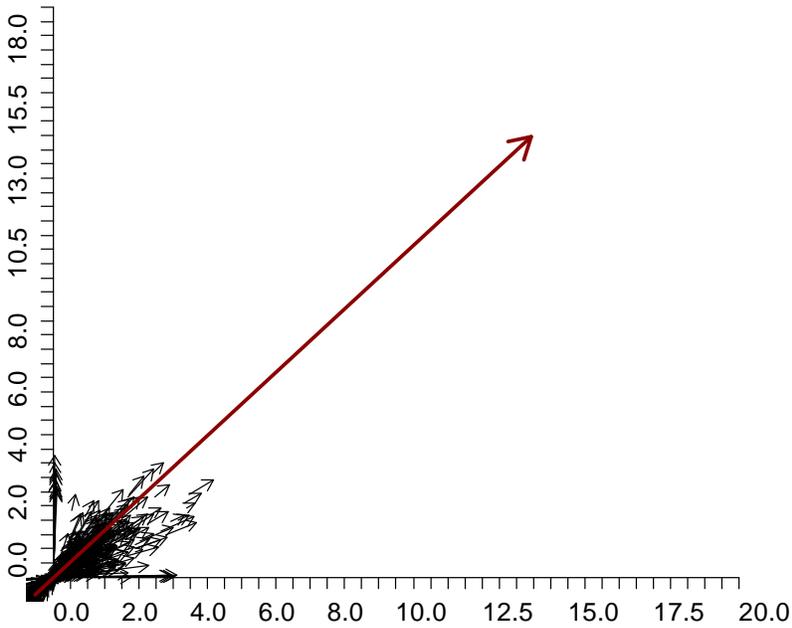


Figure 12. Item Vector Plot for ELA/literacy Grades 3 and 4 (Across Grades)

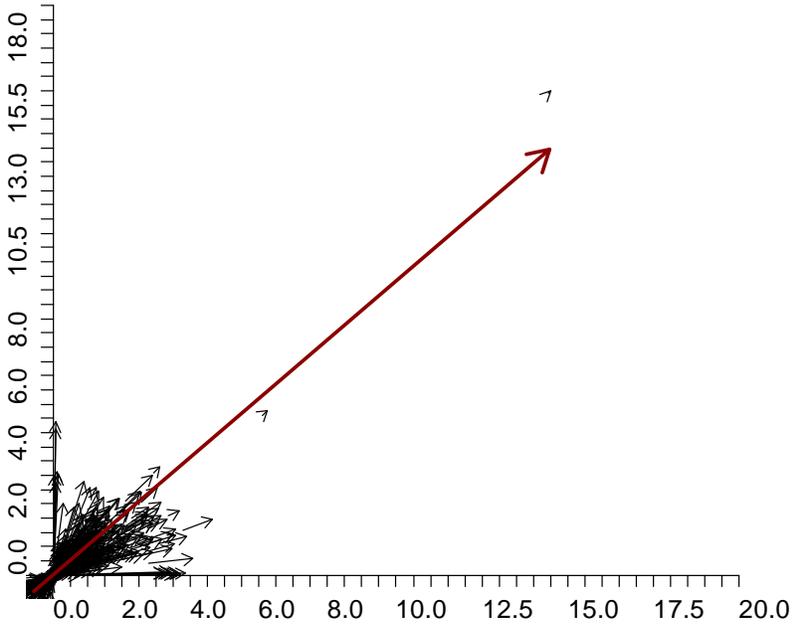


Figure 13. Item Vector Plot for ELA/literacy Grades 4 and 5 (Across Grades)

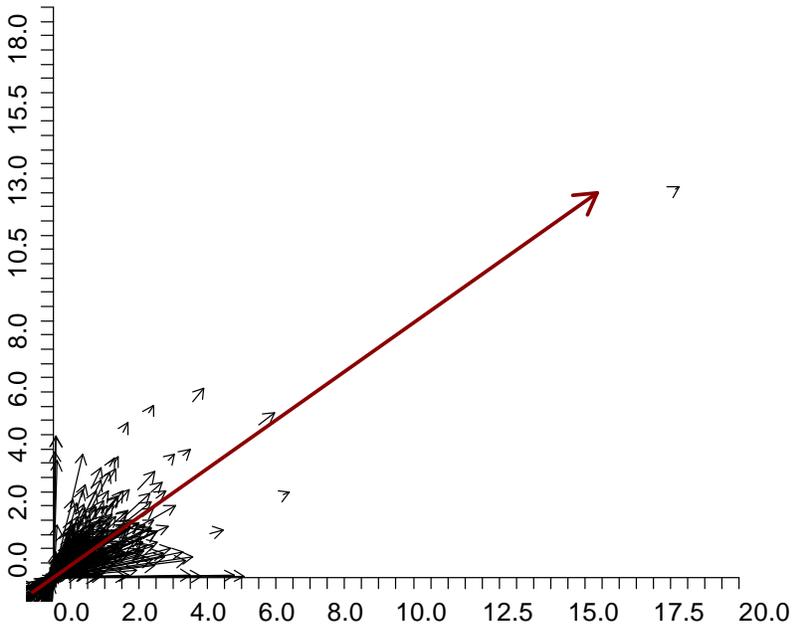


Figure 14. Item Vector Plot for ELA/literacy Grades 5 and 6 (Across Grades)

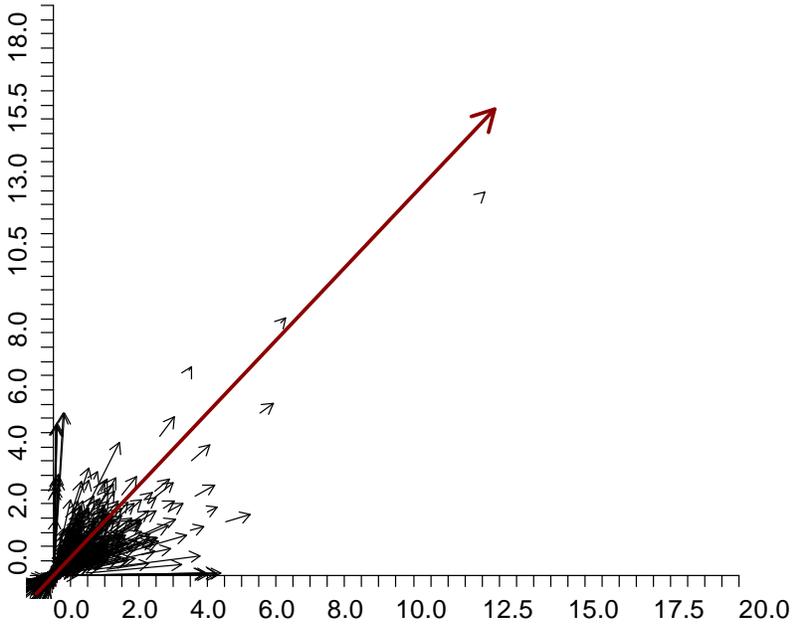


Figure 15. Item Vector Plot for ELA/literacy Grades 6 and 7 (Across Grades)

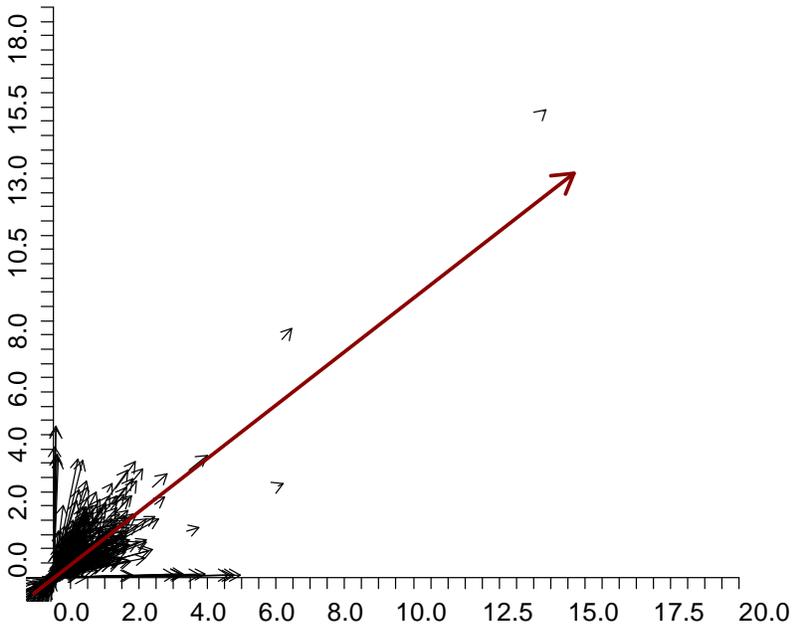


Figure 16. Item Vector Plot for ELA/literacy Grades 7 and 8 (Across Grades)

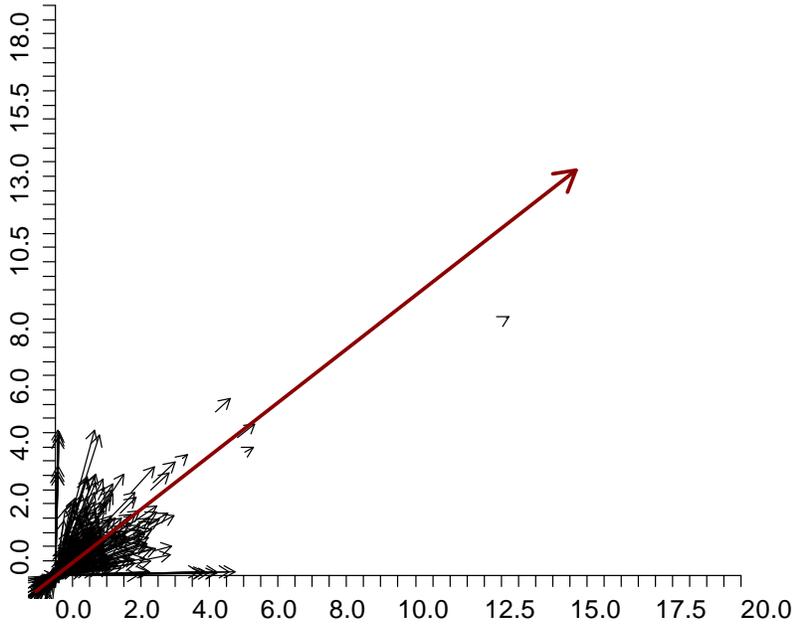


Figure 17. Item Vector Plot for ELA/literacy Grades 8 and 9 (Across Grades)

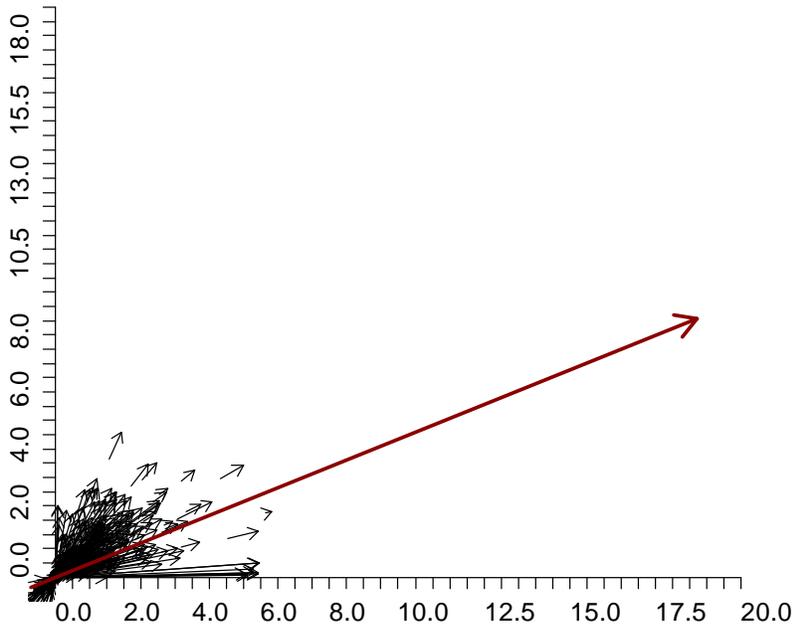


Figure 18. Item Vector Plot for ELA/literacy Grades 9 and 10 (Across Grades)

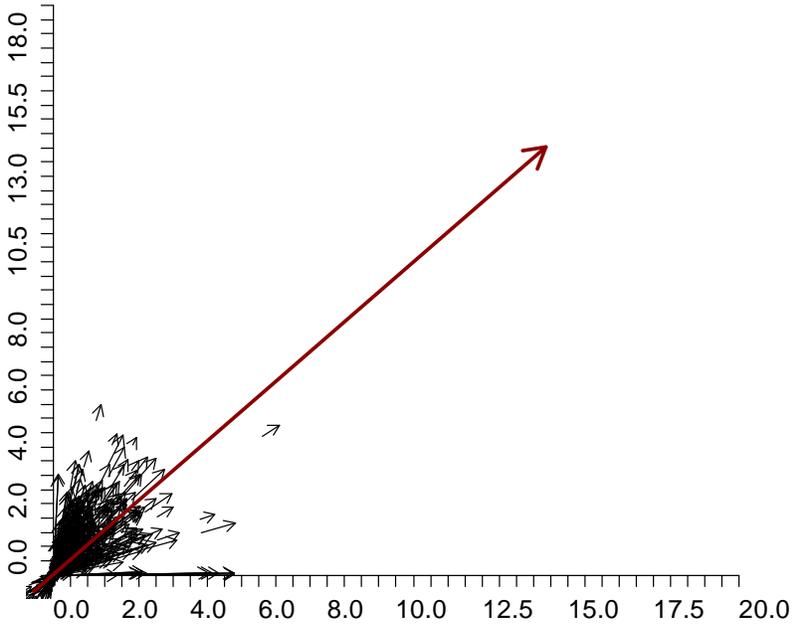


Figure 19. Item Vector Plot for ELA/literacy Grades 10 and 11 (Across Grades)

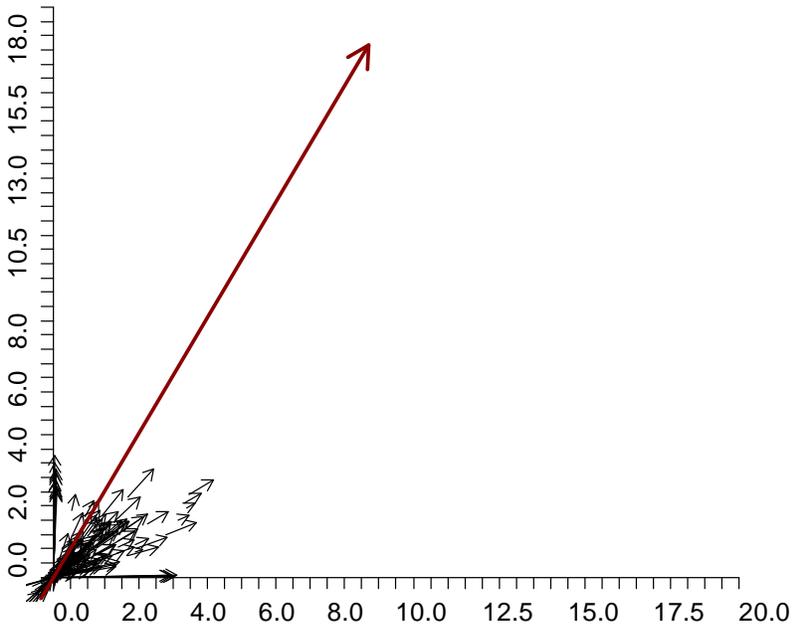


Figure 20. Item Vector Plots for the Subset of ELA/literacy Grades 3 and 4 Vertical Linking Items

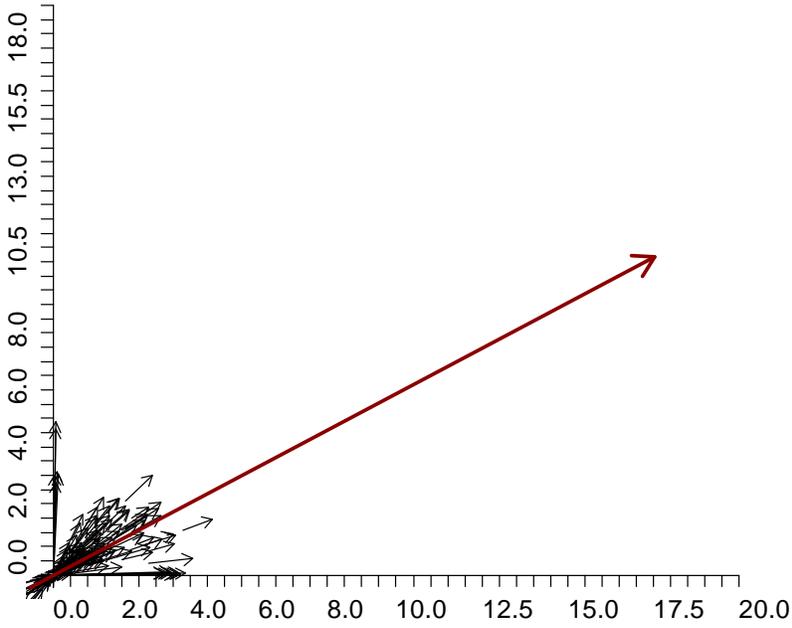


Figure 21. Item Vector Plots for the Subset of ELA/literacy Grades 4 and 5 Vertical Linking Items

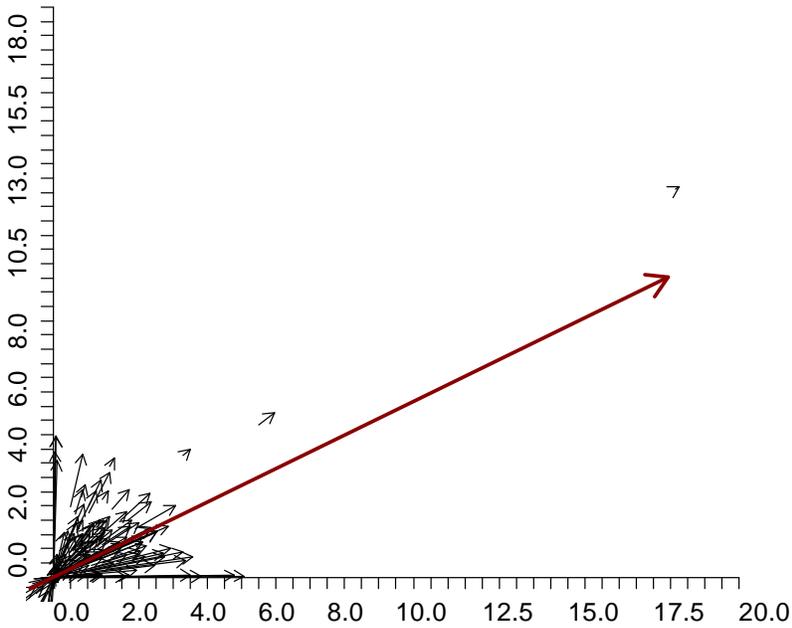


Figure 22. Item Vector Plots for the Subset of ELA/literacy Grades 5 and 6 Vertical Linking Items

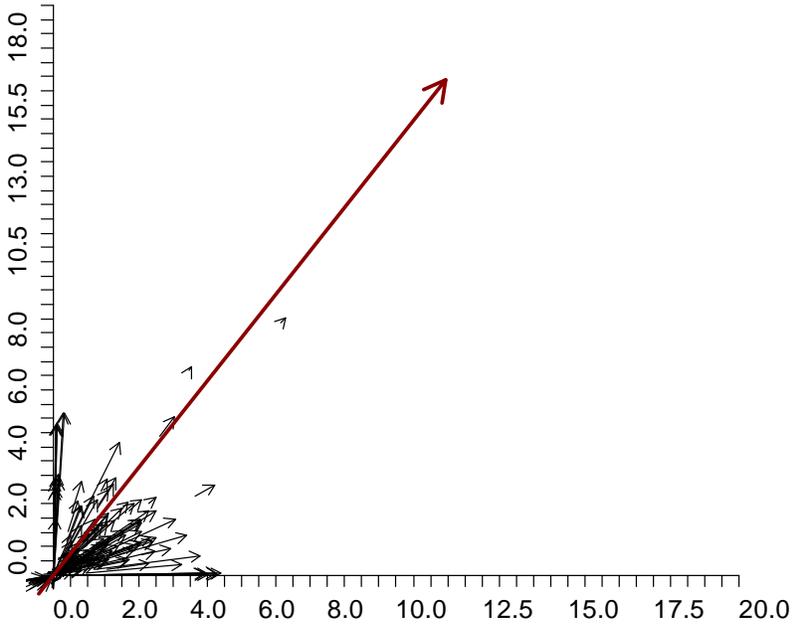


Figure 23. Item Vector Plots for the Subset of ELA/literacy Grades 6 and 7 Vertical Linking Items

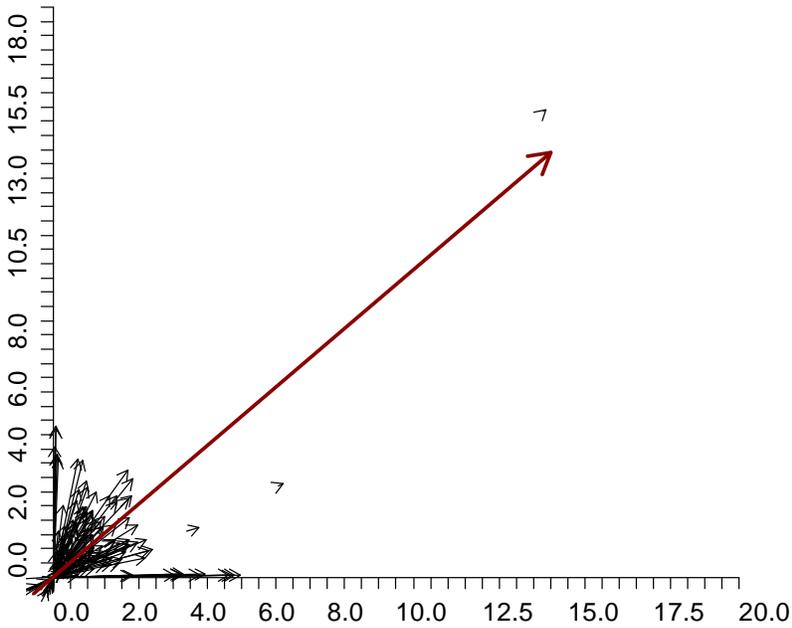


Figure 24. Item Vector Plots for the Subset of ELA/literacy Grades 7 and 8 Vertical Linking Items

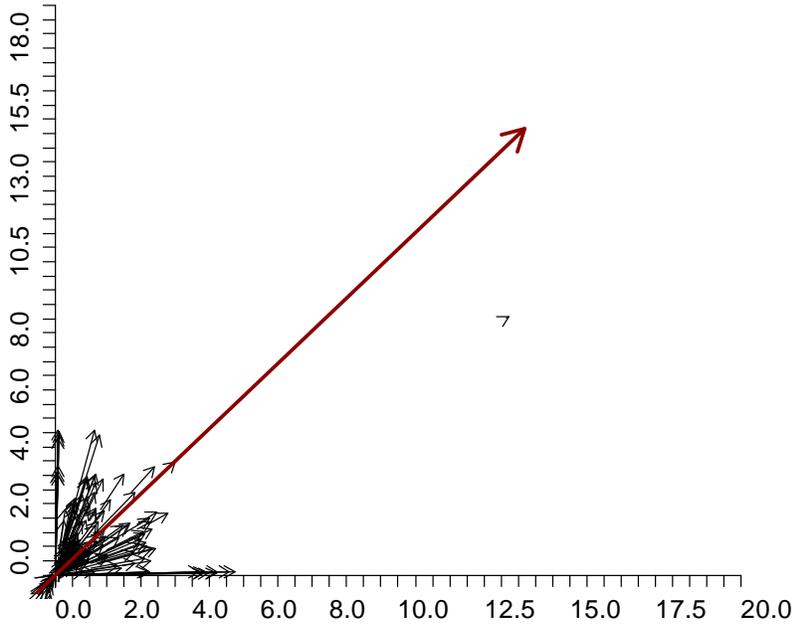


Figure 25. Item Vector Plots for the Subset of ELA/literacy Grades 8 and 9 Vertical Linking Items

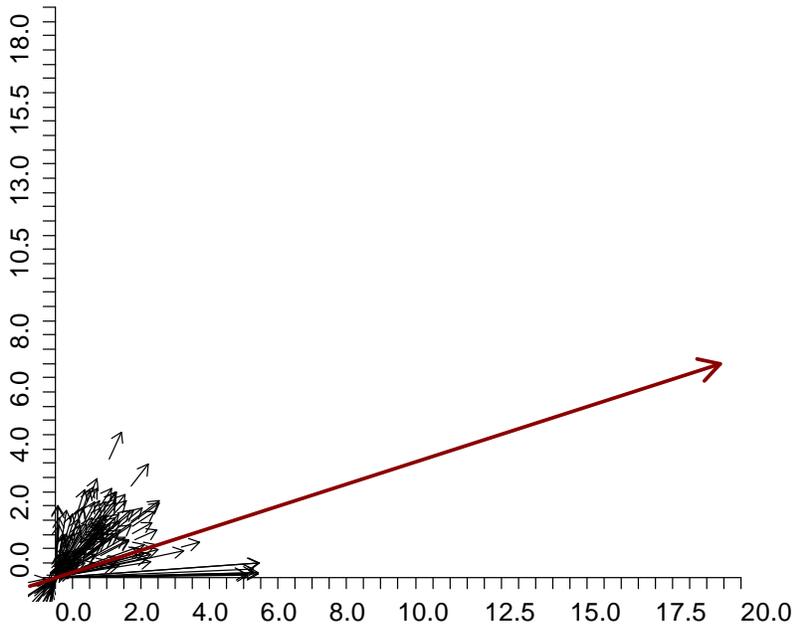


Figure 26. Item Vector Plots for the Subset of ELA/literacy Grades 9 and 10 Vertical Linking Items

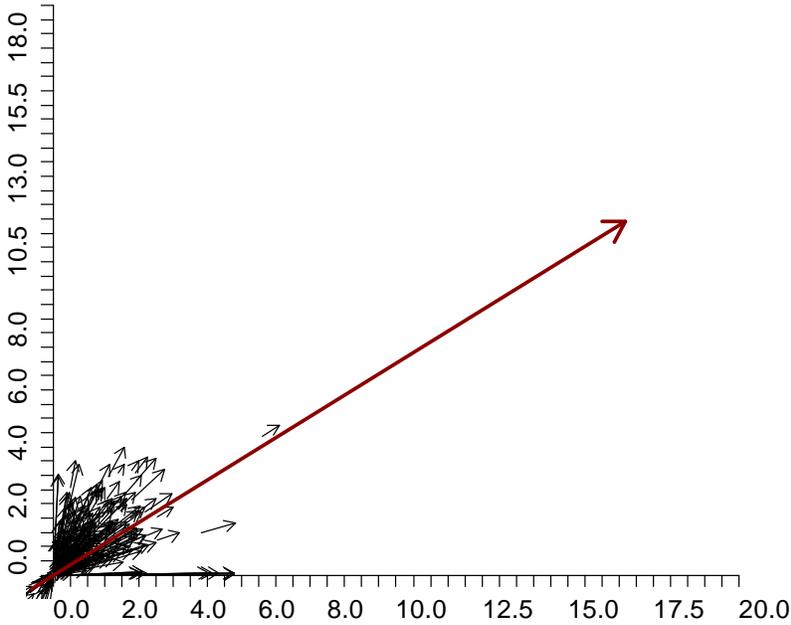


Figure 27. Item Vector Plots for the Subset of ELA/literacy Grades 10 and 11 Vertical Linking Items

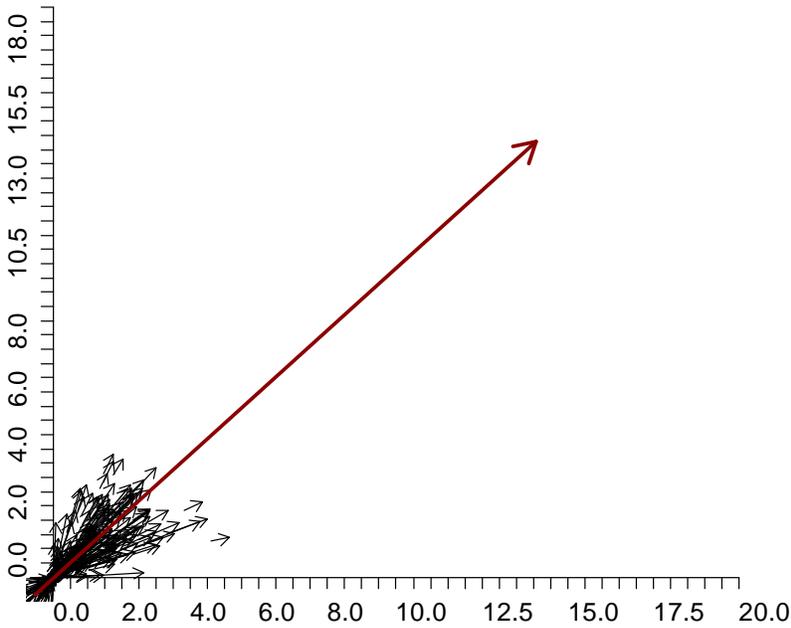


Figure 28. Item Vector Plot for Mathematics Grade 3 (Within Grade)

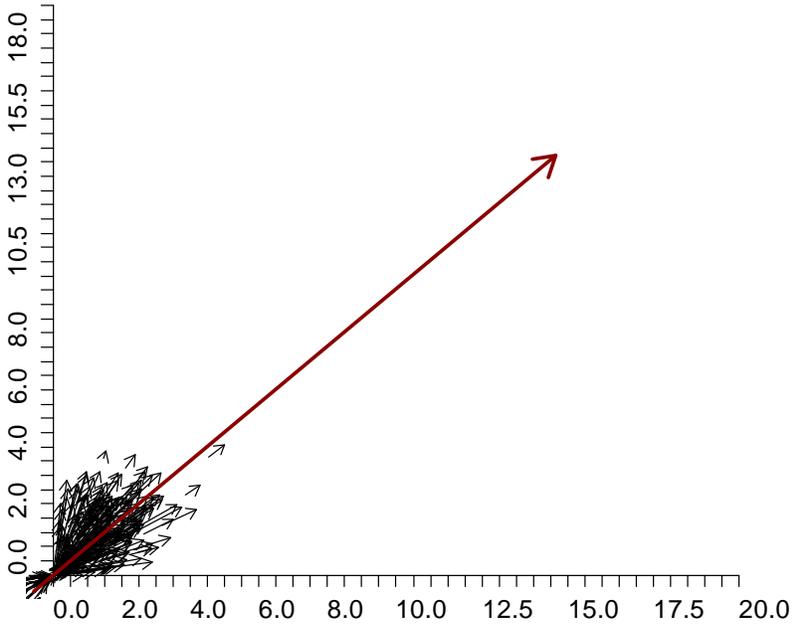


Figure 29. Item Vector Plot for Mathematics Grade 4 (Within Grade)

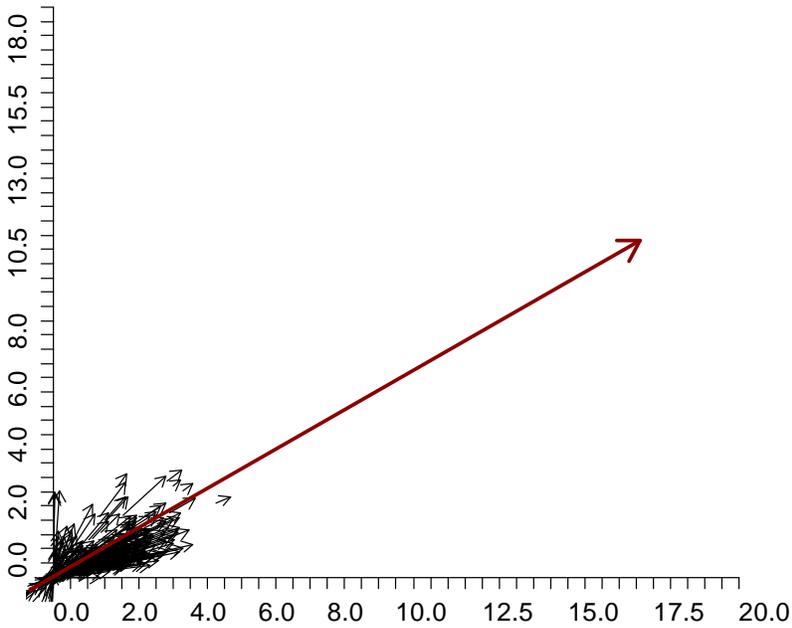


Figure 30. Item Vector Plot for Mathematics Grade 5 (Within Grade)

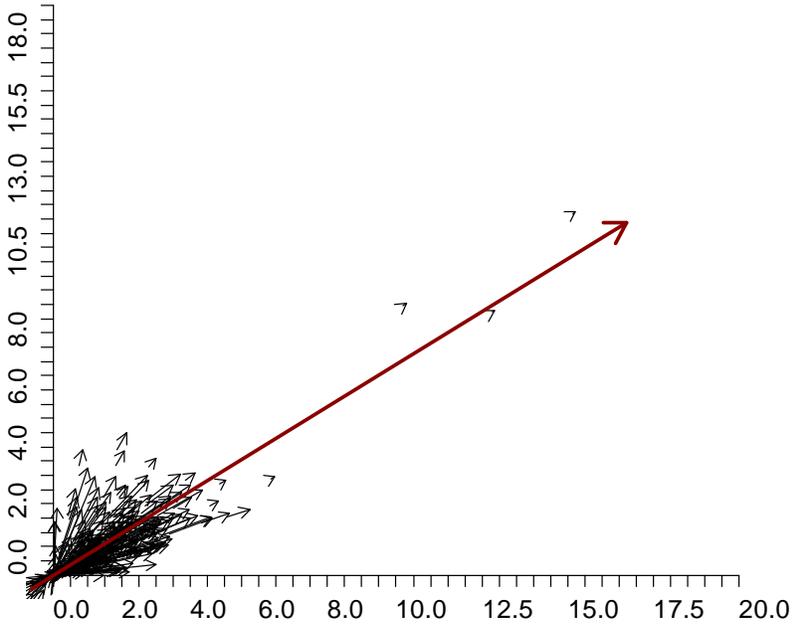


Figure 31. Item Vector Plot for Mathematics Grade 6 (Within Grade)

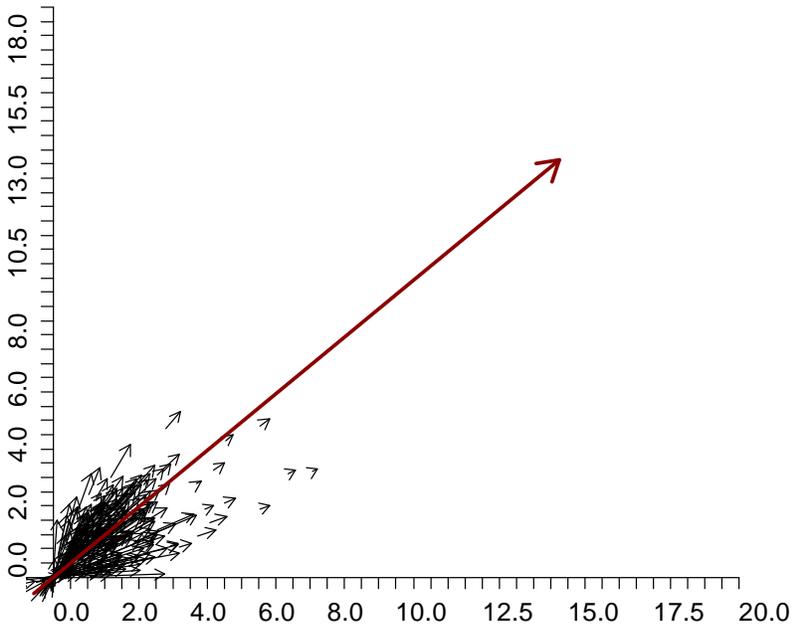


Figure 32. Item Vector Plot for Mathematics Grade 7 (Within Grade)

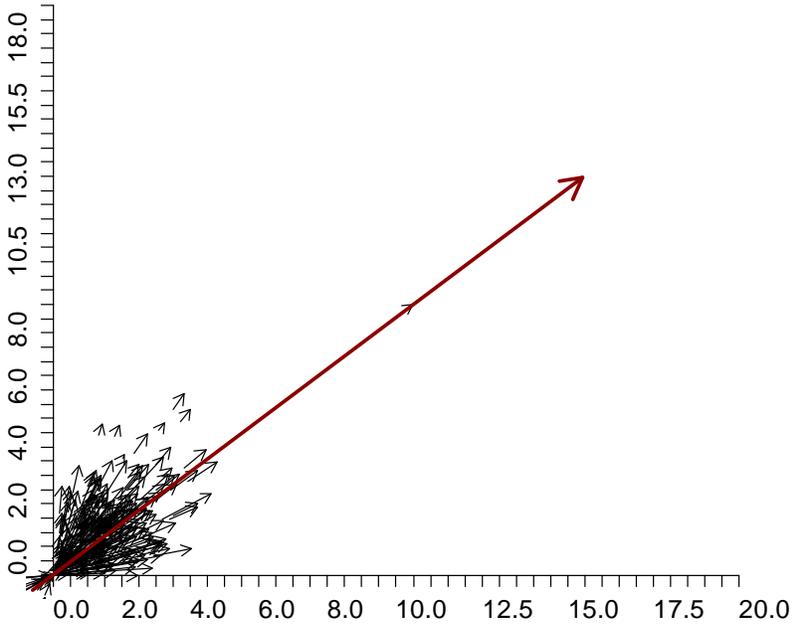


Figure 33. Item Vector Plot for Mathematics Grade 8 (Within Grade)

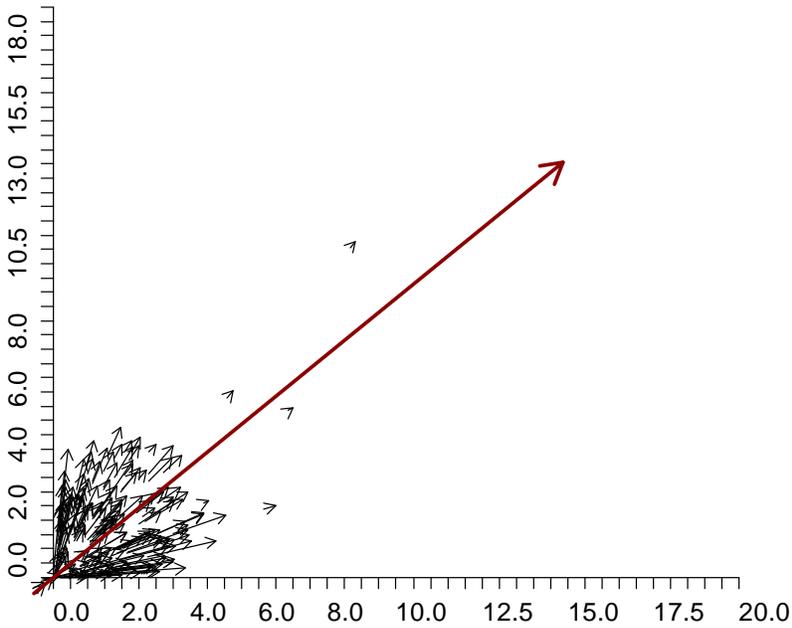


Figure 34. Item Vector Plot for Mathematics Grade 9 (Within Grade)

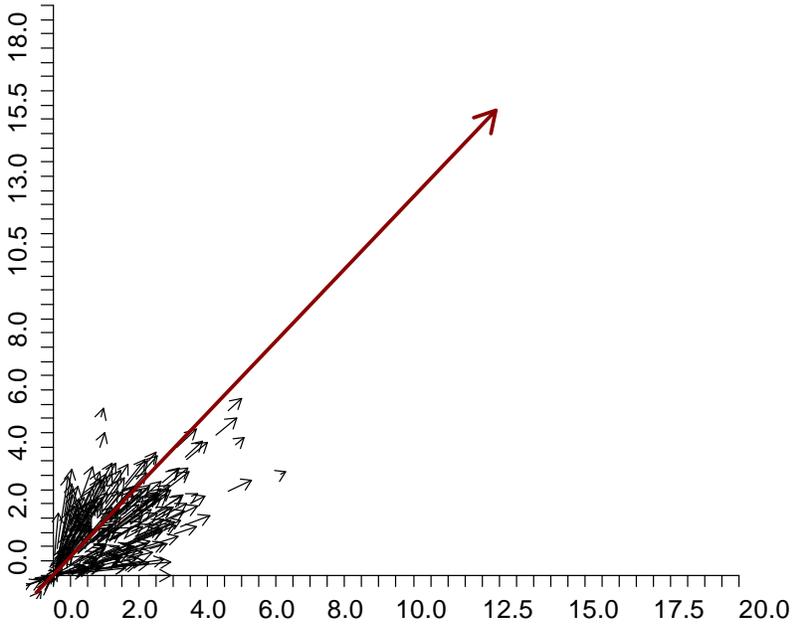


Figure 35. Item Vector Plot for Mathematics Grade 10 (Within Grade)

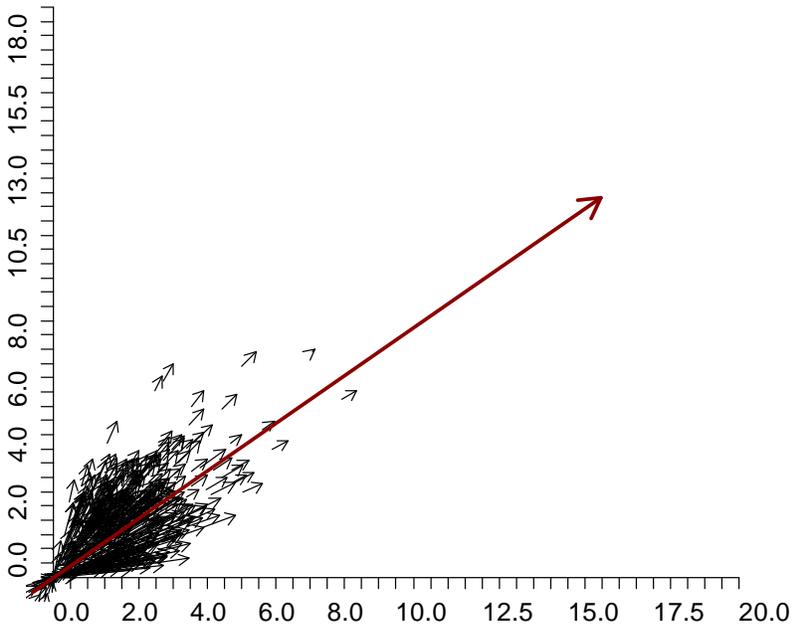


Figure 36. Item Vector Plot for Mathematics Grade 11 (Within Grade)

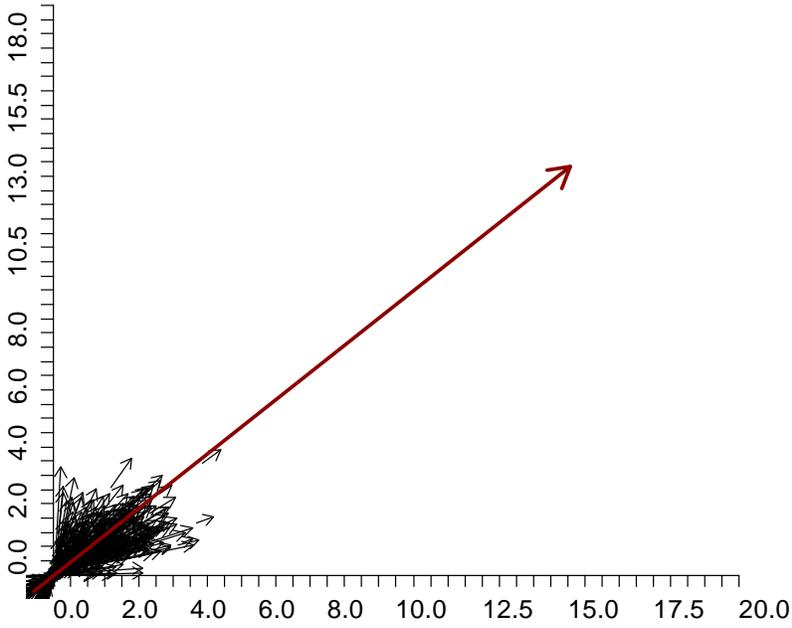


Figure 37. Item Vector Plot for Mathematics Grades 3 and 4 (Across Grades)

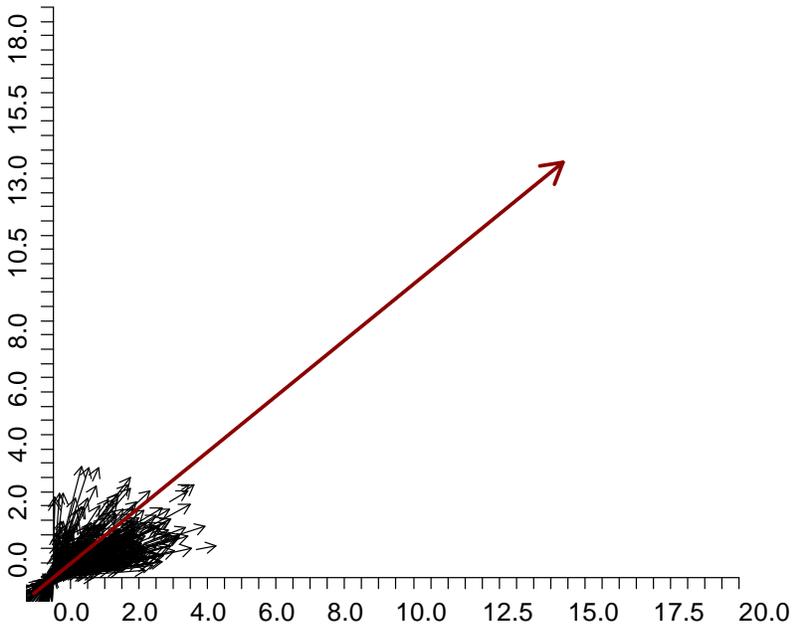


Figure 38. Item Vector Plot for Mathematics Grades 4 and 5 (Across Grades)

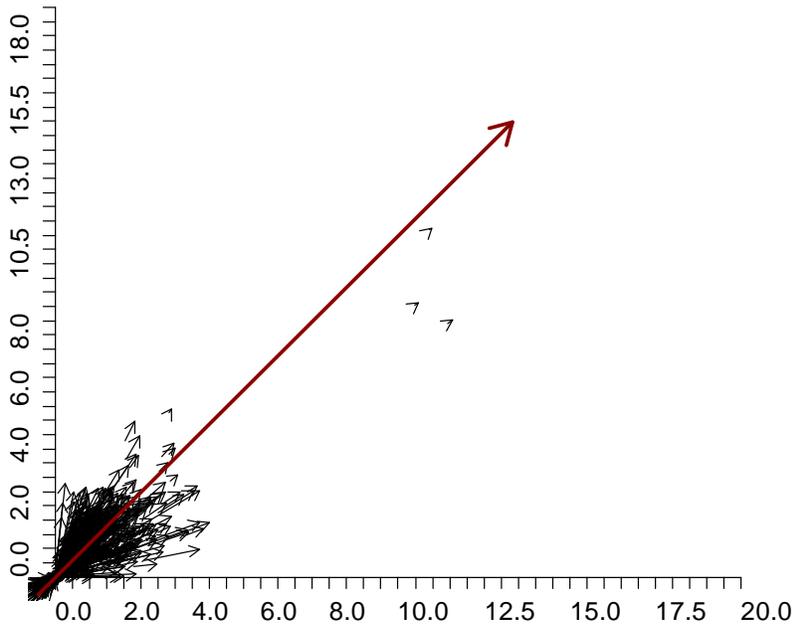


Figure 39. Item Vector Plot for Mathematics Grades 5 and 6 (Across Grades)

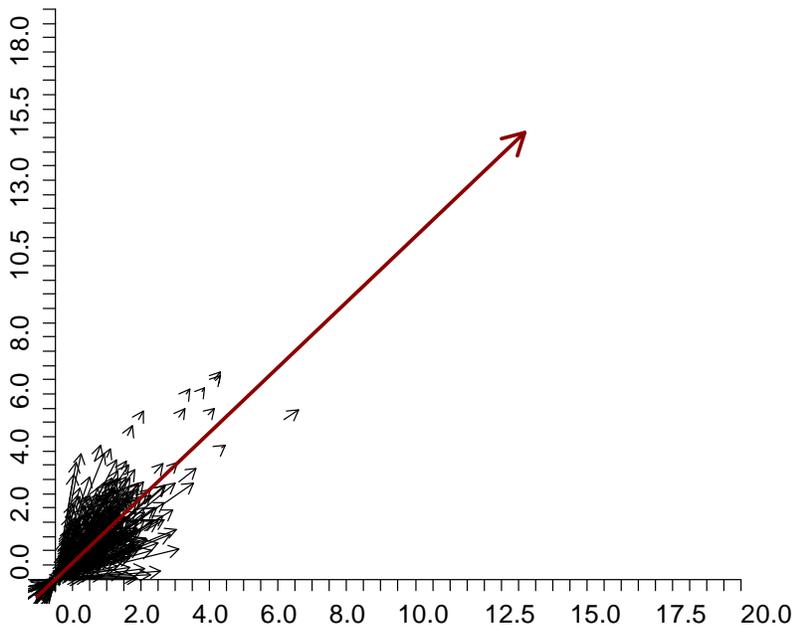


Figure 40. Item Vector Plot for Mathematics Grades 6 and 7 (Across Grades)

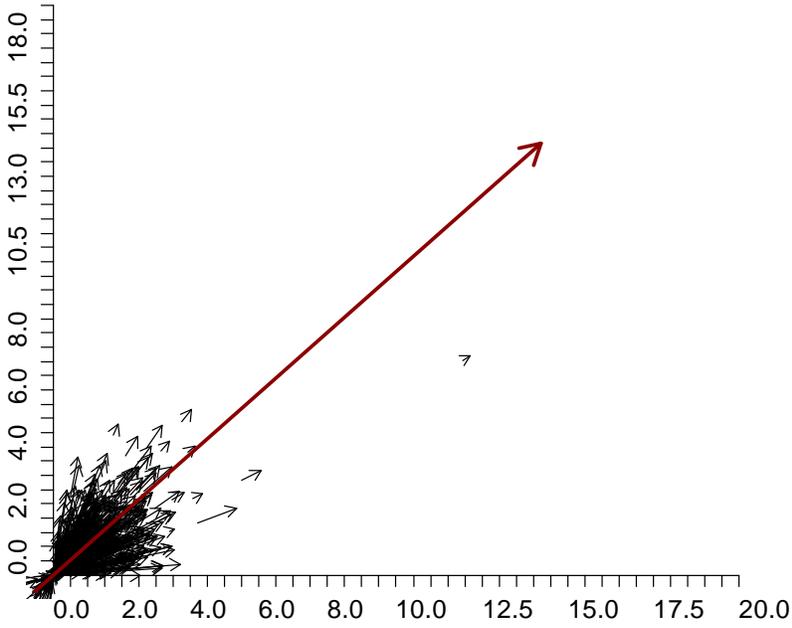


Figure 41. Item Vector Plot for Mathematics Grades 7 and 8 (Across Grades)

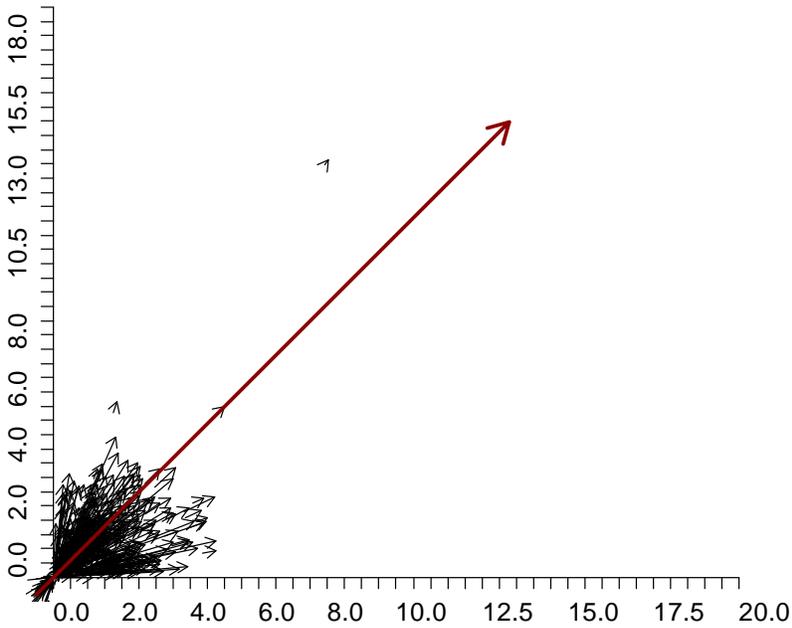


Figure 42. Item Vector Plot for Mathematics Grades 8 and 9 (Across Grades)

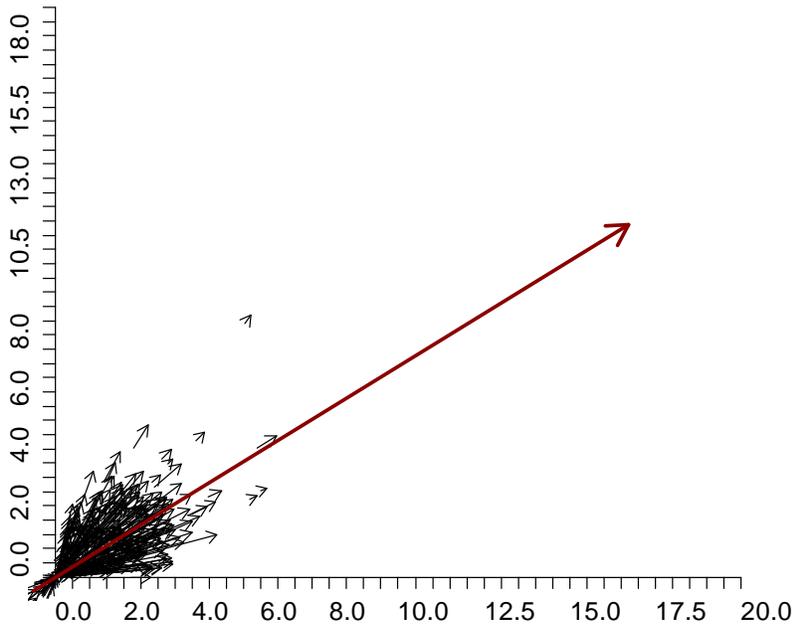


Figure 43. Item Vector Plot for Mathematics Grades 9 and 10 (Across Grades)

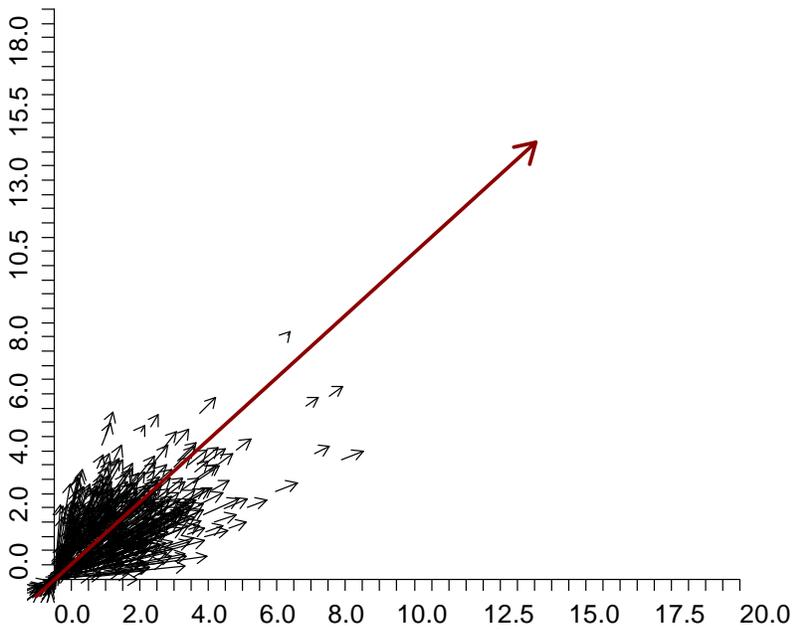


Figure 44. Item Vector Plot for Mathematics Grades 10 and 11 (Across Grades)

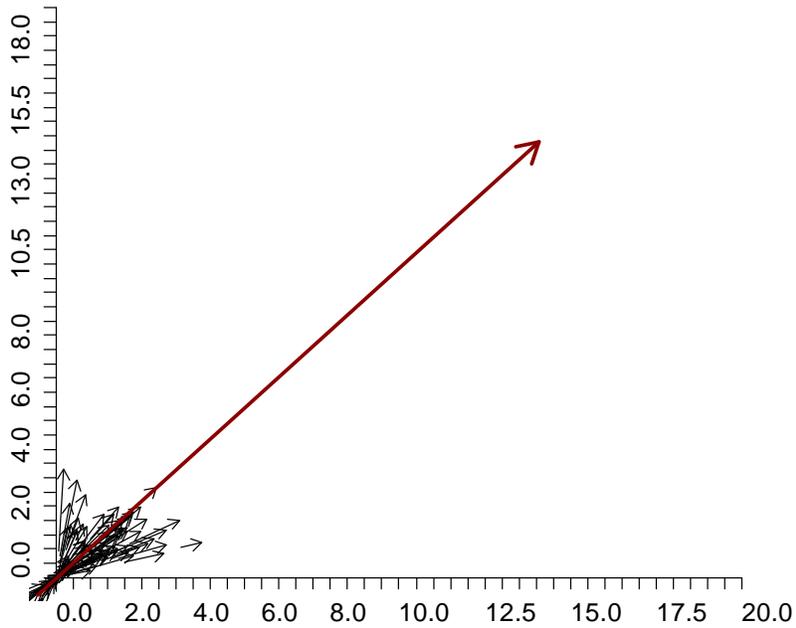


Figure 45. Item Vector Plot for the Subset of Mathematics Grades 3 and 4 (Vertical Linking Items)

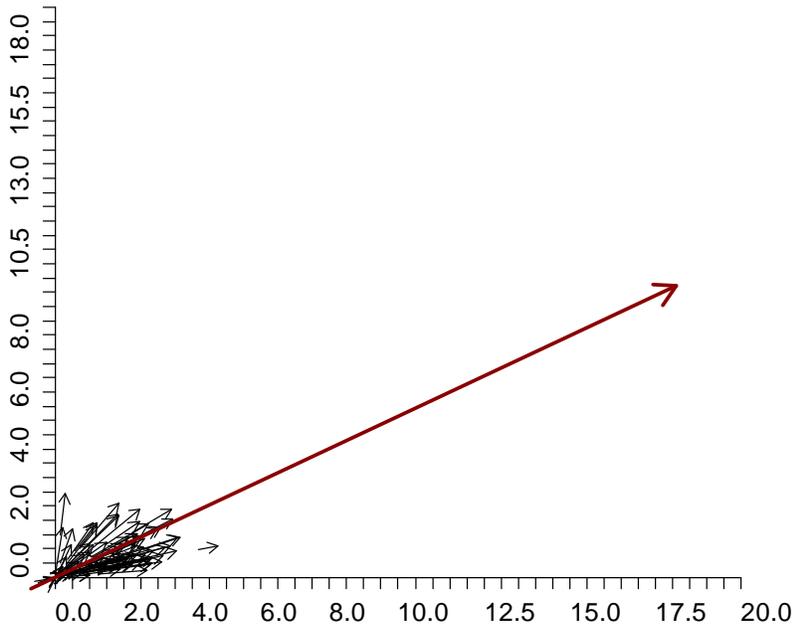


Figure 46. Item Vector Plot for the Subset of Mathematics Grades 4 and 5 (Vertical Linking Items)

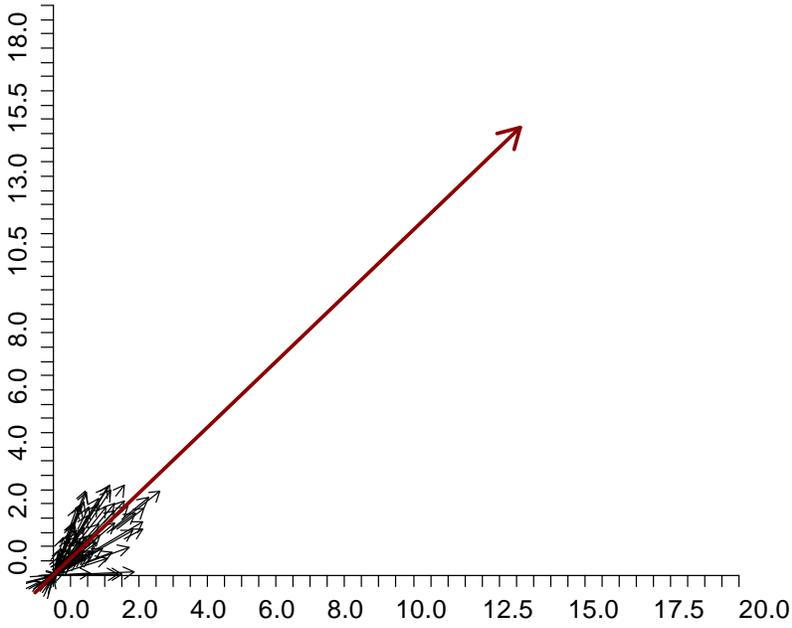


Figure 47. Item Vector Plot for the Subset of Mathematics Grades 5 and 6 (Vertical Linking Items)

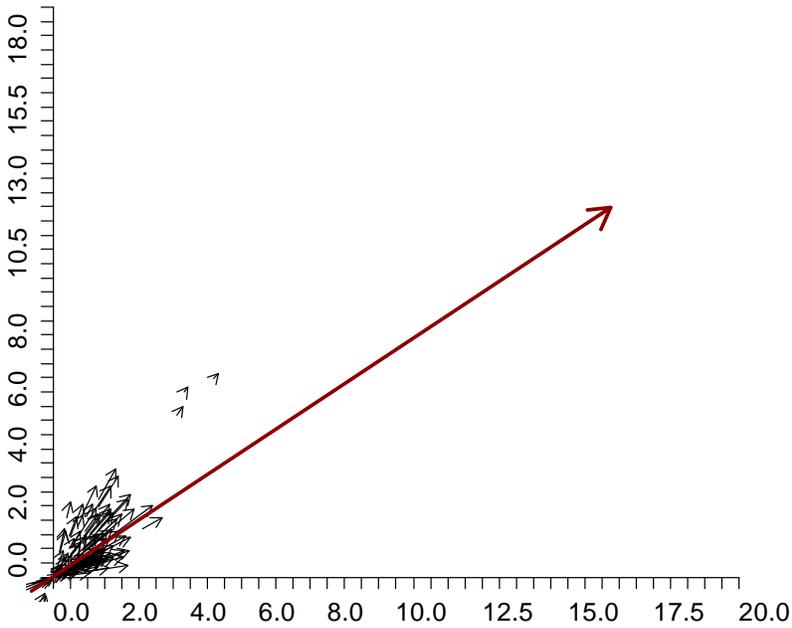


Figure 48. Item Vector Plot for the Subset of Mathematics Grades 6 and 7 (Vertical Linking Items)

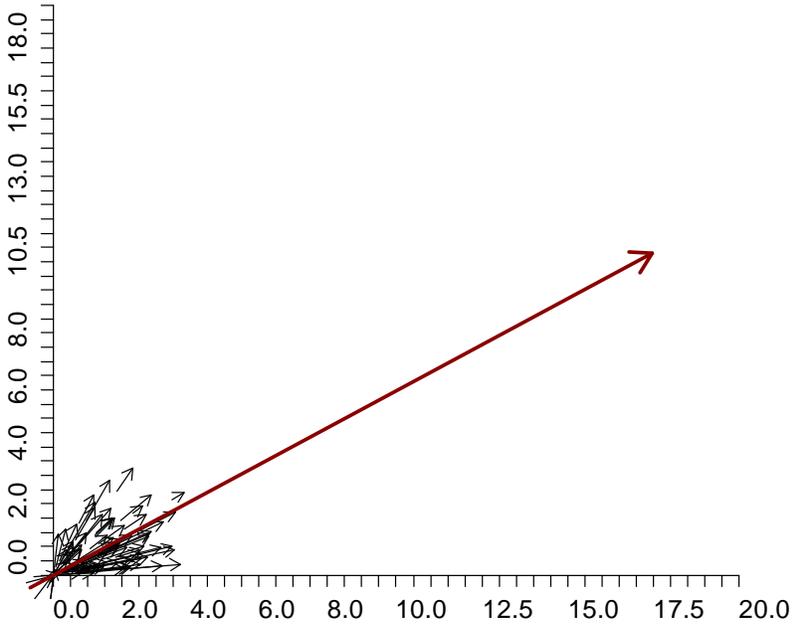


Figure 49. Item Vector Plot for the Subset of Mathematics Grades 7 and 8 (Vertical Linking Items)

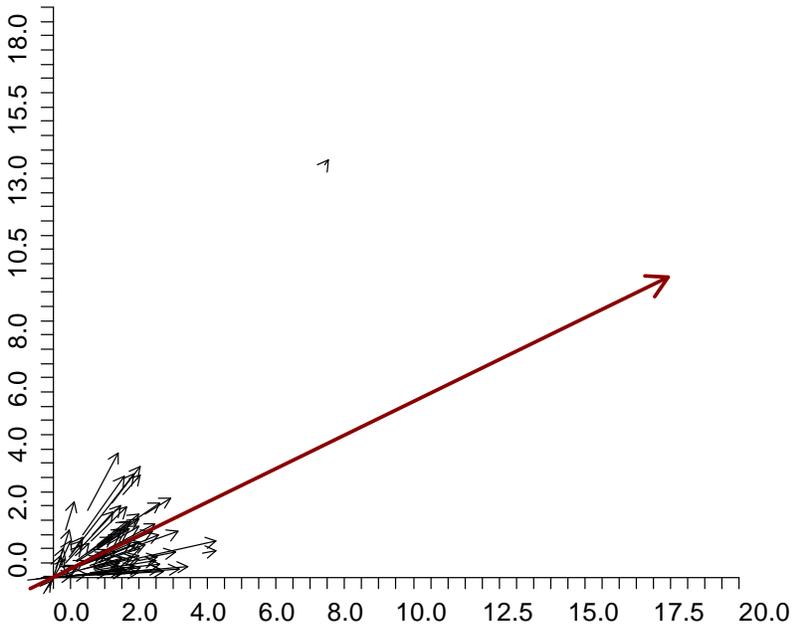


Figure 50. Item Vector Plot for the Subset of Mathematics Grades 8 and 9 (Vertical Linking Items)

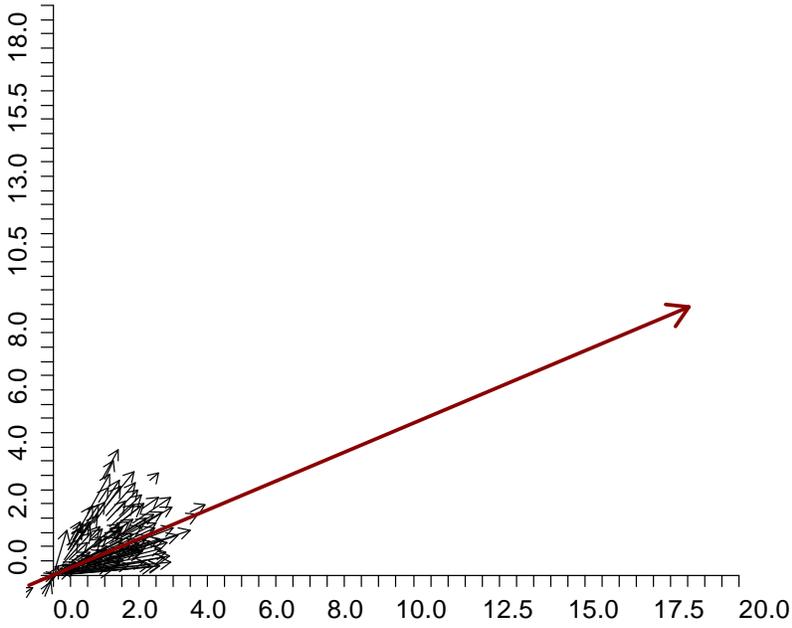


Figure 51. Item Vector Plot for the Subset of Mathematics Grades 9 and 10 (Vertical Linking Items)

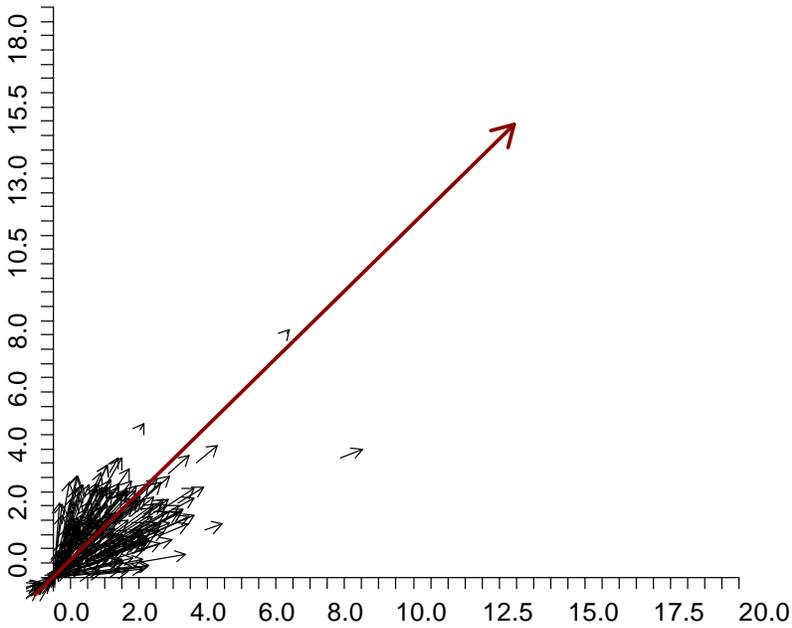


Figure 52. Item Vector Plot for the Subset of Mathematics Grades 10 and 11 (Vertical Linking Items)

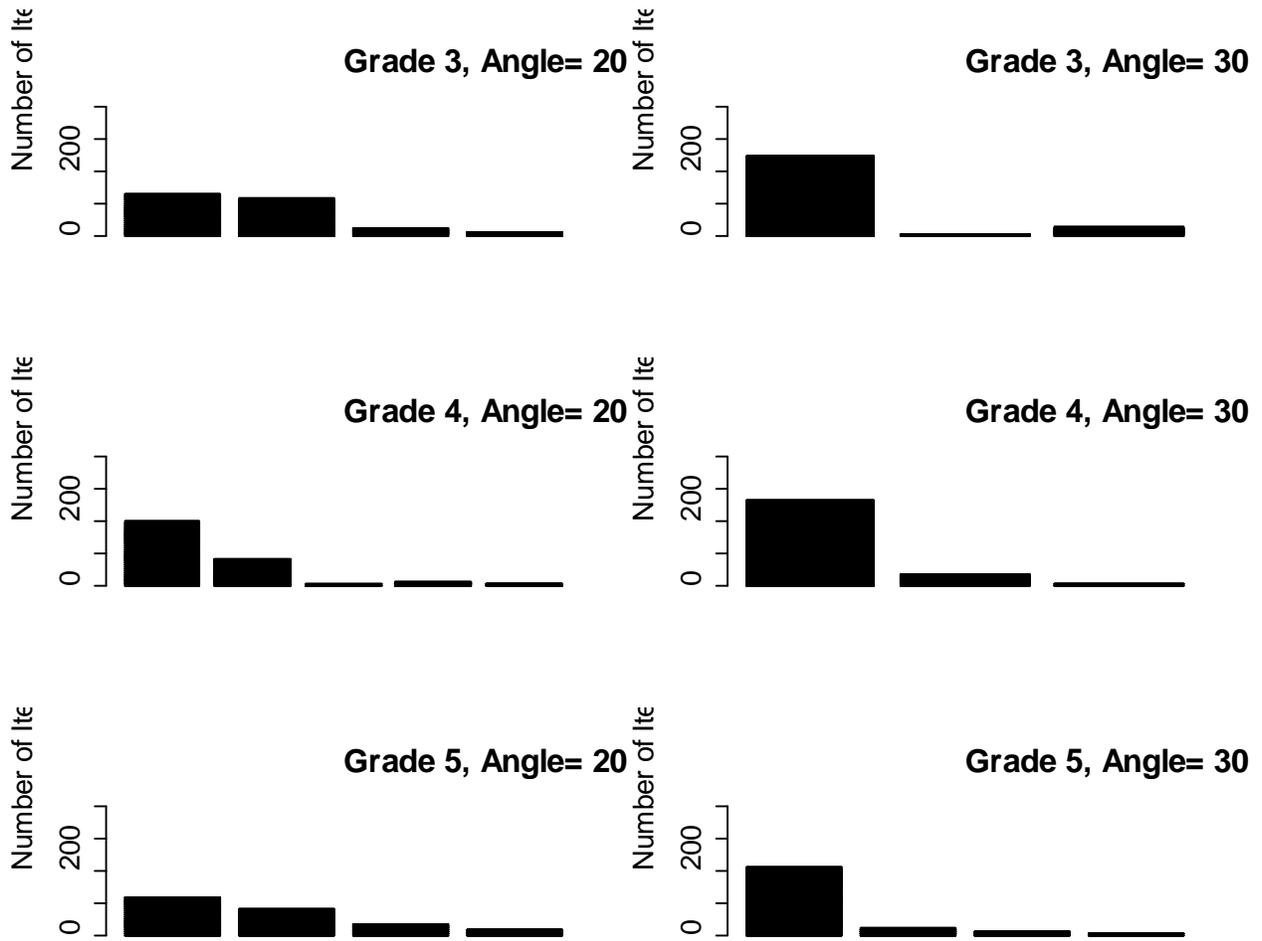


Figure 53. Clustering of Item Angle Measures for Grades 3 to 5, ELA/literacy (within grade)

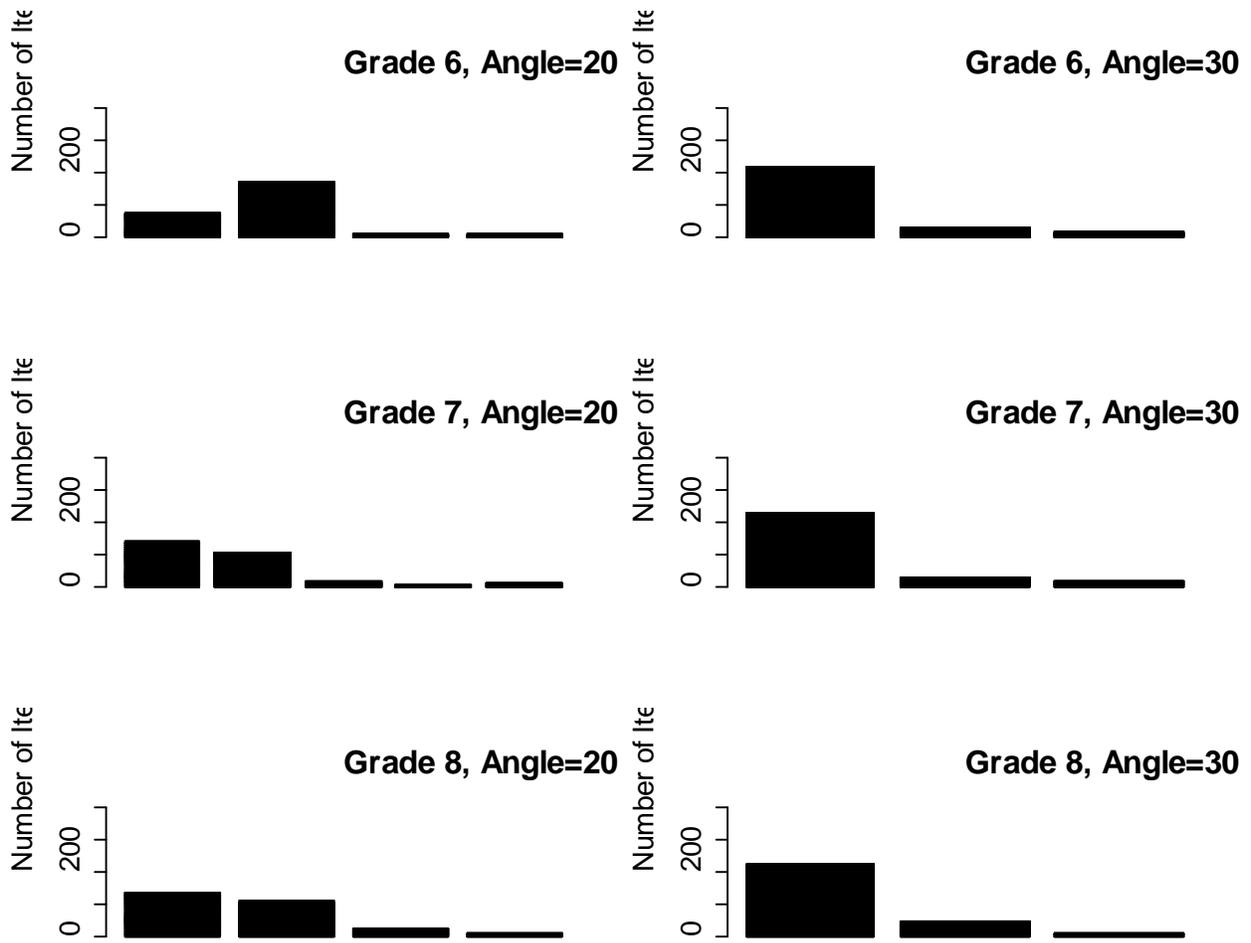


Figure 54. Clustering of Item Angle Measures for Grades 6 to 8, ELA/literacy (within grade)

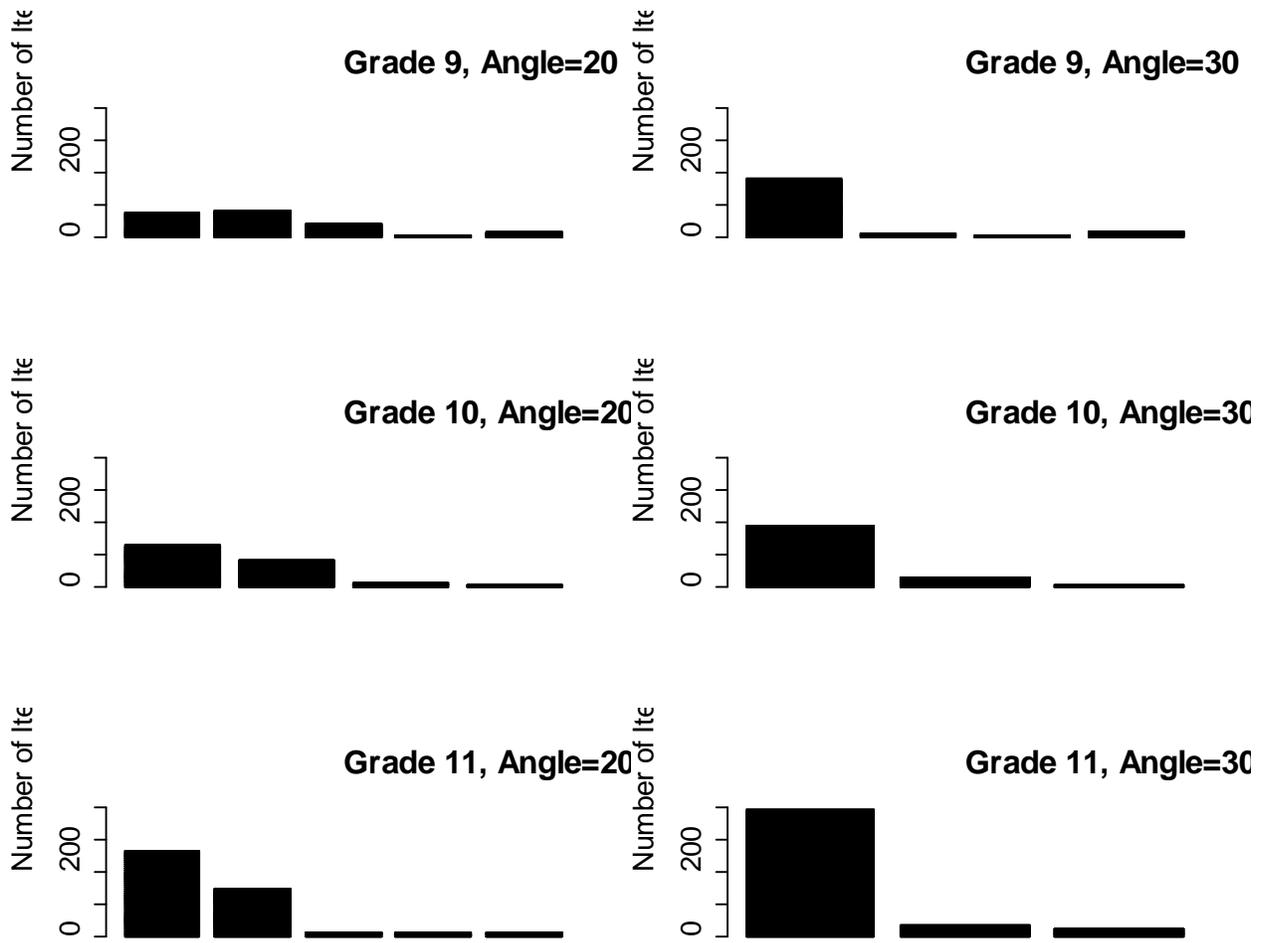


Figure 55. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (within grade)

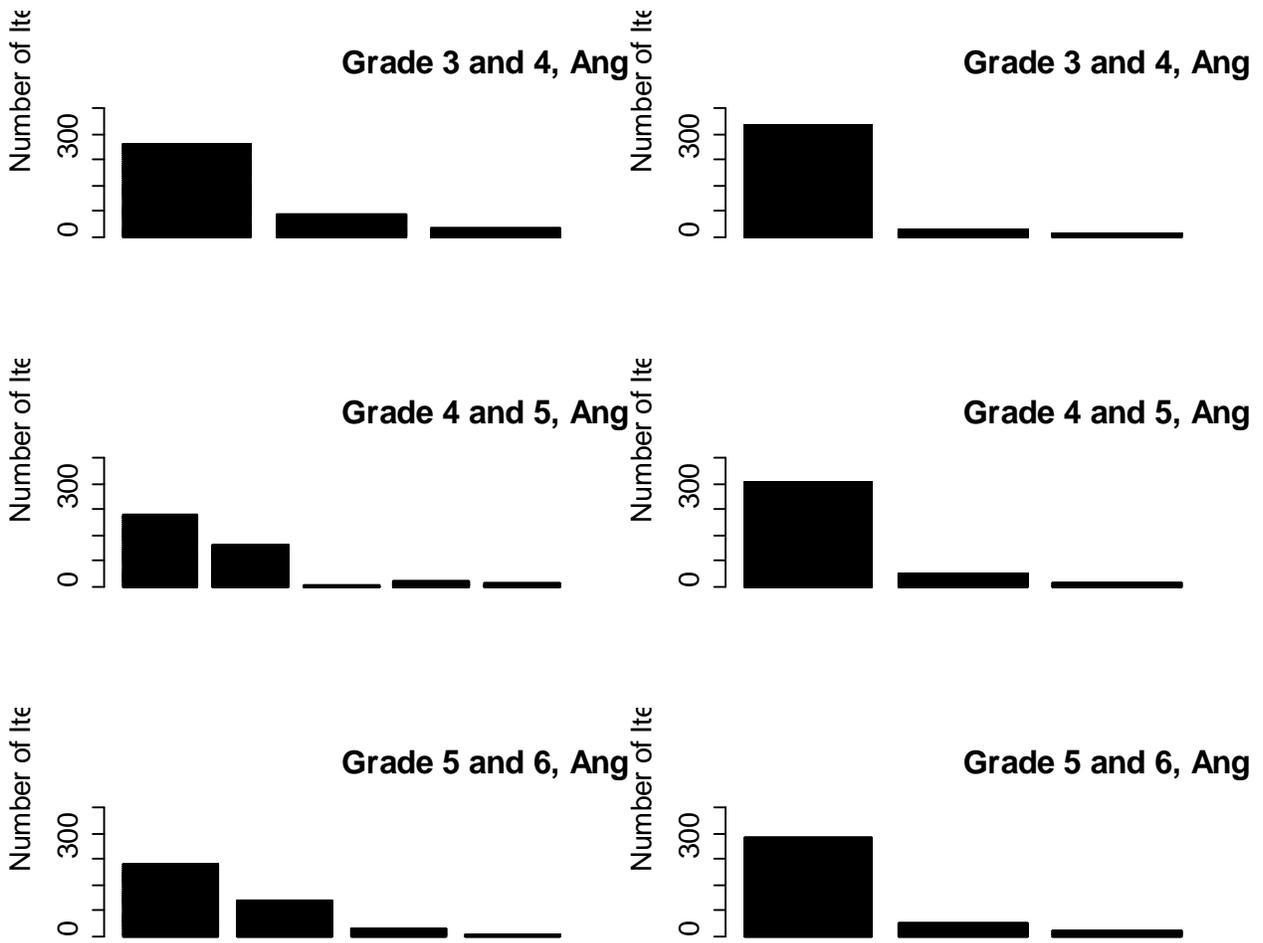


Figure 56. Clustering of Item Angle Measures for Grades 3 to 6, ELA/literacy (across grades)

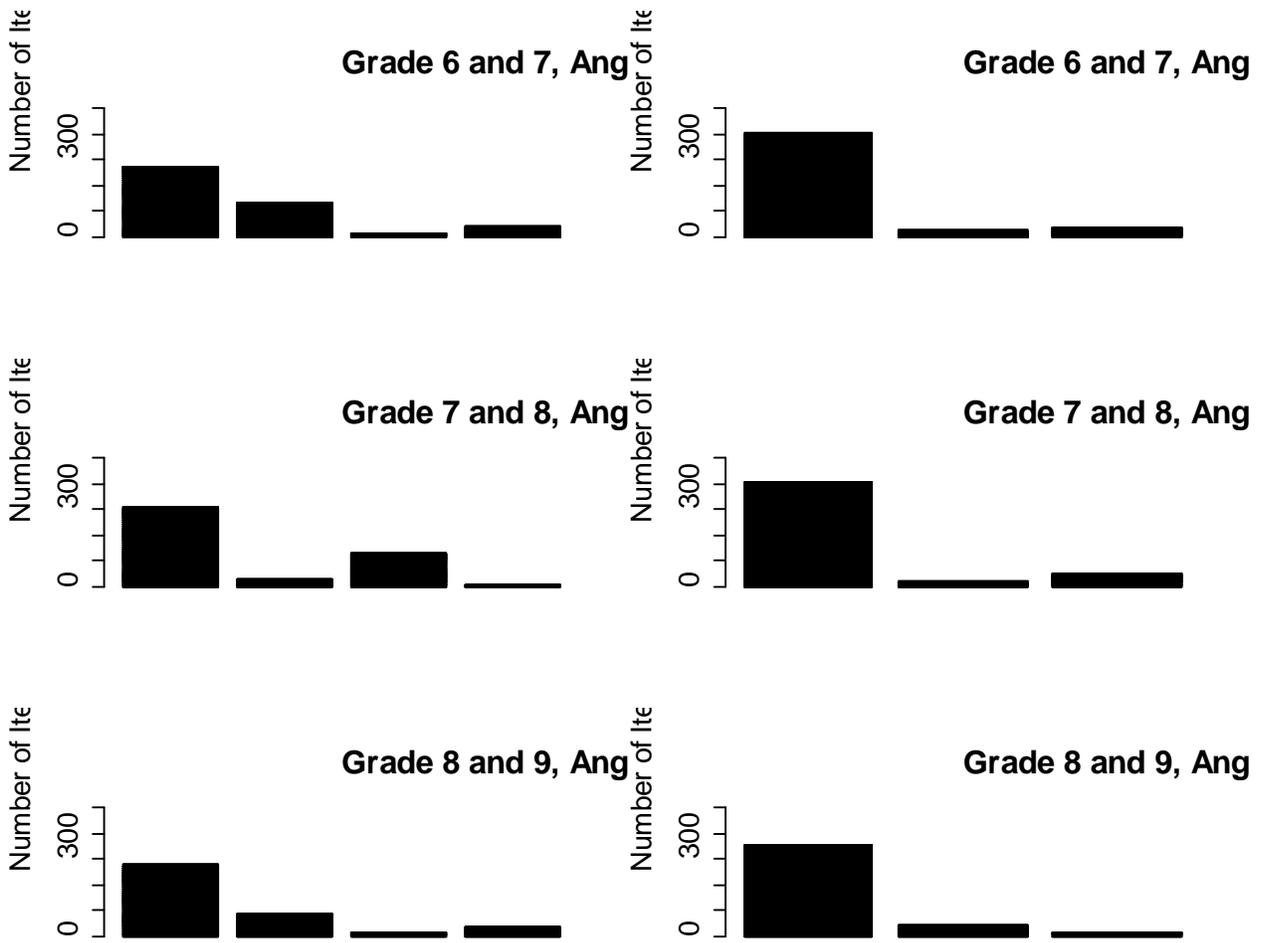


Figure 57. Clustering of Item Angle Measures for Grades 6 to 9, ELA/literacy (across grades)

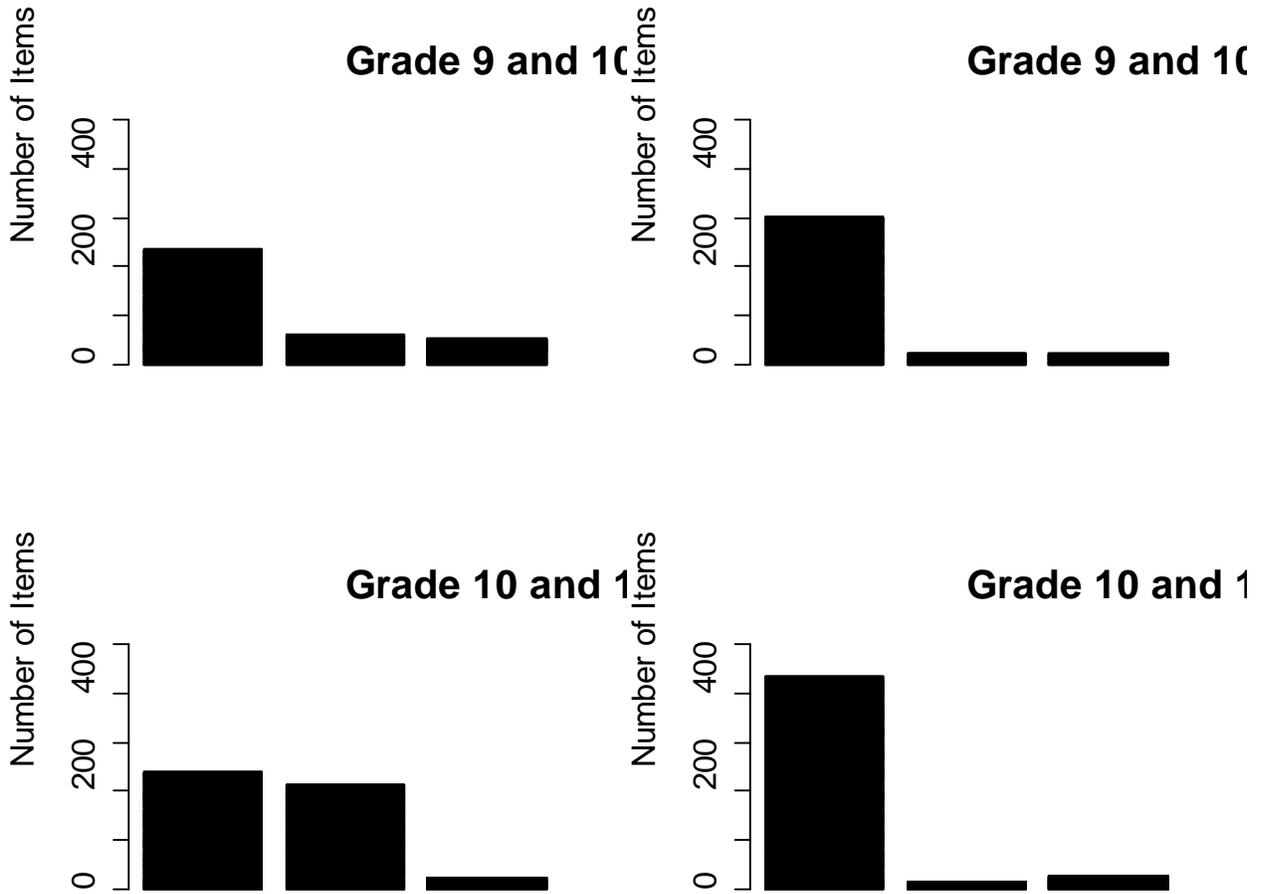


Figure 58. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (across grades)

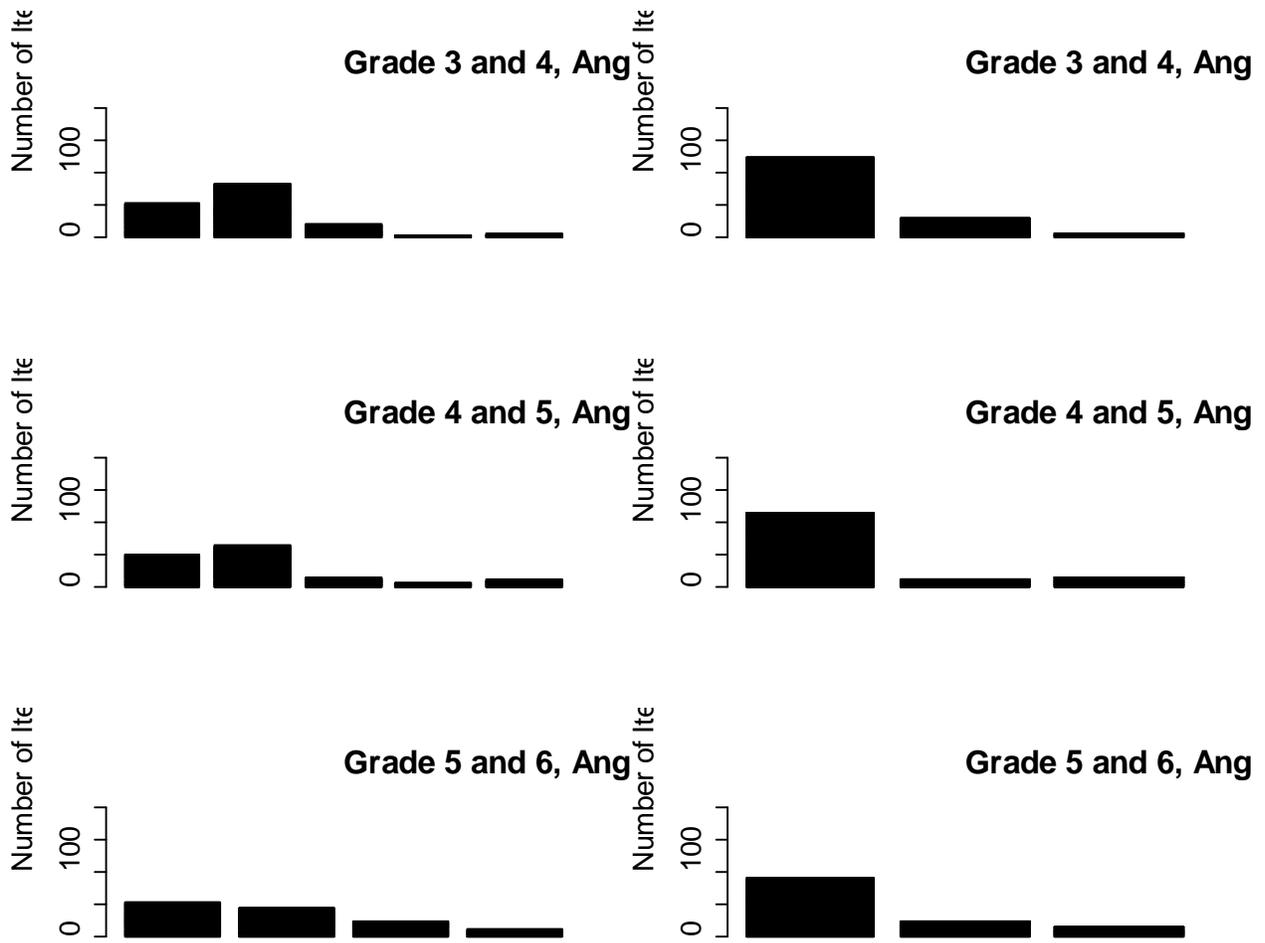


Figure 59. Clustering of Item Angle Measures for Grades 3 to 6, ELA/literacy (vertical linking)

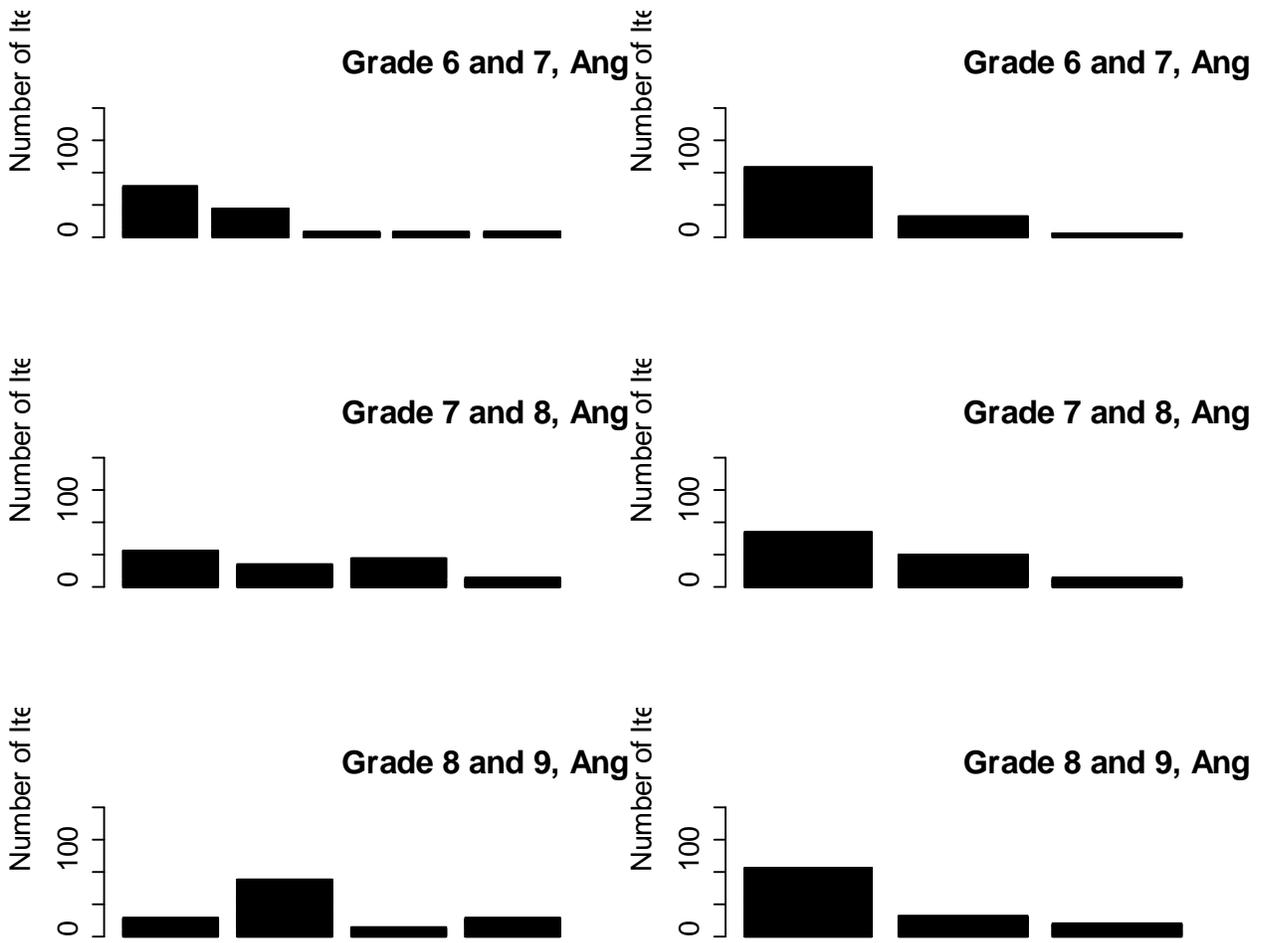


Figure 60. Clustering of Item Angle Measures for Grades 6 to 9, ELA/literacy (vertical linking)

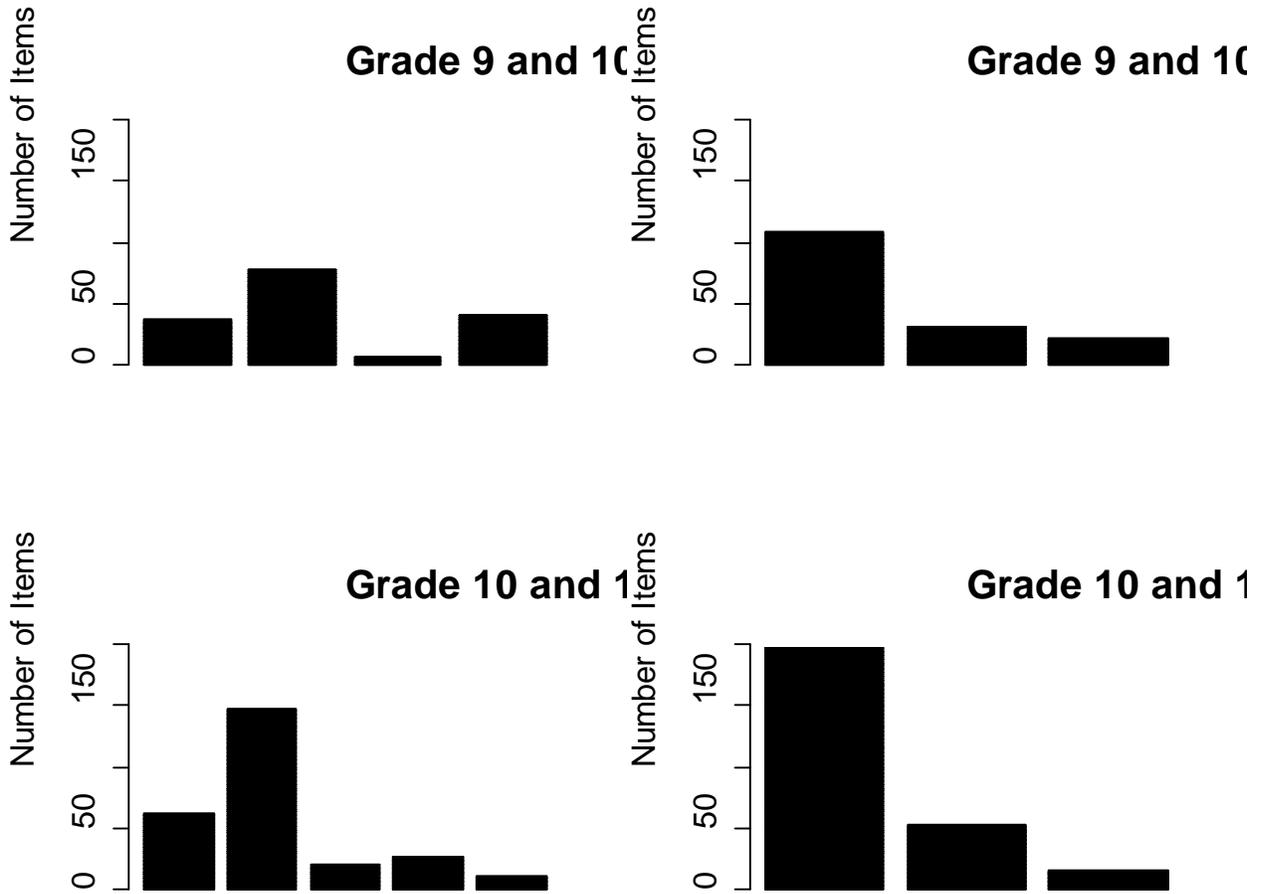


Figure 61. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (vertical linking)

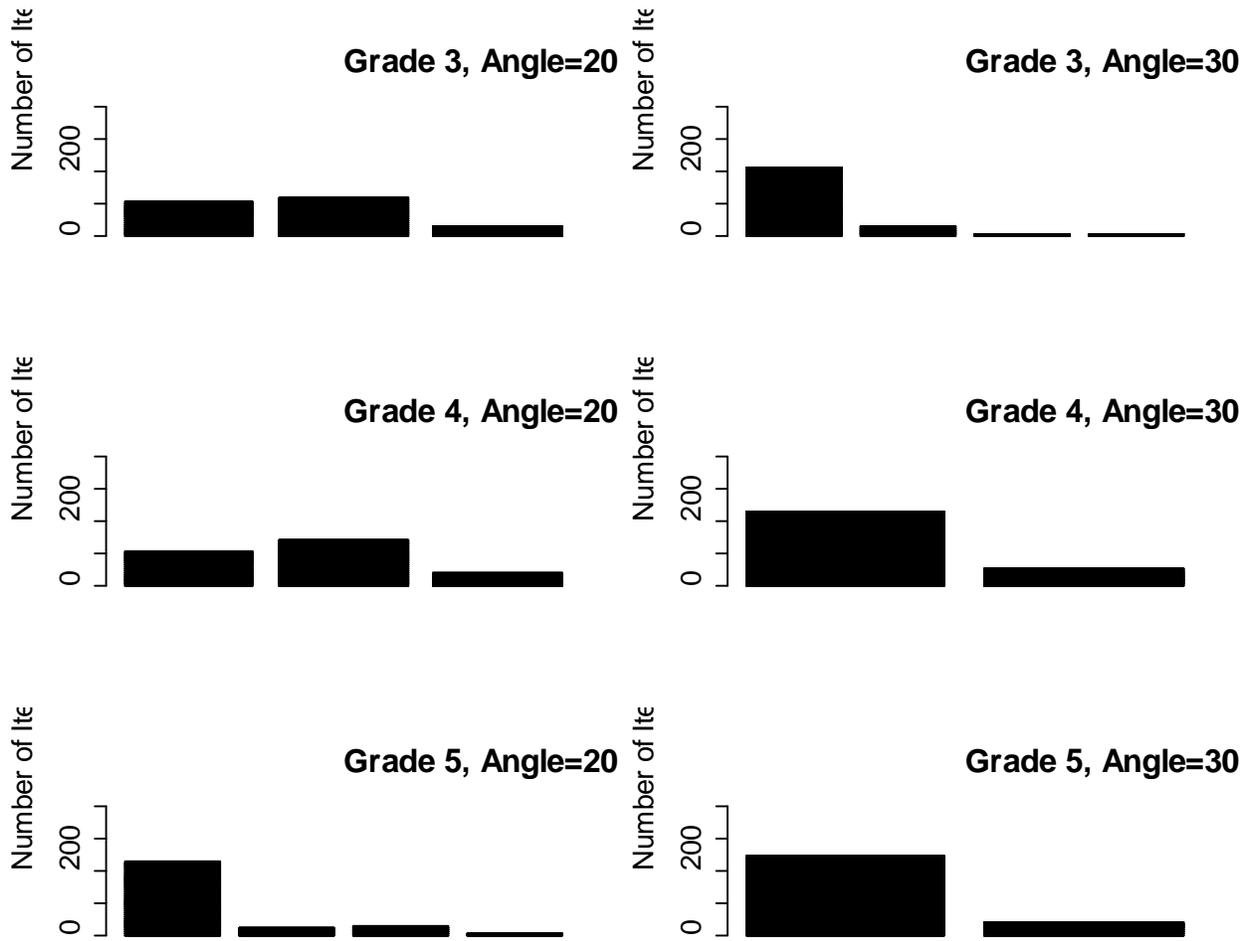


Figure 62. Clustering of Item Angle Measures for Grades 3 to 5, Mathematics (within grade)

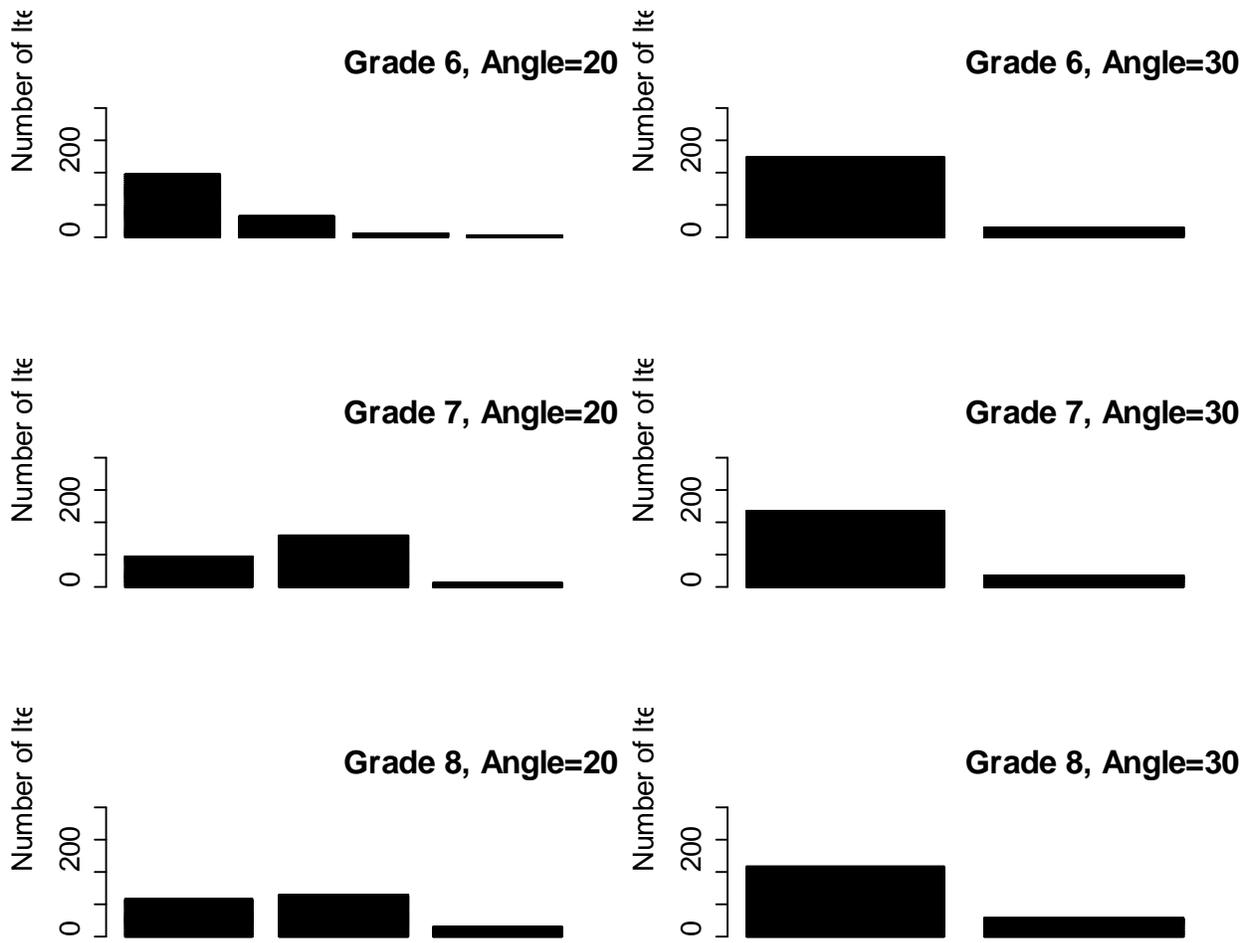


Figure 63. Clustering of Item Angle Measures for Grades 6 to 8, Mathematics (within grade)

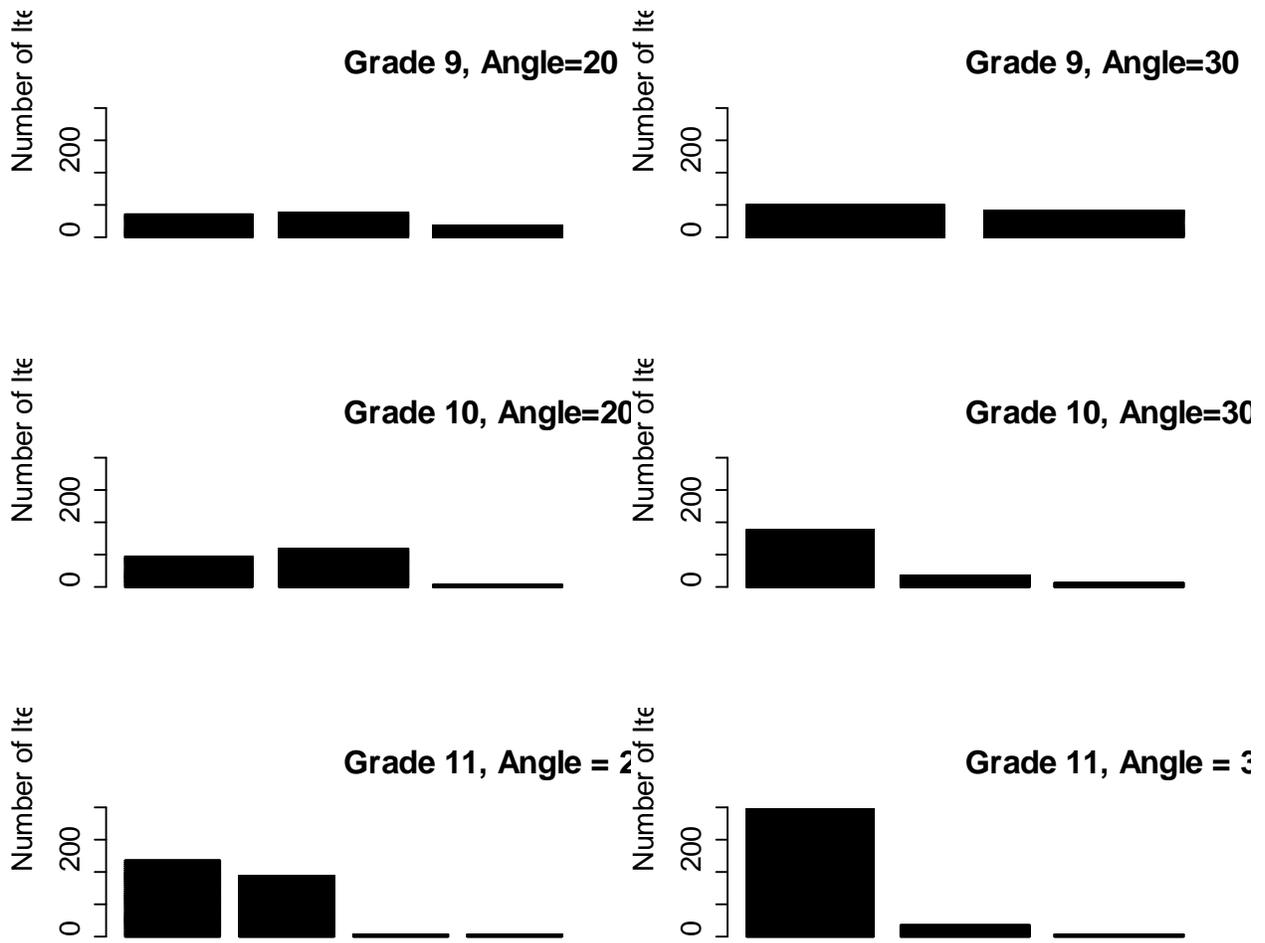


Figure 64. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (within grade)

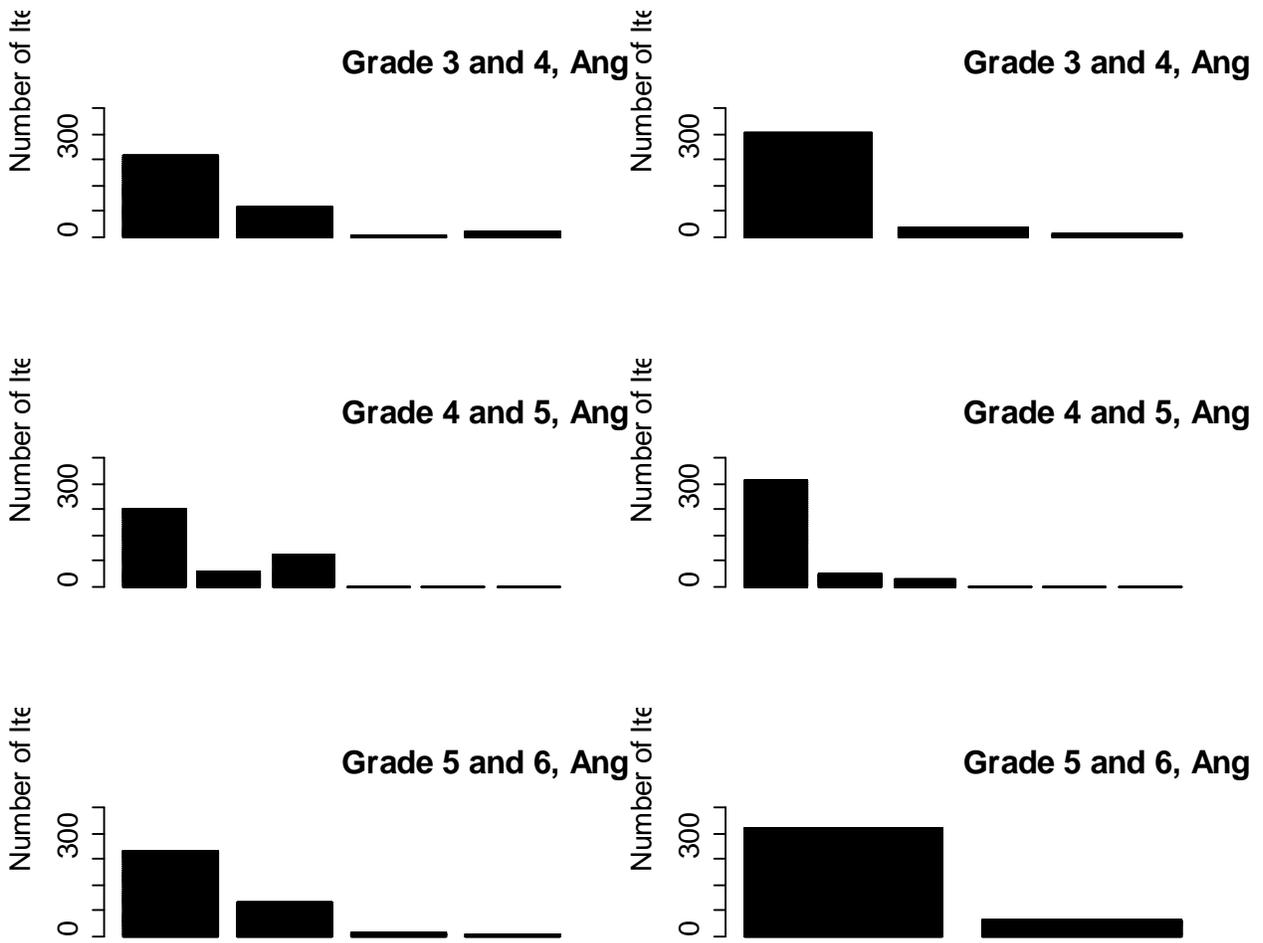


Figure 65. Clustering of Item Angle Measures for Grades 3 to 6, Mathematics (across grades)

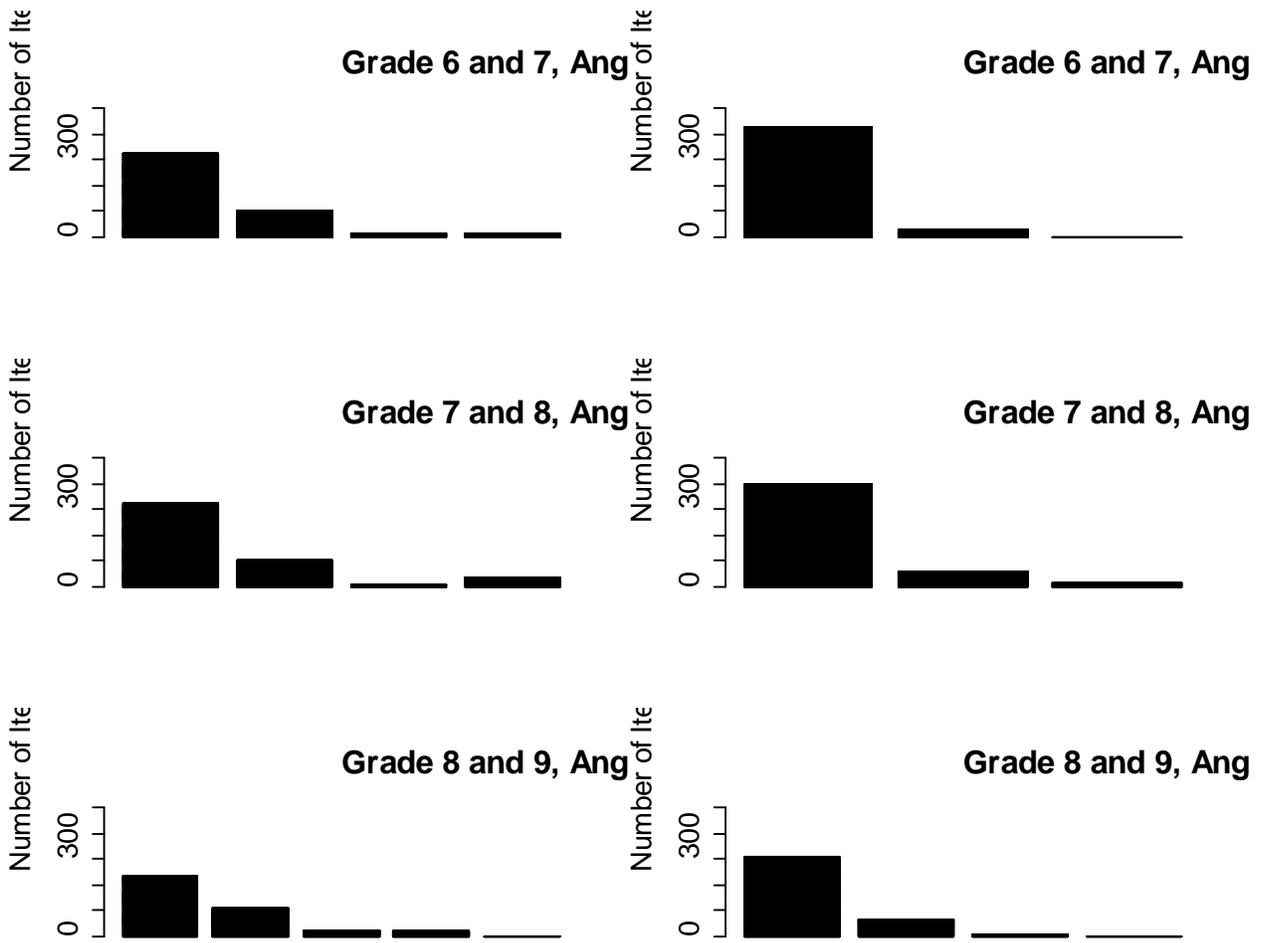


Figure 66. Clustering of Item Angle Measures for Grades 6 to 9, Mathematics (across grades)

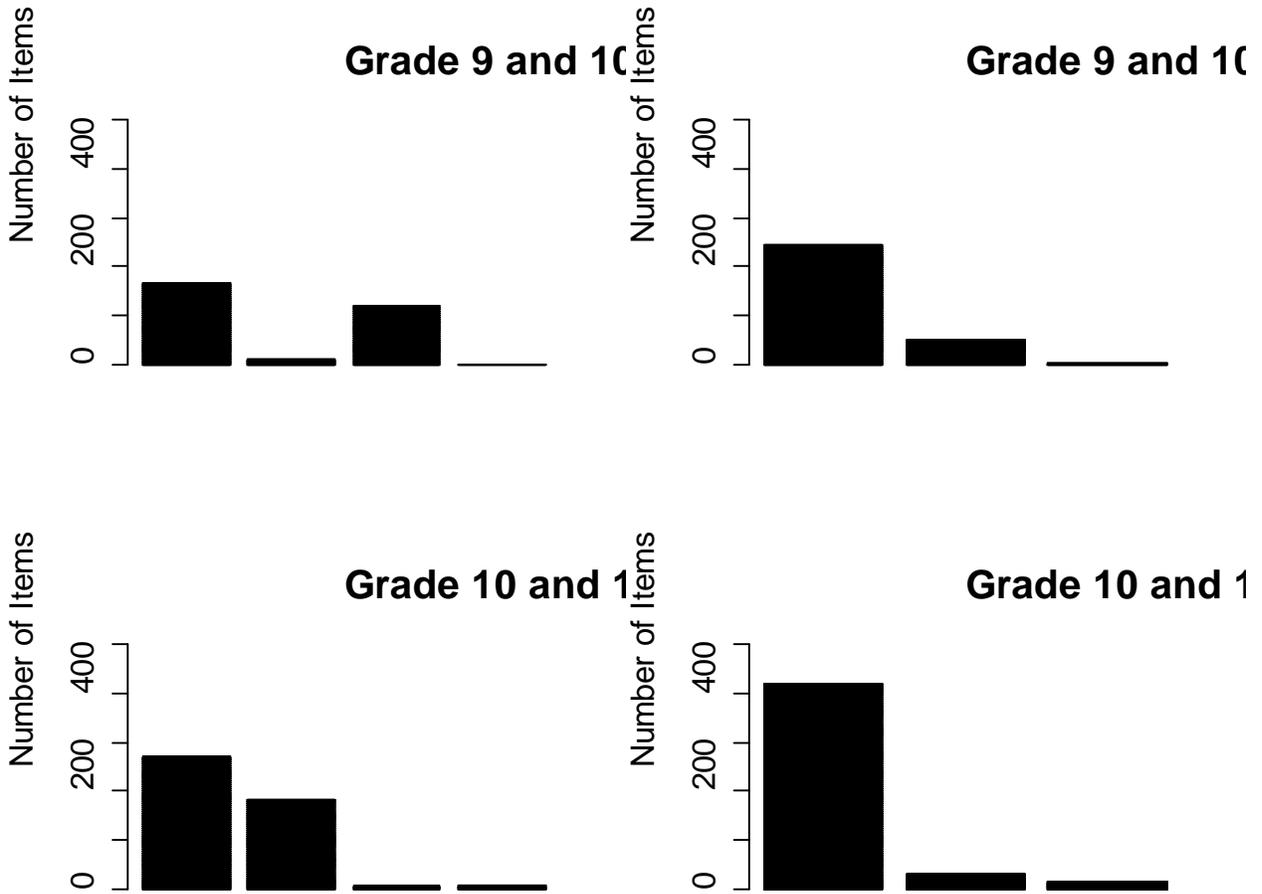


Figure 67. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (across grades)

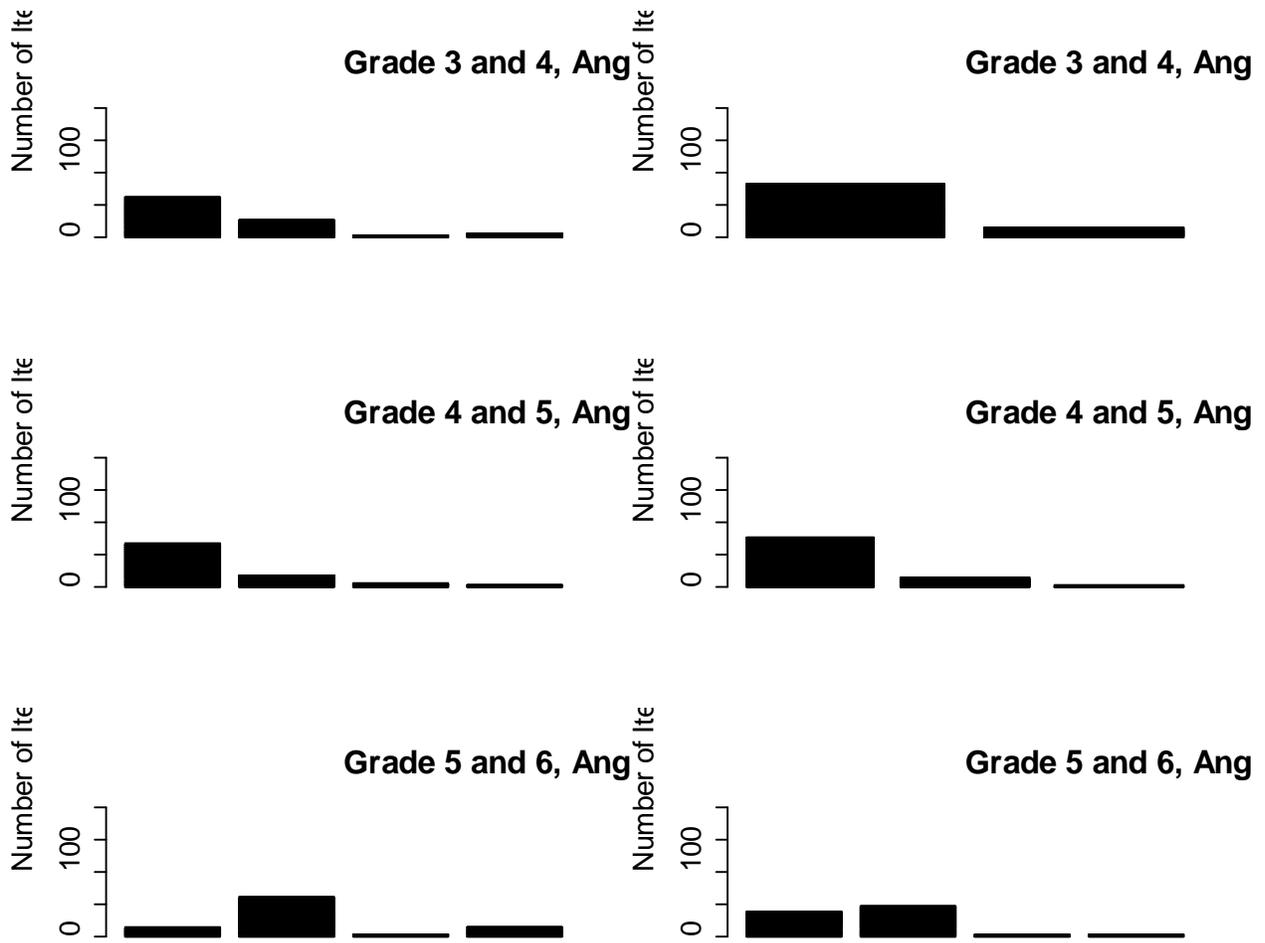


Figure 68. Clustering of Item Angle Measures for Grades 3 to 6, Mathematics (vertical linking)

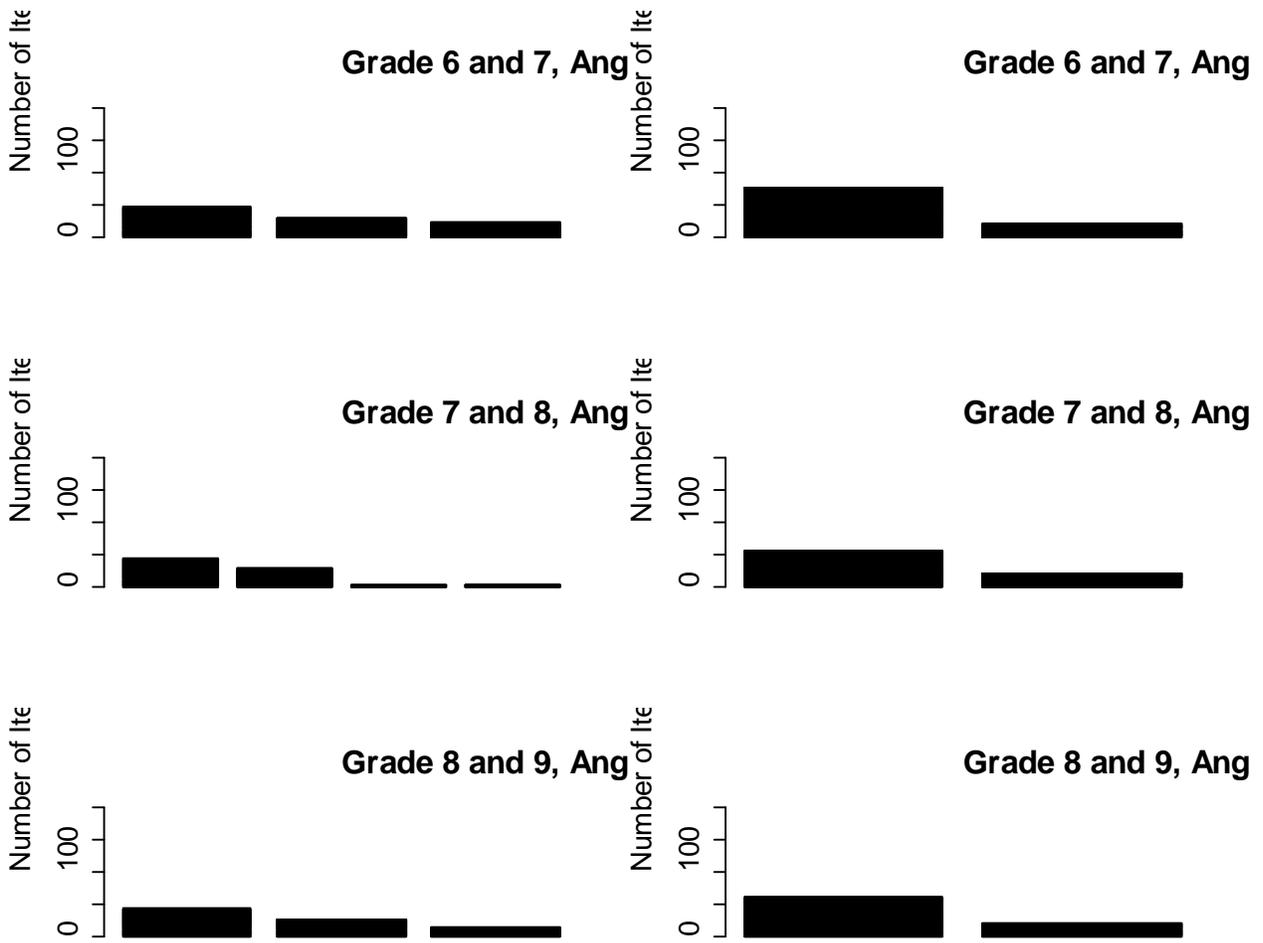


Figure 69. Clustering of Item Angle Measures for Grades 6 to 9, Mathematics (vertical linking)

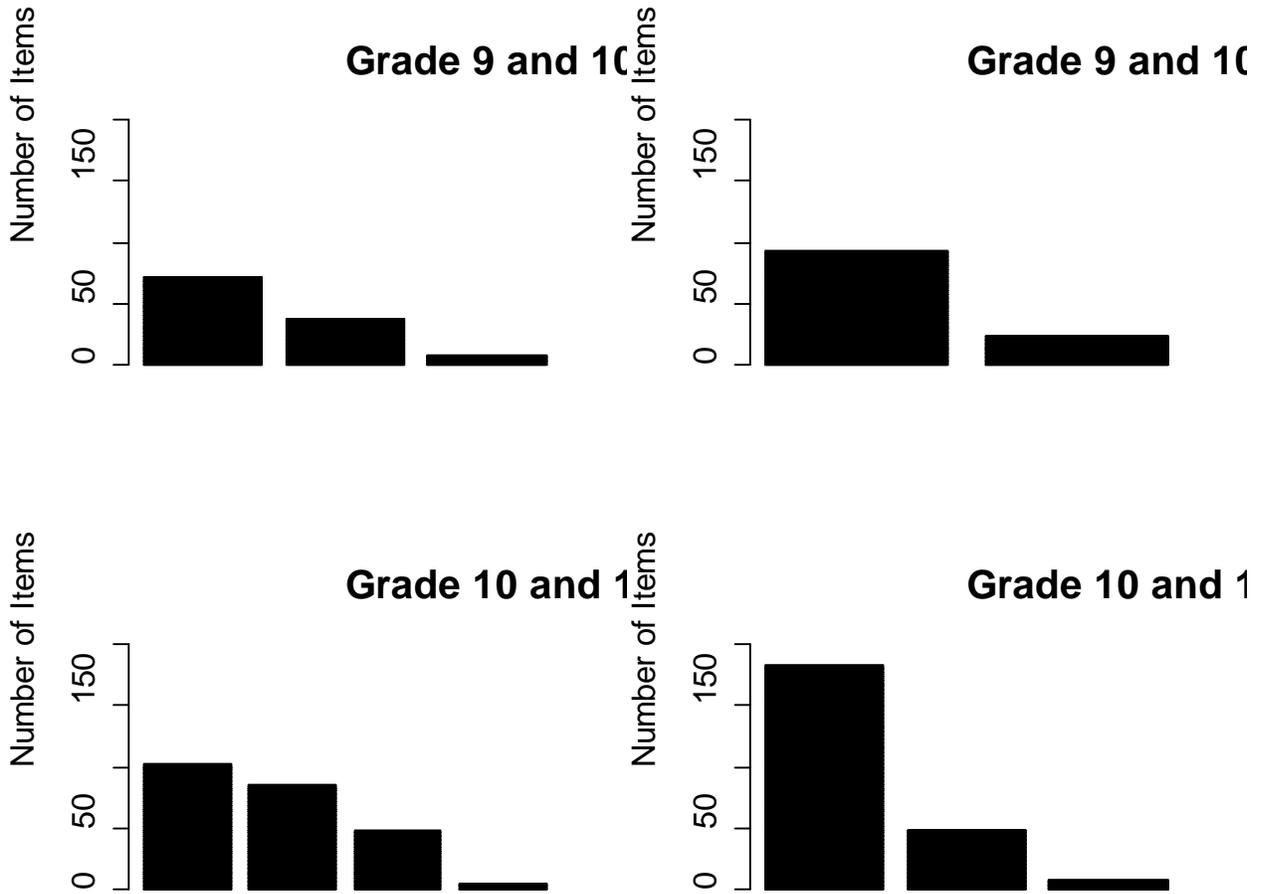


Figure 70. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (vertical linking)

Item Response Theory (IRT) Model Comparison

Within the family of IRT models, there are two major choices to be made:

1. use of a unidimensional or multidimensional model and
2. within the category of unidimensional models, the use of a Rasch one-parameter/partial credit model (Rasch/PC) combination, a two-parameter logistic/generalized partial credit model (2PL/GPC) combination, or a three-parameter logistic/generalized partial credit (3PL/GPC) combination.

It is highly desirable that a unidimensional model be used since the properties of these models are well known for scaling and are ones that have been used extensively in K-12 programs to make critical decisions concerning students, teachers, and schools. Also, the IRT models selected must be implemented in the context of an operational CAT. A multidimensional CAT with many constraints and performance tasks would be more difficult to implement and maintain.

This model comparison study has the limitations shared by the dimensionality in its reliance on Pilot data. The number and types of items and the scale properties changed significantly from the Pilot to the Field Test. The dimensionality study results from the previous section suggest that a unidimensional IRT model with a single vertical scale within each content area could be used. Three unidimensional IRT model combinations were evaluated for dichotomous and polytomous item calibration. Specifically, these combinations are the Rasch one-parameter/partial credit model (1PL/PC) combination, the two-parameter logistic/generalized partial credit model (2PL/GPC) combination, and the three-parameter logistic/generalized partial credit model (3PL/GPC) combination. Calibration and scaling results based on all three IRT model combinations are presented and compared, and they are used for making recommendations for IRT model choice for the Field Test and operational use and for determining the set of item parameter estimates to be stored in the item bank.

The Smarter Balanced assessment includes CAT-selected and constructed-response items, and items associated with performance tasks. For selected-response items, a 3PL, 2PL, or 1PL or Rasch model is used. The 3PL model is given by

$$P_i(\theta_j) = c_i + (1 - c_i) / \left\{ 1 + \exp \left[-Da_i(\theta_j - b_i) \right] \right\}$$

where $P_i(\theta_j)$ is the probability of a correct response to item i by a test taker with ability θ_j ; a_i , b_i , and c_i are the discrimination, difficulty, and lower asymptote parameters, respectively, for item i ; and

D is a constant that puts the θ ability scale in the same metric as the normal ogive model ($D = 1.7$). The 3PL model can be constrained to equal the Rasch model by setting the discrimination parameter to $1/D$ and the c parameter to 0. If the discrimination parameter is free to vary by item and $c_i = 0$, then the 2PL model results.

For constructed-response items, the generalized partial credit model (Muraki, 1992) or partial credit model (Masters, 1982) is employed. The generalized partial credit model is given by

$$P_{ih}(\theta_j) = \frac{\exp \left[\sum_{v=1}^h Da_i(\theta_j - b_i + d_{iv}) \right]}{\sum_{c=1}^{n_i} \exp \left[\sum_{v=1}^c Da_i(\theta_j - b_i + d_{iv}) \right]}$$

where $P_h(\theta_j)$ is the probability of test taker j obtaining a score of h on item i , n_i is the number of score categories item i contains, b_i is the location parameter for item i , d_{iv} is the category parameter for item i for category v , and D is a scaling constant. The generalized partial credit model can be constrained to equal the partial credit model by setting the discrimination parameter to $1/D$. The generalized partial credit model is equivalent to the two-parameter partial credit model used in the dimensionality study in the previous section (Yen and Fitzpatrick, 2006).

The choice of a family of IRT models within a unidimensional framework should include several considerations consisting of model simplicity, model fit, model stability, and reasonableness.

- **Model simplicity or parsimony.** Model selection should balance goodness-of-fit and model simplicity. The Rasch model is simpler than the 2PL/GPC and 3PL/GPC and has worked well in many K-12 applications. The Rasch one parameter logistic (1-PL) model is the most parsimonious followed by the 2-PL and 3-PL models. Likewise, Master's partial credit (1982) is a more parsimonious model than the generalized version, which includes an item specific discrimination parameter.
- **Model fit.** Because the 3PL/GPC is a more general model, it provides better statistical model fit than the 2PL/GPC and the 1PL/PC; the 2PL/GPC provides better fit than 1PL/PC. Often, the improvement in fit from 2PL to 3PL can be far smaller than from 1PL to 2PL (Haberman, 2010). However, statistical model fit, by itself, is not a sufficient basis for model choice. The practical implications of model choice should also be considered. For example, for CAT administration that delivers items targeted at a specific student's ability level, fit of the IRT item characteristic curve (ICC) in the middle range may be more consequential than fit of the curve at the two ends. The primary practical implication of model misfit is a systematic difference between observed and predicted item characteristic functions, which affects the accuracy of scoring (i.e., the relationship of raw scores and trait estimates). Some item properties that affect model fit include the following:
 - Discriminations that vary systematically by item difficulty or trait level. Rasch model assumes that the discrimination is constant across all items and that item discrimination is uncorrelated with item difficulty. By examining plots or correlations of item discrimination versus item difficulty for the 2PL/GPC, one can determine if the Rasch assumption is suitable for the Smarter Balanced assessments. This result affects vertical scaling, since item discriminations for the same items are administered across grade levels.
 - Discriminations that vary systematically by item type (SR versus CR), number of score categories or claims. Constructed-response items with multiple score levels and/or ones based on the sum of multiple raters might be expected to have varying discriminations and may not be adequately represented by the Rasch model (Sykes & Yen, 2000; Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996). The results of the 2PL/GPC can be examined to see if there is a systematic relationship between item type/number of score categories/claim area and item discrimination.
- **Model stability.** Holland (1990) indicated that unconstrained three-parameter logistic (i.e., 3-PL) models are expected to have stability problems. His study revealed that in the typical case of a standard normal prior, a unidimensional IRT model for dichotomous responses can be approximated by a log-linear model with only main effects and interactions. For a test of q items, the approximation is determined by $2q$ parameters, while the 3PL model would require $3q$ parameters. This stability issue can be addressed by having appropriate priors on

the c parameters, including holding them constant at logical values, particularly when sample sizes are small.

- **Reasonableness of the vertical scale.** Since the selected IRT model will be used to establish a vertical scale, it is important to evaluate the reasonableness of the vertical scale, including expected growth from one grade to another, before making final decisions on the model for adoption. As suggested by research, the choice of the IRT scaling model may shrink or stretch out a measurement scale (Yen, 1981) and will impact how growth is depicted by the vertical scale (Briggs & Weeks, 2009). Both the Rasch and 3PL have been used for developing K-12 vertical scales, and in the last two decades, their scale properties have been broadly accepted by K-12 users (Yen & Fitzpatrick, 2006).

To support the Smarter Balanced Consortium in the IRT model selection process, the following results, including dimensionality analysis, IRT calibration, fit comparison, guessing evaluation, common discrimination evaluation, and ability estimates and results, are provided using the Pilot data. Both ELA/literacy and mathematics results are described. However, mathematics performance-task items were not included in the analysis. A considerable portion of the Pilot Test vertical linking items administered to upper grade levels showed reverse growth patterns likely due to common core implementation differences. That is, items were harder in an upper grade and easier in a lower one. Given these vertical linking item issues, it was not productive to evaluate the reasonableness of the vertical scale as part of this model comparison analyses. For this reason, vertical scaling results were not provided as part of the model comparison analysis at this time.

IRT Data Step

The additional IRT related data steps described below were conducted prior to performing the calibrations. As stated previously, students took either multiple CAT components or a combination of CAT items and a performance task during the Pilot Test administration. The CAT or performance task administered might be on-grade or off-grade to facilitate vertical linking, but each participating student was administered at least one on-grade CAT component. Performance tasks were included in the ELA/literacy IRT model comparison analyses but not for mathematics. The first step was to create a sparse data matrix for IRT analysis reflecting item scores as well as missing item information by design. For a given grade, the dimension of the sparse matrix is the total number of students times the total number of unique items (i.e., scorable units). The remaining cells, representing items not administered to a student, were treated as “not presented” items in the IRT calibration. The following item exclusion rules were implemented:

- Items that have no scored responses, or items that have scored responses in only one category, were excluded.
- CAT items that have on-grade item total correlations < 0.15 were removed from on-grade *AND* off-grade data sets regardless of their off-grade performance.
- CAT items that have been recommended for “rejection” per content experts during data review meetings were removed from on-grade *AND* off-grade data sets.
- Performance task items that have negative on-grade item total correlations were removed from on-grade *AND* off-grade data sets.
- CAT or performance task items with negative off-grade but reasonable on-grade item-total correlations were removed from the specific off-grade data sets only. For the dimensionality study, off-grade responses were calibrated together with on-grade responses in one part of the study.

The following item score category treatments for constructed-response were followed:

- Categories that have a reversed pattern of average criterion score progression (i.e., the average criterion score for a lower score category was higher than the average criterion score for a higher score category) at the on-grade level were collapsed in both on-grade *AND* off-grade data sets.
- Categories with fewer than ten test takers at on-grade level were collapsed with neighboring categories in both on-grade *AND* off-grade data sets. If the score category that needed to be collapsed was a middle category, it was collapsed with the adjacent lower score category.
- Categories that had a reversed pattern of average criterion score progression (i.e., the average criterion score for a lower score category was higher than the average criterion score for a higher score category) at the off-grade level but not at the on-grade level were collapsed in the specific off-grade data sets.
- Categories with fewer than ten test takers at the off-grade level but ten or more test takers at the on-grade level were collapsed with neighboring categories in the specific off-grade data sets.

Of all the items that required category collapsing due to sparse responses, more than 70 of them had fewer than 1,500 valid responses from the Pilot administration. The number of CAT/performance task items that entered into IRT analyses after the application of these rules and the student sample sizes associated with them are presented in Tables 30 to 33. Table 30 shows the number of items dropped due to implementing these rules. Table 31 shows the overall number of items with collapsed score levels by content area. Tables 32 and 33 present further detail on item collapsing by grade, vertical linking grade (off-grade), and item type. Linking grade refers to the off-grade item administration for vertical scaling. For the most part, items had collapsed score levels due to no or insufficient number of student responses in the highest (hardest) category. Table 34 shows the number of item by type for ELA/literacy and mathematics that contributed to the IRT calibration. Table 35 shows descriptive statistics that include the percentile distribution for the number of student observations per item in ELA/literacy and mathematics. For these items, there was a large variation in the number of student observations per item. A small percentage of items did not have sufficient observations for accurate calibration.

Table 30. Number of Items Dropped from the Calibration (On-grade).

Grade	ELA/literacy	Mathematics
3	10	5
4	19	5
5	9	6
6	25	24
7	15	40
8	30	33
9	20	32
10	24	17
11	26	43

Table 31. Number of Constructed-response and Performance tasks with Collapsed Score Levels (On-grade).

Grade	ELA/literacy	Mathematics
3	13	2
4	13	0
5	10	3
6	15	2
7	10	1
8	16	1
9	16	1
10	8	1
11	18	8

Table 32. Number of Constructed-response and Performance tasks with Collapsed Score Levels for ELA/literacy (Detail).

Grade	Linking Grade	Item Type	No. Collapsed
3	3	CAT	2
	3	PT	11
4	3	CAT	1
	4	PT	13
5	5	PT	10
	6	PT	2
6	6	CAT	2
	6	PT	13
7	6	CAT	2
	7	PT	10
	8	CAT	3
8	8	CAT	6
	8	PT	10
9	8	CAT	1
	9	PT	16
	10	CAT	2
10	9	CAT	2
	10	PT	8
	11	CAT	2
	11	PT	1
11	9	CAT	3
	10	CAT	7
	11	CAT	5
	11	PT	13

Table 33. Number of Constructed-response with Collapsed Score Levels for Mathematics (Detail).

Grade	Linking Grade	No. Collapsed
3	3	2
4		NA
5	5	3
6	6	2
	7	1
7	7	1
	8	3
8	8	1
	9	1
9	8	3
	9	1
	10	3
10	9	6
	10	1
	11	4
11	9	5
	10	9
	11	8

Table 34. Number of ELA/literacy and Mathematics Items in the IRT Calibration.

Grade	Item Grade	ELA/Literacy			Mathematics
		Total	CAT	PT	Total (CAT only)
3	3	231	200	31	207
	4	48	44	4	47
4	3	48	44	4	38
	4	217	179	38	209
	5	36	35	1	37
5	4	40	36	4	41
	5	175	144	31	204
	6	34	31	3	39
6	5	23	23		41
	6	202	161	41	189
	7	38	36	2	48
7	6	37	35	2	41
	7	195	163	32	190
	8	43	41	2	37
8	7	38	36	2	33
	8	202	168	34	191
	9	39	39		47
9	8	38	35	3	23
	9	126	80	46	103
	10	46	46		56
10	9	41	41		51
	10	133	109	24	122

Grade	Item Grade	ELA/Literacy			Mathematics
		Total	CAT	PT	Total (CAT only)
	11	50	48	2	48
11	9	80	80		80
	10	107	107		90
	11	261	221	40	263

Table 35. Descriptive Statistics for Number of Students per Item for ELA/literacy and Mathematics.

Percentile												
Grade	No. Items	Min	1st	10th	25th	50th	75th	90th	99th	Max	Mean	SD
ELA/literacy												
3	279 (35*)	864	986	1,377	3,333	4,450	4,794	8,464	9,846	9,846	4,451	2,380
4	301 (43)	897	943	1,171	1,642	4,130	6,346	10,342	16,301	16,343	4,467	3,318
5	249 (38)	950	1,099	1,226	1,419	4,050	4,311	8,591	18,347	18,373	4,177	3,384
6	263 (43)	929	1,124	1,342	1,421	4,678	5,009	8,682	12,721	12,760	4,202	2,699
7	275 (36)	1,060	1,066	1,117	1,555	3,824	4,042	7,893	8,653	12,078	3,603	2,317
8	279 (36)	492	511	1,009	1,074	2,059	4,382	8,152	13,072	13,077	3,316	2,874
8	210 (49)	553	591	662	1,197	1,344	4,848	4,945	5,008	5,008	2,624	1,860
10	224 (26)	369	401	511	556	1,197	2,915	2,965	3,013	3,013	1,693	1,167
11	448 (40)	249	251	271	291	1,423	1,674	3,362	3,706	3,729	1,219	1,026
Mathematics												
3	254	416	431	1,772	3,540	4,360	5,335	6,245	14,008	14,735	4,305	2,315
4	284	497	498	1,970	4,030	4,702	4,792	4,827	9,633	9,642	4,342	2,054
5	284	496	496	2,164	4,335	5,019	5,125	10,110	10,336	10,338	4,842	2,349
6	278	483	494	1,872	1,991	4,403	4,515	4,610	9,209	9,213	3,939	1,953
7	267	441	454	946	1,074	3,206	3,906	7,527	10,743	11,138	3,533	2,483
8	271	473	481	1,471	1,696	4,152	4,346	5,350	8,542	8,556	3,858	1,996
9	182	484	494	1,352	1,422	2,794	3,451	5,654	6,496	6,497	2,761	1,437
10	221	493	497	700	764	1,705	2,125	2,162	3,877	3,889	1,538	847
11	433	422	456	569	607	1,426	1,882	2,594	3,889	5,407	1,438	889

Note: * refers to the number of performance task items for ELA/literacy.

IRT Model Calibration

IRT calibration was conducted based on 1PL/PC, 2PL/GPC, and 3PL/GPC model combinations using **PARSCALE** (Muraki & Bock, 2003). **PARSCALE** properties are well known, and a variety of unidimensional IRT models can be implemented with it.

Additional Rules for Items in the Calibration. Some additional IRT based rules were necessary in the case of item nonconvergence or unreasonably large standard errors for item parameter estimates. Nonconvergence was defined by either not achieving the criterion of largest parameter change lower than 0.005 or an erratic pattern of $-2\log$ likelihood values. Standard errors were evaluated as part of the reasonableness procedures. Calibration issues in the Pilot Test analyses were caused by the following issues.

- Local item dependence (LID). Many performance tasks for writing scores (i.e., long-writes) were highly correlated. These items involved the same student responses scored with different trait rubrics. The local item dependence made these items appear highly discriminating and caused problems for **PARSCALE** in locating slope parameter estimates.
- Low item discrimination. While CAT items with item-total correlations lower than 0.15 were removed from the pool, items with poor IRT discrimination, especially ones that are difficult, caused convergence issues in calibrations using the 3PL model.
- Guessing parameter indeterminacy in the 3PL model. Starting values for the “guessing” sometimes lead to large standard errors for difficulty estimates (> 1.0) or unreasonable guessing parameter estimates (zero guessing parameter estimates associated with standard errors larger than 0.04).

To address these calibration issues and permit accurate estimation, the following rules were implemented when a specific item was identified as being problematic.

For selected-response items:

- For the 3PL model, the guessing parameter starting values were changed. First, the guessing parameter starting values were changed to 0.25, then 0.10, and finally to 0.0, if calibration issues persisted.
- For the 3PL model, the guessing parameter was held at a fixed value if changing the guessing parameter starting value did not solve the calibration issues. The guessing parameter was first fixed to 0.25, next to 0.10, and finally to 0.0, if estimation issues persisted.
- If none of the above actions solved the calibration issue, then the item was removed.

For constructed-response items:

- Starting values were changed for the item. For polytomous items, there is an option to use category starting values that are constant values for “scores for ordinal or ranked data” instead of the **PARSCALE** default category starting values.
- Score categories were collapsed for polytomous items.
- If none of the above steps solved the calibration issue, then the item was removed.
- Usually when **PARSCALE** encountered convergence issues due to local item dependence, one item trait score out of the pair was removed for the trait scoring of writing (i.e., long-writes).

No items were deleted from the 1PL analyses and a few items were deleted from the 2PL analyses, largely due to local item dependence issues. The additional item steps in 3PL model analyses were primarily due to c-parameter estimation issues. As a result, there were some differences in the item sets included in the following results comparing the three models.

Under each model combination, IRT parameter estimates as well as standard errors associated with them, and item goodness-of-fit results were evaluated as were the ability parameter estimates. In

general, convergence under each IRT model combination was reached and the resulting IRT item/ability parameter estimates under each model combination were reasonable.

IRT Model Fit Comparison

To allow comparison of item fit across different IRT model combinations, PARSCALE G^2 statistics were evaluated. In PARSCALE, a likelihood ratio G^2 test statistic can be used to compare the frequencies of correct and incorrect responses in the intervals on the θ continuum with those expected based on the fitted model (du Toit, 2003):

$$G_i^2 = 2 \sum_{h=1}^{n_g} \left[r_{ih} \log_e \frac{r_{ih}}{N_h P_i(\bar{\theta}_h)} + (N_h - r_{ih}) \log_e \frac{N_h - r_{ih}}{N_h (1 - P_i(\bar{\theta}_h))} \right],$$

where n_g is the total number of intervals, r_{ih} is the observed frequency of correct responses to item i in interval h , N_h is the number of students in interval h , $\bar{\theta}_h$ is the average ability of students in interval h , and $P_i(\bar{\theta}_h)$ is the value of the fitted response function for item i at $\bar{\theta}_h$.

Since the G^2 statistic tends to be sensitive to sample size (i.e., flagging more items with larger sample size), it is used as a descriptive statistic in this study instead of one for significance testing. Since there are many items for any grade/content area combination, the distributions of G^2 are compared across IRT model combinations. Tables 36 and 37 present the summary of G^2 statistics across 1PL/PC, 2PL/GPC, and 3PL/GPC models for ELA/literacy and mathematics, respectively. Although G^2 statistics may not be strictly comparable across models due to the difference in degrees of freedom, the size of the G^2 statistics in general still provides some evidence for comparing fit across models, considering that the degrees of freedom for each item is roughly comparable across different models. The tables show that for most of the tests the mean value of G^2 for the 1PL/PC is substantially greater than the mean values for the other two model combinations, indicating considerable average improvement in fit with 2PL/GPC and 3PL/GPC in comparison with 1PL/PC.

Table 36. Summary of G^2 Statistics of On-Grade ELA/literacy Items across 1PL, 2PL, and 3PL IRT Models.

Item Grade	1PL/PC			2PL/GPC			3PL/GPC		
	No. of Items	G^2 Mean	G^2 SD	No. of Items	G^2 Mean	G^2 SD	No. of Items	G^2 Mean	G^2 SD
3	231	151	114	231	79	58	231	79	60
4	217	128	93	216	72	38	216	70	41
5	175	121	87	171	75	42	171	73	43
6	202	132	99	197	79	51	197	78	51
7	195	127	87	190	84	57	190	84	58
8	202	135	118	199	85	73	199	84	73
9	126	103	67	119	72	44	119	72	45
10	133	93	56	129	63	31	129	62	33
11	261	79	48	259	57	34	259	57	35

 Table 37. Summary of G^2 Statistics of On-Grade Mathematics Items across 1PL, 2PL, and 3PL IRT Models.

Item Grade	1PL/PC			2PL/GPC			3PL/GPC		
	No. of Items	G^2 Mean	G^2 SD	No. of Items	G^2 Mean	G^2 SD	No. of Items	G^2 Mean	G^2 SD
3	207	127	88	207	86	58	207	84	58
4	209	139	99	209	92	82	209	90	84
5	204	167	127	204	95	77	204	93	80
6	189	145	106	189	96	69	189	93	69
7	190	162	123	190	113	94	190	110	97
8	191	152	111	191	110	86	191	114	99
9	103	111	66	103	95	62	103	94	60
10	122	97	52	122	71	42	122	71	44
11	263	72	58	263	72	88	263	68	74

Guessing Evaluation

The single-selection selected-response items in the Pilot Test had four answer choices. Since 1PL and 2PL models assume minimal guessing, the amount of guessing involved for selected-response items is evaluated by examining the size of guessing parameter estimates under the 3PL/GPC model combinations. Large guessing parameter estimates provide evidence for the use of 3PL models, and small guessing parameter estimates allow the possible use of 1PL and 2PL models. Tables 38 and 39 present the mean, standard deviation, minimum, maximum, and range of guessing parameter estimates for items administered on-grade for ELA/literacy and mathematics, respectively. Results indicate that the average guessing is below .20 for most tests. The range of the guessing values showed a consistent pattern across grade levels in that the majority of selected-response items had guessing parameter estimates below .20 but greater than .10.

Table 38. Summary of Guessing Parameter Estimates for On-Grade ELA/literacy Items.

Grade	No. of Items	c Estimate Summary				c Estimate Range			
		Mean	SD	Min	Max	0–0.10	0.10–0.20	0.20–0.30	>0.30
3	76	0.16	0.07	0.06	0.39	16	43	14	3
4	111	0.17	0.07	0.04	0.36	20	53	31	7
5	77	0.15	0.07	0.00	0.31	16	40	20	1
6	75	0.15	0.07	0.05	0.33	23	35	14	3
7	76	0.18	0.07	0.06	0.38	9	39	25	3
8	77	0.15	0.07	0.00	0.34	16	46	10	5
9	36	0.16	0.08	0.04	0.31	10	15	9	2
10	46	0.16	0.08	0.00	0.35	9	24	10	3
11	91	0.18	0.07	0.04	0.39	12	48	25	6

Table 39. Summary of Guessing Parameter Estimates for On-Grade Mathematics Items.

Grade	No. of Items	c Estimate Summary				c Estimate Range			
		Mean	SD	Min	Max	0–0.10	0.10–0.20	0.20–0.30	>0.30
3	34	0.18	0.07	0.05	0.36	3	21	8	2
4	31	0.17	0.06	0.03	0.29	3	18	10	0
5	39	0.18	0.10	0.02	0.43	13	10	11	5
6	41	0.21	0.09	0.08	0.38	5	14	13	9
7	31	0.20	0.08	0.07	0.39	3	12	13	3
8	34	0.18	0.07	0.07	0.32	3	18	10	3
9	14	0.20	0.08	0.09	0.35	1	8	3	2
10	19	0.26	0.11	0.06	0.46	2	3	8	6
11	32	0.19	0.08	0.05	0.37	4	15	9	4

Common Discrimination Evaluation

The Rasch model assumes common item discrimination across all items. Analyses were conducted to evaluate if item discrimination varied systematically with difficulty, item type (SR vs. CR), number of item score categories, or by claim. This evaluation was done by plotting item discrimination versus item difficulty estimates from the 2PL/GPC model. When the distribution of item discrimination is reasonably homogeneous, the selection of a model that assumes equal item discrimination may be viable. An advantage of the 2PL/GPC in comparison to the 1PL/PC is that it would permit using items with a range of item discriminations, while the 1PL/PC might flag items with both very high and very low discriminations for exhibiting poor fit and requiring further content review.

Tables 40 and 41 summarize discrimination and difficulty parameter estimates and correlations between them under the 2PL/GPC for ELA/literacy and mathematics items administered on-grade. These summary statistics are provided for the overall set of items as well as groups of items characterized by item type, score categories, and claim areas. Figures 71 and 72 present, for ELA/literacy and mathematics and at each grade level, plots of item discrimination versus item difficulty under the 2PL/GPC with item type, score category, and claim area highlighted for each item. Results show that for the 2PL/GPC model there is moderate negative correlation between item difficulty and discrimination for ELA/literacy. There is less evidence for either positive nor negative correlation between item difficulty and discrimination for mathematics. These tables also show sizable standard deviations for discrimination parameter estimates above 0.20 for all subjects and grade levels, which indicate a substantially wide range of discrimination parameter estimates for the items in the pool. The average discriminations vary somewhat, but not considerably, across item groupings. The constructed-response items were slightly more discriminating on average than selected-response ones. The pattern of item discrimination across different numbers of score categories was inconsistent across subjects. For ELA/literacy, items with two or three score categories had comparable discrimination, while items with four score categories generally had higher average discrimination (which might be due to local item dependence issues for PT items). For mathematics, the fewer the number of score categories, the higher the item discrimination was. ELA/literacy items in claims two and four had slightly higher average discriminations than items in claims one and three for most of the grade levels. Mathematics items did not show a noticeable pattern of differential discrimination across different claims.

Table 40. Summary of 2PL/GPC Slope and Difficulty Estimates and Correlations for ELA/literacy.

Grade	Item Category		No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation	
				Mean	SD	Min.	Max.	Mean	SD	Min.	Max.		
3	Overall		231	0.63	0.23	0.15	1.24	0.32	1.22	-1.87	5.00	-0.29	
	Item Type	SR	76	0.64	0.25	0.16	1.23	-0.44	1.09	-1.87	5.00	-0.64	
		CR	155	0.62	0.22	0.15	1.24	0.69	1.11	-1.80	4.35	-0.12	
	Score Categories	2	134	0.65	0.24	0.16	1.23	0.08	1.27	-1.87	5.00	-0.39	
		3	91	0.56	0.20	0.15	1.09	0.61	1.09	-1.80	3.38	-0.18	
		4	6	1.04	0.16	0.86	1.24	1.39	0.14	1.22	1.60	-0.39	
	Claim Area	1	85	0.63	0.23	0.18	1.12	0.10	1.12	-1.84	3.14	-0.51	
		2	64	0.68	0.26	0.18	1.24	0.33	0.99	-1.25	2.98	0.03	
		3	44	0.60	0.21	0.15	1.06	-0.22	0.96	-1.87	2.23	-0.24	
		4	38	0.57	0.22	0.16	1.06	1.44	1.40	-1.80	5.00	-0.41	
	4	Overall		216	0.57	0.23	0.20	1.40	0.33	1.21	-1.93	4.14	-0.15
		Item Type	SR	111	0.54	0.21	0.20	1.24	-0.32	0.89	-1.93	2.18	-0.59
CR			105	0.61	0.24	0.20	1.40	1.01	1.13	-1.28	4.14	-0.06	
Score Categories		2	148	0.56	0.21	0.20	1.24	0.00	1.12	-1.93	3.54	-0.30	
		3	59	0.53	0.21	0.20	1.26	0.97	1.16	-1.28	4.14	-0.11	
		4	9	1.02	0.25	0.73	1.40	1.48	0.44	1.01	2.00	-0.91	
Claim Area		1	78	0.58	0.22	0.20	1.24	-0.16	1.02	-1.85	2.48	-0.48	
		2	58	0.62	0.25	0.27	1.40	0.42	1.07	-1.93	2.71	0.04	
		3	40	0.49	0.17	0.22	0.83	-0.05	0.91	-1.55	2.54	-0.21	
	4	40	0.57	0.24	0.20	1.26	1.51	1.20	-0.89	4.14	-0.10		
5	Overall		171	0.61	0.20	0.19	1.15	0.34	1.21	-2.14	3.38	-0.16	
	Item Type	SR	77	0.57	0.21	0.19	1.05	-0.46	0.84	-2.14	1.87	-0.53	

Grade	Item Category	No. of Items	a Estimate Summary				b Estimate Summary				a and b Correlation		
			Mean	SD	Min.	Max.	Mean	SD	Min.	Max.			
	Score Categories	CR	94	0.63	0.18	0.20	1.15	1.00	1.06	-1.06	3.38	-0.16	
		2	115	0.59	0.19	0.19	1.05	-0.01	1.15	-2.14	3.38	-0.25	
		3	50	0.61	0.19	0.20	1.12	1.01	1.02	-1.01	2.96	-0.16	
		4	6	0.80	0.26	0.57	1.15	1.51	0.63	0.80	2.14	-0.80	
	Claim Area	1	55	0.56	0.18	0.19	0.92	0.21	1.15	-1.98	2.90	-0.20	
		2	51	0.62	0.20	0.27	1.15	0.39	1.05	-1.74	2.75	-0.07	
		3	32	0.62	0.20	0.28	1.05	-0.51	0.84	-2.14	1.42	-0.59	
		4	33	0.65	0.21	0.20	1.12	1.31	1.18	-1.13	3.38	-0.18	
	6	Overall		197	0.58	0.28	0.17	2.06	0.65	1.48	-1.79	8.05	-0.10
		Item Type	SR	75	0.51	0.20	0.17	1.01	-0.31	0.98	-1.79	2.65	-0.54
			CR	122	0.63	0.31	0.19	2.06	1.23	1.44	-1.26	8.05	-0.16
		Score Categories	2	128	0.58	0.25	0.17	1.34	0.41	1.47	-1.79	5.29	-0.11
3			66	0.58	0.29	0.19	2.06	1.06	1.44	-1.26	8.05	-0.15	
4			3	1.09	0.61	0.59	1.77	1.59	0.24	1.40	1.86	-0.46	
Claim Area		1	77	0.55	0.19	0.19	1.04	0.52	1.27	-1.79	3.54	-0.41	
		2	56	0.61	0.35	0.18	2.06	0.54	1.32	-1.34	4.80	-0.02	
		3	29	0.52	0.17	0.17	0.85	-0.38	0.96	-1.74	2.65	-0.46	
		4	35	0.68	0.34	0.19	1.34	1.93	1.69	-0.29	8.05	-0.23	
7		Overall		190	0.53	0.21	0.11	1.18	0.57	1.34	-2.25	6.61	-0.30
		Item Type	SR	76	0.52	0.24	0.19	1.18	-0.13	1.10	-2.25	3.29	-0.56
	CR		114	0.53	0.19	0.11	1.14	1.04	1.29	-1.76	6.61	-0.18	
		2	115	0.55	0.22	0.19	1.18	0.26	1.38	-2.25	5.81	-0.32	

Grade	Item Category	No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation	
			Mean	SD	Min.	Max.	Mean	SD	Min.	Max.		
	Score Categories	3	70	0.49	0.19	0.11	1.07	1.00	1.15	-1.32	6.61	-0.23
		4	5	0.61	0.08	0.51	0.72	1.68	0.38	1.24	2.10	0.26
	Claim Area	1	70	0.52	0.20	0.12	1.18	0.47	1.41	-2.21	5.81	-0.38
		2	46	0.52	0.16	0.21	0.96	0.40	1.23	-2.25	2.71	-0.14
		3	42	0.47	0.23	0.19	1.06	0.29	1.13	-1.90	3.29	-0.60
		4	32	0.61	0.23	0.11	1.14	1.42	1.31	-0.25	6.61	-0.38
	8	Overall		199	0.56	0.27	0.08	1.58	0.53	1.21	-2.87	6.17
Item Type		SR	77	0.50	0.20	0.08	1.02	-0.11	0.98	-2.01	2.53	-0.50
		CR	122	0.59	0.30	0.13	1.58	0.93	1.17	-2.87	6.17	-0.11
Score Categories		2	119	0.56	0.24	0.08	1.26	0.23	1.17	-2.01	6.17	-0.17
		3	74	0.49	0.24	0.13	1.25	0.93	1.17	-2.87	4.47	-0.19
		4	6	1.24	0.35	0.69	1.58	1.49	0.21	1.30	1.83	-0.26
Claim Area		1	75	0.49	0.17	0.13	0.90	0.38	1.40	-2.01	6.17	-0.36
		2	50	0.64	0.35	0.18	1.58	0.36	1.02	-2.87	2.18	0.19
		3	40	0.47	0.21	0.17	1.02	0.44	1.16	-1.78	2.95	-0.61
		4	34	0.69	0.30	0.08	1.26	1.21	0.82	-0.53	3.30	-0.29
9	Overall		119	0.60	0.24	0.20	1.20	0.64	1.33	-2.24	6.04	0.01
	Item Type	SR	36	0.54	0.20	0.22	0.99	-0.43	0.78	-1.60	1.21	-0.51
		CR	83	0.63	0.25	0.20	1.20	1.10	1.25	-2.24	6.04	-0.03
	Score Categories	2	64	0.58	0.23	0.20	1.08	0.38	1.36	-1.60	3.54	0.01
		3	51	0.60	0.26	0.21	1.20	0.89	1.28	-2.24	6.04	-0.06
		4	4	0.87	0.15	0.73	1.06	1.45	0.22	1.25	1.74	-0.44

Grade	Item Category		No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation
				Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	
	Claim Area	1	56	0.58	0.27	0.20	1.20	0.54	1.46	-2.24	6.04	-0.13
		2	25	0.65	0.20	0.29	1.00	0.46	1.11	-1.60	2.61	0.07
		3	14	0.49	0.19	0.28	0.99	-0.19	1.08	-1.23	2.12	-0.46
		4	24	0.67	0.23	0.22	1.10	1.55	0.86	-0.30	3.30	0.24
10	Overall		129	0.60	0.25	0.19	1.33	0.75	1.26	-1.78	4.70	-0.18
	Item Type	SR	46	0.56	0.25	0.22	1.11	-0.10	0.92	-1.78	2.78	-0.55
		CR	83	0.63	0.24	0.19	1.33	1.23	1.17	-0.76	4.70	-0.16
	Score Categories	2	73	0.61	0.24	0.22	1.12	0.53	1.40	-1.78	4.70	-0.21
		3	53	0.57	0.24	0.19	1.32	1.02	1.01	-0.76	3.25	-0.17
		4	3	1.00	0.30	0.73	1.33	1.53	0.22	1.28	1.71	-0.99
	Claim Area	1	59	0.55	0.20	0.19	1.05	0.74	1.40	-1.78	4.70	-0.28
		2	30	0.73	0.28	0.22	1.33	0.90	1.18	-1.34	3.92	-0.15
		3	20	0.52	0.25	0.21	1.11	0.00	0.84	-1.21	1.91	-0.72
		4	20	0.65	0.26	0.23	1.30	1.32	0.96	-0.16	2.78	-0.14
11	Overall		259	0.54	0.22	0.18	1.32	1.01	1.20	-1.97	5.09	-0.15
	Item Type	SR	91	0.49	0.17	0.19	0.91	0.24	0.89	-1.97	2.85	-0.55
		CR	168	0.57	0.23	0.18	1.32	1.43	1.14	-1.29	5.09	-0.18
	Score Categories	2	142	0.54	0.20	0.19	1.21	0.75	1.26	-1.97	5.09	-0.09
		3	110	0.52	0.22	0.18	1.18	1.34	1.07	-0.68	4.71	-0.27
		4	7	0.89	0.28	0.69	1.32	1.36	0.10	1.26	1.50	-0.27
	Claim Area	1	95	0.47	0.20	0.19	1.19	1.20	1.26	-1.97	4.83	-0.22
		2	54	0.65	0.24	0.26	1.32	0.65	1.00	-1.25	2.92	0.04

Grade	Item Category	No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation	
			Mean	SD	Min.	Max.	Mean	SD	Min.	Max.		
		3	65	0.48	0.16	0.18	0.86	0.72	1.17	-1.29	4.71	-0.46
		4	45	0.65	0.20	0.26	1.21	1.49	1.13	-0.52	5.09	-0.02

Table 41. Summary of 2PL/GPC Slope and Difficulty Estimates and Correlations for Mathematics.

Grade	Item Category		No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation	
				Mean	SD	Min	Max	Mean	SD	Min	Max		
3	Overall		207	0.69	0.21	0.21	1.31	0.31	1.43	-4.06	4.42	0.01	
	Item Type	SR	34	0.65	0.23	0.21	1.21	-0.81	1.25	-4.06	1.84	-0.33	
		CR	173	0.69	0.21	0.21	1.31	0.53	1.36	-3.16	4.42	0.04	
	Score Categories	2	126	0.75	0.21	0.21	1.31	0.18	1.50	-4.06	3.63	0.05	
		3	66	0.61	0.16	0.32	0.98	0.45	1.28	-2.77	4.42	0.08	
		4	15	0.50	0.18	0.21	0.77	0.77	1.36	-1.68	3.25	0.28	
	Claim Area	1	154	0.71	0.21	0.21	1.31	0.20	1.43	-4.06	3.63	0.04	
		2	28	0.65	0.18	0.37	1.22	0.50	1.15	-1.26	3.53	0.02	
		3	17	0.60	0.19	0.21	0.86	1.00	1.79	-2.77	4.42	-0.16	
		4	8	0.60	0.21	0.28	0.95	0.37	1.08	-1.68	1.55	0.71	
	4	Overall		209	0.72	0.25	0.19	1.32	0.72	1.20	-3.42	3.97	0.01
		Item Type	SR	31	0.63	0.25	0.19	1.10	-0.09	1.36	-1.91	3.86	-0.58
CR			178	0.73	0.25	0.27	1.32	0.86	1.12	-3.42	3.97	0.08	
Score Categories		2	141	0.78	0.25	0.19	1.32	0.70	1.28	-3.42	3.97	-0.02	
		3	55	0.59	0.17	0.28	1.09	0.64	1.03	-1.66	2.45	0.30	
		4	13	0.50	0.14	0.28	0.77	1.32	0.82	0.05	2.46	0.10	
Claim Area		1	158	0.72	0.24	0.24	1.32	0.54	1.23	-3.42	3.58	0.09	
		2	30	0.70	0.28	0.19	1.22	1.23	0.91	-0.10	3.86	-0.30	
		3	14	0.72	0.31	0.28	1.26	1.36	0.98	0.01	3.97	-0.26	
		4	7	0.70	0.20	0.50	1.08	1.41	0.62	0.40	2.44	0.30	
5	Overall		204	0.71	0.26	0.23	1.38	0.55	1.10	-3.34	4.17	0.17	
	Item Type	SR	39	0.62	0.21	0.23	1.13	-0.11	0.73	-1.83	1.72	0.00	

Grade	Item Category		No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation	
				Mean	SD	Min	Max	Mean	SD	Min	Max		
		CR	165	0.73	0.27	0.27	1.38	0.70	1.12	-3.34	4.17	0.14	
	Score Categories	2	144	0.76	0.27	0.23	1.38	0.47	1.15	-3.34	3.43	0.25	
		3	53	0.60	0.20	0.27	1.11	0.71	0.99	-1.29	4.17	0.04	
		4	7	0.47	0.09	0.30	0.58	0.82	0.70	-0.26	1.68	0.48	
	Claim Area	1	156	0.71	0.25	0.23	1.31	0.43	1.12	-3.34	4.17	0.19	
		2	26	0.76	0.28	0.38	1.38	1.04	0.99	-0.70	3.43	-0.01	
		3	15	0.56	0.24	0.30	1.09	0.69	0.85	-1.29	1.72	-0.15	
		4	7	0.77	0.33	0.34	1.13	1.00	1.03	-0.26	2.33	0.68	
	6	Overall		189	0.70	0.27	0.19	1.58	0.95	1.19	-1.77	4.09	0.01
		Item Type	SR	41	0.55	0.21	0.19	1.10	0.21	1.17	-1.77	2.98	-0.55
			CR	148	0.74	0.27	0.20	1.58	1.16	1.12	-1.54	4.09	0.00
		Score Categories	2	133	0.75	0.28	0.19	1.58	0.94	1.28	-1.77	4.09	0.05
3			49	0.61	0.18	0.20	0.99	1.02	0.99	-0.78	3.69	-0.23	
4			7	0.43	0.12	0.32	0.64	0.74	0.80	-0.36	2.07	0.03	
Claim Area		1	149	0.68	0.27	0.19	1.58	0.84	1.21	-1.77	4.09	-0.08	
		2	22	0.74	0.21	0.41	1.17	1.03	1.13	-1.19	3.05	0.48	
		3	11	0.63	0.26	0.30	1.10	1.84	0.73	0.63	3.10	-0.05	
		4	7	0.97	0.32	0.63	1.50	1.84	0.75	0.67	2.97	0.33	
7		Overall		190	0.66	0.26	0.15	1.43	1.38	1.19	-1.81	6.38	-0.11
		Item Type	SR	31	0.46	0.15	0.23	0.91	0.82	1.12	-1.81	3.84	-0.68
	CR		159	0.70	0.26	0.15	1.43	1.49	1.18	-1.02	6.38	-0.16	
		2	101	0.73	0.27	0.23	1.43	1.38	1.03	-1.81	4.20	0.11	

Grade	Item Category	No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation		
			Mean	SD	Min	Max	Mean	SD	Min	Max			
	Score Categories	3	74	0.60	0.22	0.15	1.15	1.40	1.40	-1.02	6.38	-0.45	
		4	15	0.50	0.21	0.21	0.96	1.25	1.11	-0.62	3.89	0.07	
	Claim Area	1	148	0.67	0.26	0.15	1.43	1.41	1.17	-1.81	6.38	-0.09	
		2	20	0.74	0.22	0.27	1.17	0.90	0.83	-0.87	2.59	0.22	
		3	17	0.54	0.26	0.21	1.06	1.65	1.65	-0.92	5.46	-0.33	
		4	5	0.70	0.16	0.51	0.96	1.44	1.02	0.29	2.56	0.14	
	8	Overall		191	0.65	0.27	0.13	1.47	1.25	1.17	-1.49	5.12	-0.08
Item Type		SR	34	0.48	0.17	0.20	0.76	0.79	1.09	-0.99	4.54	-0.60	
		CR	157	0.69	0.28	0.13	1.47	1.35	1.16	-1.49	5.12	-0.09	
Score Categories		2	121	0.70	0.30	0.18	1.47	1.35	1.22	-1.20	5.12	-0.13	
		3	62	0.57	0.20	0.13	1.07	1.02	1.06	-1.49	4.95	-0.02	
		4	8	0.50	0.16	0.28	0.82	1.40	0.96	0.12	3.04	-0.62	
Claim Area		1	149	0.63	0.27	0.13	1.45	1.20	1.17	-1.49	5.12	-0.11	
		2	26	0.74	0.31	0.34	1.47	1.58	1.26	-0.97	5.12	-0.05	
		3	12	0.65	0.16	0.48	0.88	1.04	1.00	-1.01	2.49	-0.15	
		4	4	0.72	0.25	0.45	1.04	1.50	0.42	0.97	1.85	-0.18	
9		Overall		103	0.60	0.27	0.15	1.42	1.92	1.27	-0.62	7.34	0.00
		Item Type	SR	14	0.46	0.20	0.21	0.77	0.99	1.04	-0.29	3.76	-0.62
	CR		89	0.62	0.28	0.15	1.42	2.07	1.25	-0.62	7.34	-0.01	
	Score Categories	2	63	0.68	0.28	0.21	1.42	1.89	1.31	-0.62	7.34	0.09	
		3	34	0.50	0.21	0.15	1.02	1.94	1.18	0.29	6.10	-0.14	
		4	6	0.33	0.09	0.23	0.44	2.13	1.54	-0.44	3.80	-0.30	
	Claim Area	1	84	0.62	0.28	0.15	1.42	1.93	1.31	-0.62	7.34	-0.02	

Grade	Item Category	No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation	
			Mean	SD	Min	Max	Mean	SD	Min	Max		
		2	11	0.47	0.22	0.20	0.77	1.77	1.36	-0.44	4.21	0.18
		3	6	0.51	0.18	0.24	0.69	2.14	0.84	1.02	3.24	-0.11
		4	2	0.52	0.08	0.47	0.58	1.71	0.30	1.50	1.93	-1.00
10	Overall		122	0.67	0.36	0.17	1.76	1.32	1.10	-1.15	5.49	0.12
	Item Type	SR	19	0.48	0.22	0.18	1.12	0.88	1.48	-0.71	5.49	-0.35
		CR	103	0.71	0.37	0.17	1.76	1.40	1.00	-1.15	3.84	0.15
	Score Categories	2	68	0.81	0.40	0.18	1.76	1.40	1.15	-0.71	5.49	0.06
		3	42	0.53	0.19	0.17	0.91	1.22	1.00	-1.15	3.67	0.19
		4	12	0.37	0.17	0.17	0.75	1.18	1.16	-0.36	3.23	0.22
	Claim Area	1	94	0.69	0.38	0.17	1.76	1.13	1.08	-1.15	5.49	0.21
		2	13	0.67	0.22	0.26	1.10	1.80	0.69	-0.02	2.54	0.03
		3	10	0.50	0.32	0.17	1.33	1.99	1.06	0.37	3.84	-0.19
		4	5	0.59	0.29	0.36	1.09	2.35	1.05	1.27	3.67	0.03
11	Overall		263	0.84	0.39	0.21	2.20	2.18	1.29	-1.11	5.48	-0.04
	Item Type	SR	32	0.45	0.21	0.21	1.22	1.05	1.27	-1.06	3.63	-0.43
		CR	231	0.90	0.38	0.22	2.20	2.33	1.21	-1.11	5.48	-0.17
	Score Categories	2	213	0.90	0.40	0.21	2.20	2.28	1.29	-1.06	5.48	-0.10
		3	41	0.62	0.23	0.22	1.19	1.78	1.19	-0.87	4.45	-0.16
		4	9	0.59	0.24	0.35	1.01	1.45	1.15	-1.11	2.51	0.20
	Claim Area	1	204	0.84	0.38	0.21	2.18	2.09	1.31	-1.11	5.48	-0.06
		2	31	0.85	0.49	0.24	2.20	2.81	1.20	-0.53	4.90	-0.04
		3	20	0.83	0.33	0.33	1.68	1.74	0.98	0.26	4.10	-0.03

Grade	Item Category	No. of Items	<i>a</i> Estimate Summary				<i>b</i> Estimate Summary				<i>a</i> and <i>b</i> Correlation	
			Mean	SD	Min	Max	Mean	SD	Min	Max		
		4	8	0.92	0.54	0.34	2.07	2.96	0.79	1.61	4.18	0.25

Evaluation of Ability Estimates

It is worthwhile to determine how ability estimates and scales vary among the three model combinations. The expectation is that the correlations of ability estimates will be very high across models for a given student since the same item responses are used for all three ability estimates. The differences are determined by the respective weighting of the item responses and how the ability scales differ in terms of being “stretched” or “compressed” in various parts of the ability scale.⁸ For this evaluation, MLE scoring table estimates were used for ability. Tables 42 and 43 summarize means and standard deviations of theta estimates and their correlations across different model combinations for ELA/literacy and mathematics, respectively. Figures 73 and 74 present scatter plots of theta estimates for different model choices for ELA/literacy and mathematics, respectively. Results show that the ability estimates across all three models are highly correlated. The scatter plots show that 2PL/GPC produced ability estimates that were most similar to the 3PL/GPC in the middle of the ability scale. Despite the difference between item-parameter estimates produced by the 1PL/PC and the 3PL/GPC, the ability scale produced by the 1PL/PC is very similar to that produced by 3PL/GPC, and the two ability scales exhibit a linear relationship.

Table 42. ELA/literacy Correlations of Ability Estimates across Different Model Combinations.

Grade	Model	Theta Summary		Theta Correlations		
		Mean	SD	1PL/PC	2PL/GPC	3PL/GPC
3	1PL/PC	-0.02	1.10	1.00	0.99	0.98
	2PL/GPC	-0.01	1.10		1.00	0.99
	3PL/GPC	-0.01	1.10			1.00
4	1PL/PC	-0.01	1.13	1.00	0.98	0.97
	2PL/GPC	0.01	1.14		1.00	0.99
	3PL/GPC	0.01	1.14			1.00
5	1PL/PC	-0.01	1.14	1.00	0.98	0.97
	2PL/GPC	0.00	1.16		1.00	0.98
	3PL/GPC	0.00	1.18			1.00
6	1PL/PC	-0.01	1.16	1.00	0.98	0.97

⁸The three models produce different scales when applied to selected-response data where it is possible for very low ability students to correctly identify the keyed answer (Yen, 1981).

Grade	Model	Theta Summary		Theta Correlations		
		Mean	<i>SD</i>	1PL/PC	2PL/GPC	3PL/GPC
	2PL/GPC	0.00	1.18		1.00	0.99
	3PL/GPC	-0.01	1.19			1.00
7	1PL/PC	-0.01	1.16	1.00	0.97	0.95
	2PL/GPC	0.01	1.19		1.00	0.98
	3PL/GPC	-0.01	1.19			1.00
8	1PL/PC	-0.01	1.17	1.00	0.98	0.97
	2PL/GPC	0.00	1.19		1.00	0.99
	3PL/GPC	0.00	1.20			1.00
9	1PL/PC	-0.01	1.17	1.00	0.97	0.96
	2PL/GPC	0.00	1.20		1.00	0.99
	3PL/GPC	-0.01	1.21			1.00
10	1PL/PC	-0.02	1.15	1.00	0.98	0.97
	2PL/GPC	0.00	1.15		1.00	0.99
	3PL/GPC	0.00	1.15			1.00
11	1PL/PC	-0.02	1.12	1.00	0.98	0.97
	2PL/GPC	-0.03	1.14		1.00	0.98
	3PL/GPC	-0.04	1.15			1.00

Table 43. Mathematics Correlations of Ability Estimates across Different Model Combinations.

Grade	Model	Theta Summary		Theta Correlations		
		Mean	SD	1PL/PC	2PL/GPC	3PL/GPC
3	1PL/PC	-0.01	1.10	1.00	0.99	0.98
	2PL/GPC	-0.03	1.11		1.00	1.00
	3PL/GPC	-0.03	1.11			1.00
4	1PL/PC	-0.01	1.07	1.00	0.99	0.98
	2PL/GPC	-0.04	1.09		1.00	0.99
	3PL/GPC	-0.06	1.06			1.00
5	1PL/PC	-0.02	1.09	1.00	0.99	0.97
	2PL/GPC	-0.04	1.11		1.00	0.99
	3PL/GPC	-0.05	1.11			1.00
6	1PL/PC	0.01	1.09	1.00	0.98	0.97
	2PL/GPC	-0.01	1.11		1.00	0.99
	3PL/GPC	0.00	1.09			1.00
7	1PL/PC	0.00	1.09	1.00	0.98	0.96
	2PL/GPC	-0.02	1.11		1.00	0.98
	3PL/GPC	-0.05	1.06			1.00
8	1PL/PC	0.01	1.09	1.00	0.97	0.96
	2PL/GPC	-0.01	1.12		1.00	0.99
	3PL/GPC	-0.01	1.11			1.00
9	1PL/PC	0.00	1.14	1.00	0.95	0.92
	2PL/GPC	-0.07	1.16		1.00	0.96
	3PL/GPC	-0.15	1.13			1.00
10	1PL/PC	-0.02	1.14	1.00	0.97	0.93
	2PL/GPC	-0.09	1.13		1.00	0.97
	3PL/GPC	-0.27	1.03			1.00
11	1PL/PC	0.06	1.01	1.00	0.95	0.92

Grade	Model	Theta Summary		Theta Correlations		
		Mean	<i>SD</i>	1PL/PC	2PL/GPC	3PL/GPC
	2PL/GPC	-0.08	1.01		1.00	0.98
	3PL/GPC	-0.08	0.95			1.00

IRT Model Recommendations

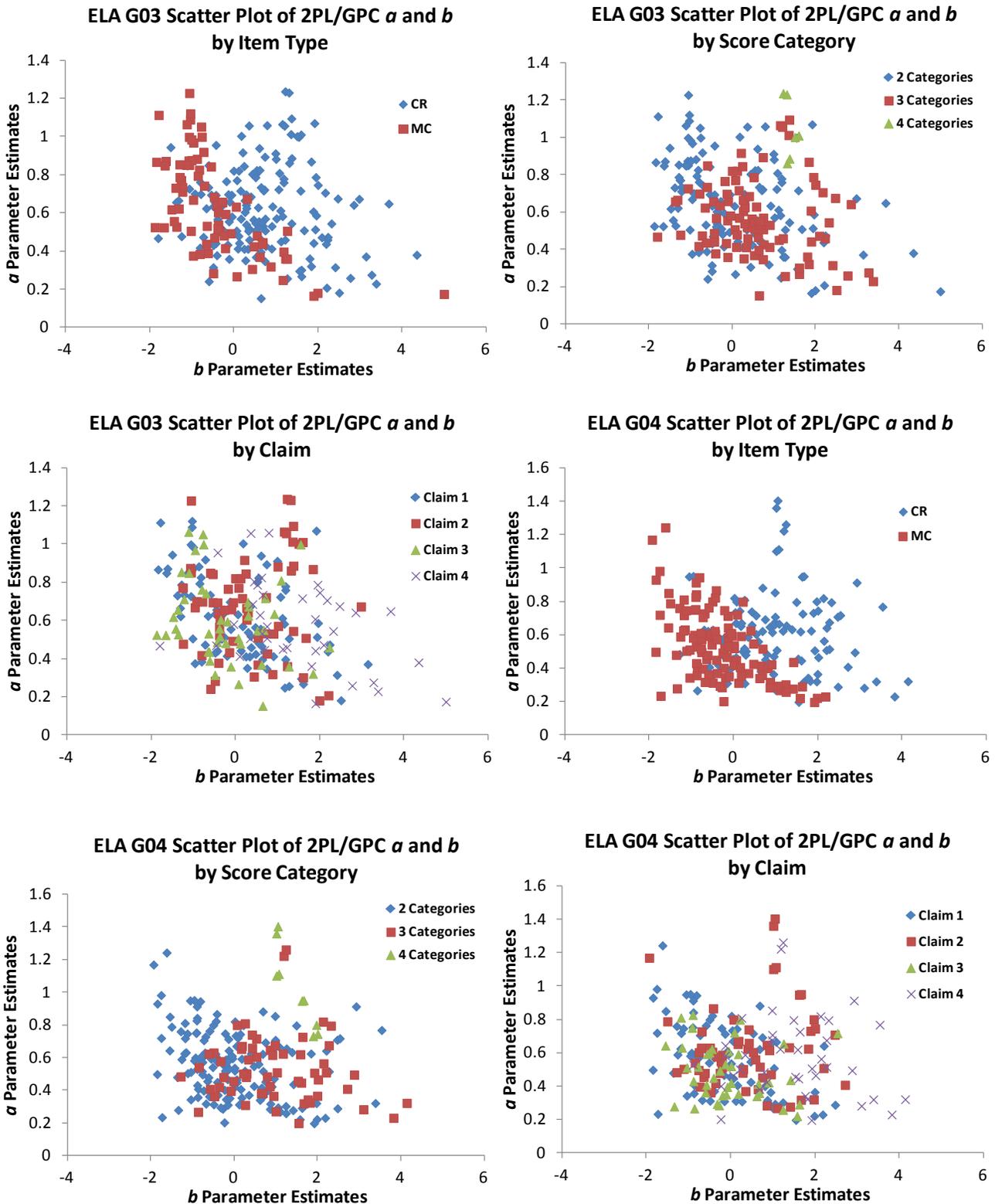
Based on the model comparison analysis results for the Pilot Test, the 2PL/GPC model could be adopted as the IRT model combination for calibrating Smarter Balanced items and establishing a vertical scale. The 2PL/GPC model provides flexibility for estimating a range of item discriminations without the complications of implementing a 3PL/GPC model. Recommendations based on the model comparison analysis should be evaluated with caution given the preliminary nature of the Pilot data. There were changes in item formats from Pilot to Field Test to operational administration, and adjustments were made to the test blueprints. In addition, performance tasks for mathematics were not available for analysis. There was no information concerning the impact of the three models for vertical scaling and growth depictions.

These results were presented to the Technical Advisory Committee Meeting held in Minneapolis, MN, in May 2014. The Smarter Balanced Executive Committee representatives accepted, on a majority-rule basis, that 2PL/GPCM was the preferred IRT model combination for the Field Test analysis. The following rationale and limitations were discussed:

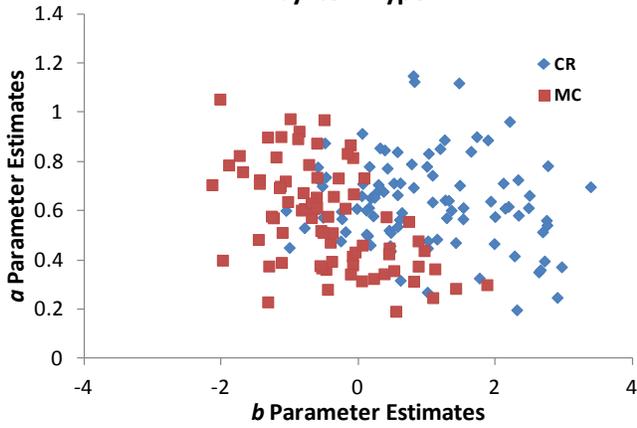
- There was a practical constraint that the scale will need to be established in the Field Test under very short timelines to support other activities such as standard setting. As a result, there will not be sufficient time to analyze the Field Test data under different IRT models. Therefore, it was necessary to determine an IRT model combination prior to the Field Test without the benefit of further examination.
- Although the Pilot data suggested that 2PL/GPC showed significant improvement in data-model fit, the Pilot Test data imposed limits on the ability to generalize the results.
- The impact of misfit under a CAT administration is minimized to some extent since items are targeted at student ability over the middle of the item characteristic curve.
- In the Field Test, 1PL/PC might be advantageous because of stability of scales under the Rasch model, particularly when a program is in the midst of significant change.
- If the conditions for additive conjoint measurement are met for Rasch, then it is assumed that interval level measurement will result. Interval level measurement is a desirable and necessary property for vertical scales.

Due in part to these considerations, the consensus was that the Smarter Balanced Field Test be scaled with 2PL/2PPC IRT model combination.

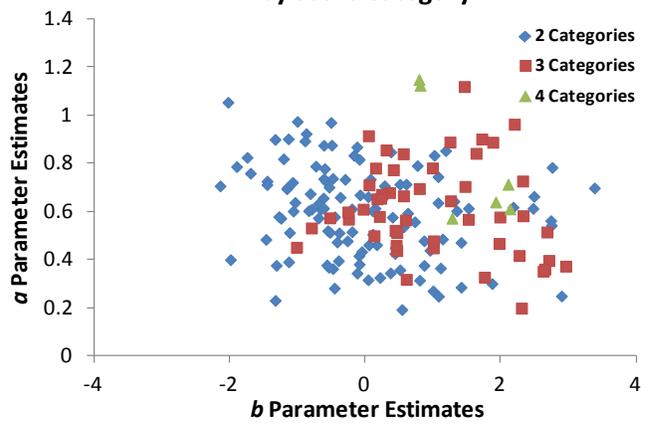
Figure 71. Scatter Plot of ELA/literacy 2PL/GPC Slope and Difficulty Estimates by Item Type, Score Category and Claim



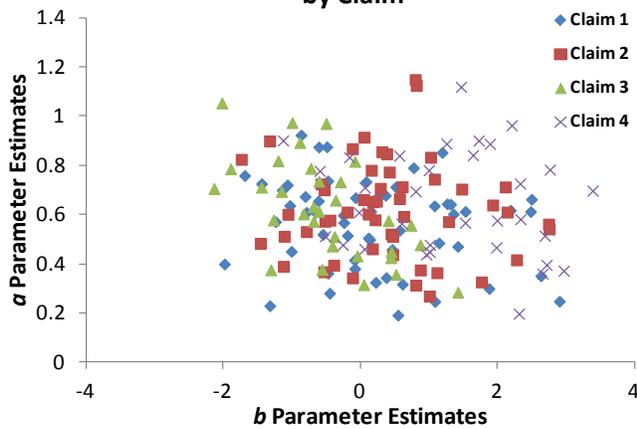
ELA G05 Scatter Plot of 2PL/GPC a and b by Item Type



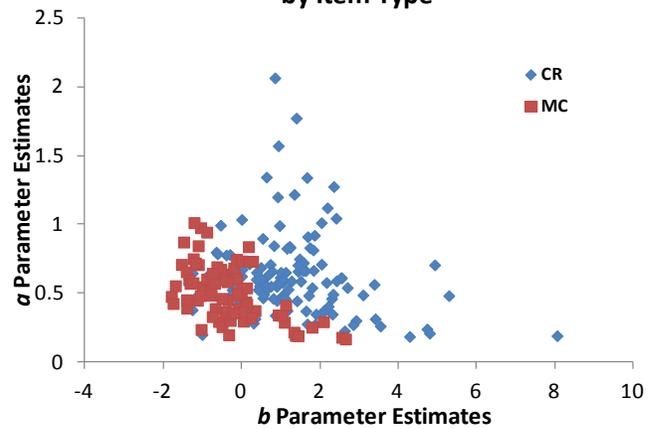
ELA G05 Scatter Plot of 2PL/GPC a and b by Score Category



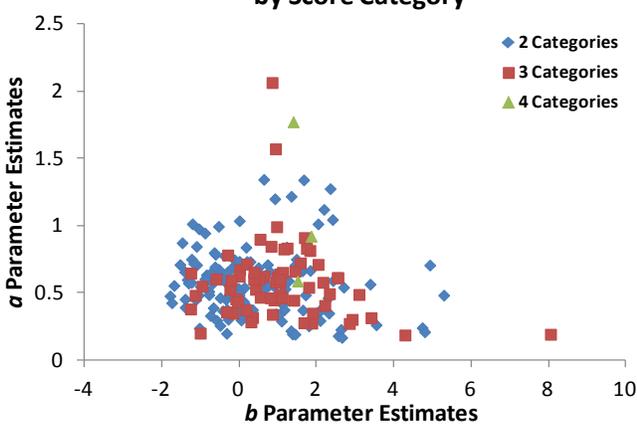
ELA G05 Scatter Plot of 2PL/GPC a and b by Claim



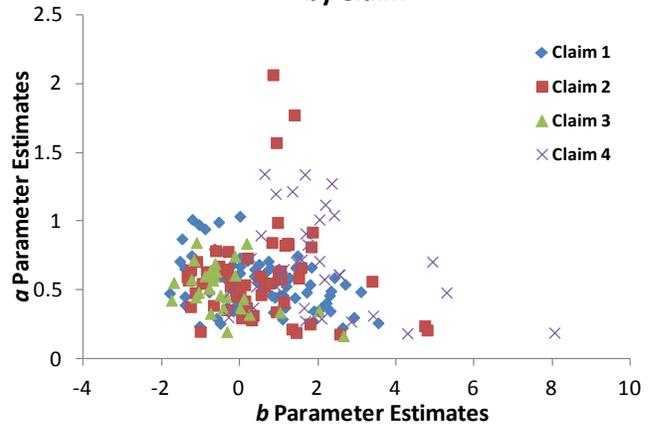
ELA G06 Scatter Plot of 2PL/GPC a and b by Item Type



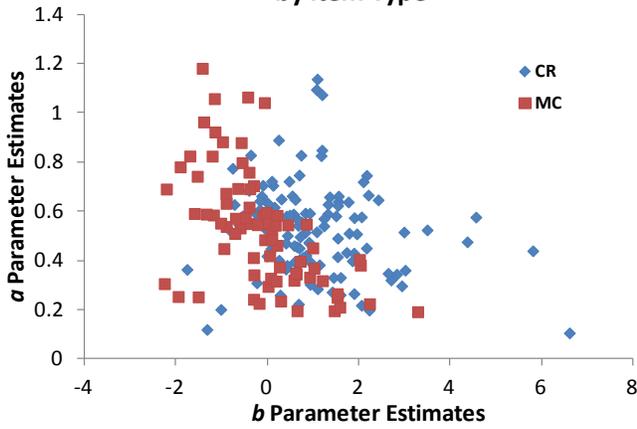
ELA G06 Scatter Plot of 2PL/GPC a and b by Score Category



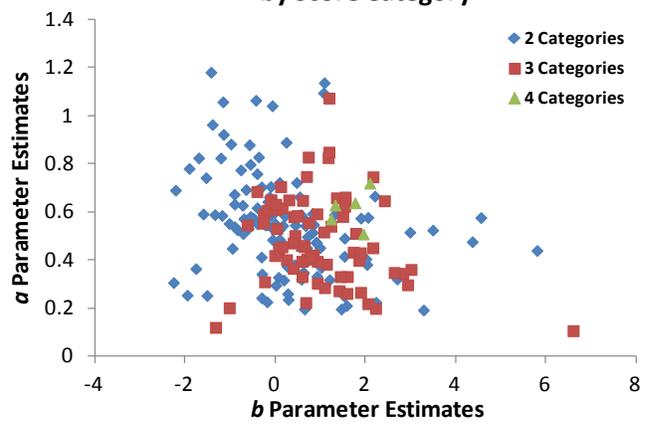
ELA G06 Scatter Plot of 2PL/GPC a and b by Claim



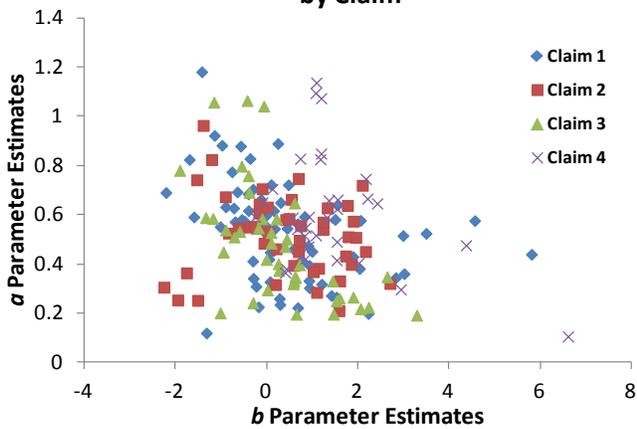
ELA G07 Scatter Plot of 2PL/GPC a and b by Item Type



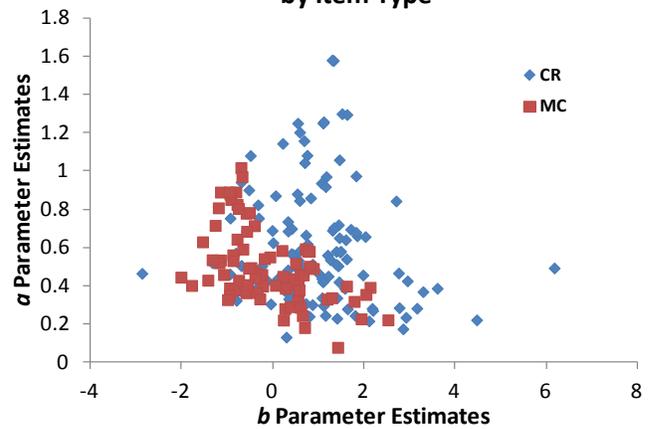
ELA G07 Scatter Plot of 2PL/GPC a and b by Score Category



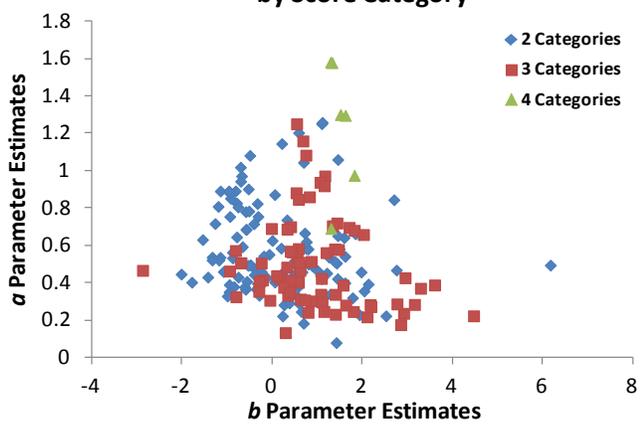
ELA G07 Scatter Plot of 2PL/GPC a and b by Claim



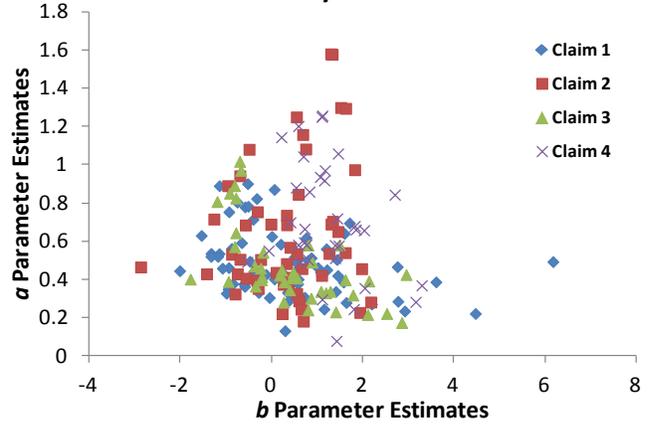
ELA G08 Scatter Plot of 2PL/GPC a and b by Item Type



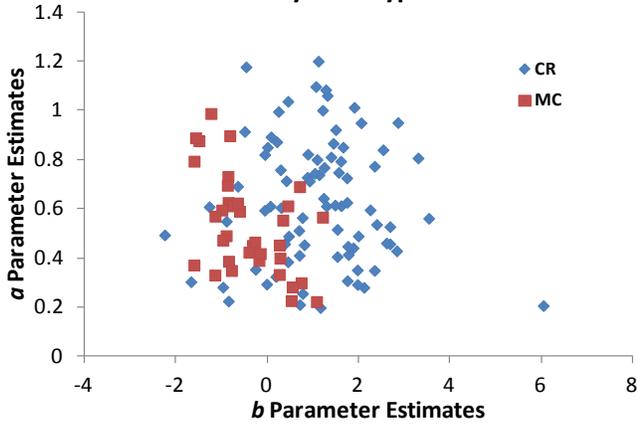
ELA G08 Scatter Plot of 2PL/GPC a and b by Score Category



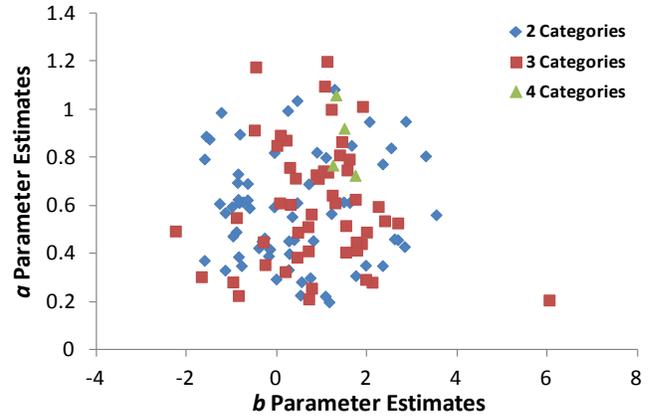
ELA G08 Scatter Plot of 2PL/GPC a and b by Claim



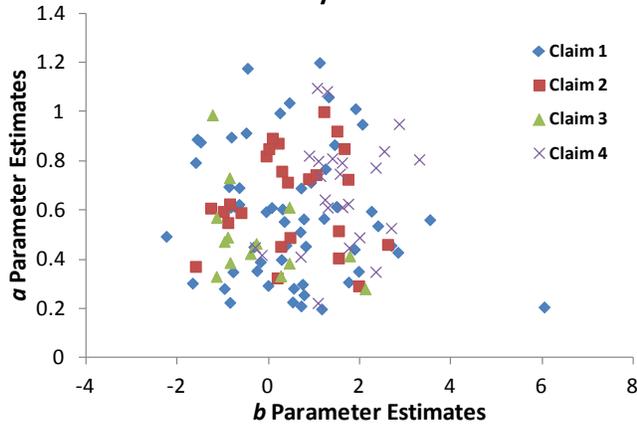
ELA G09 Scatter Plot of 2PL/GPC a and b by Item Type



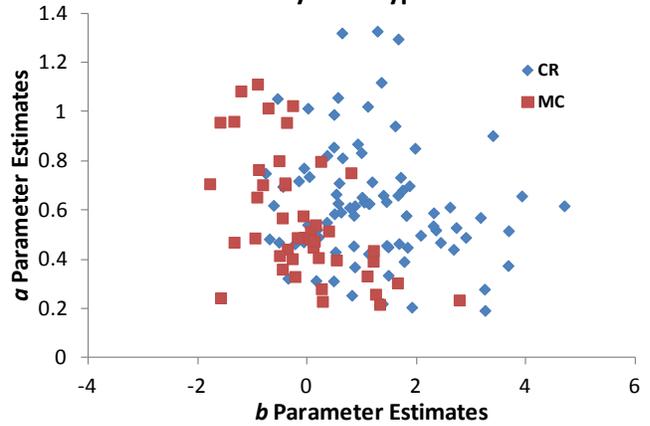
ELA G09 Scatter Plot of 2PL/GPC a and b by Score Category



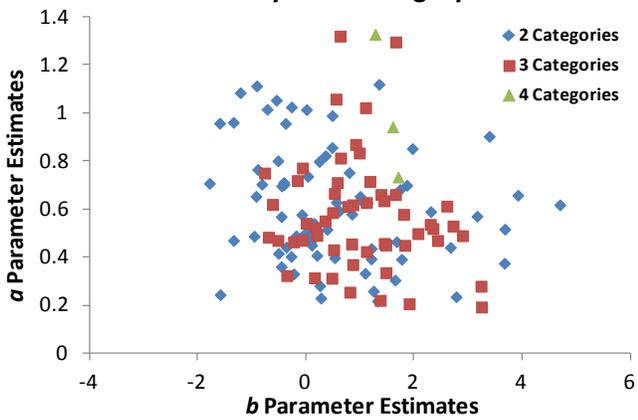
ELA G09 Scatter Plot of 2PL/GPC a and b by Claim



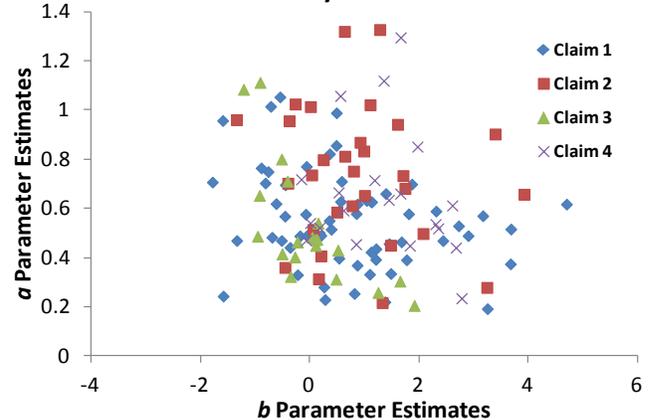
ELA G10 Scatter Plot of 2PL/GPC a and b by Item Type



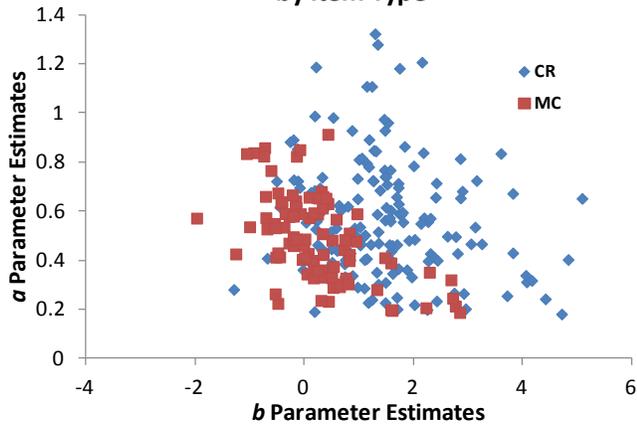
ELA G10 Scatter Plot of 2PL/GPC a and b by Score Category



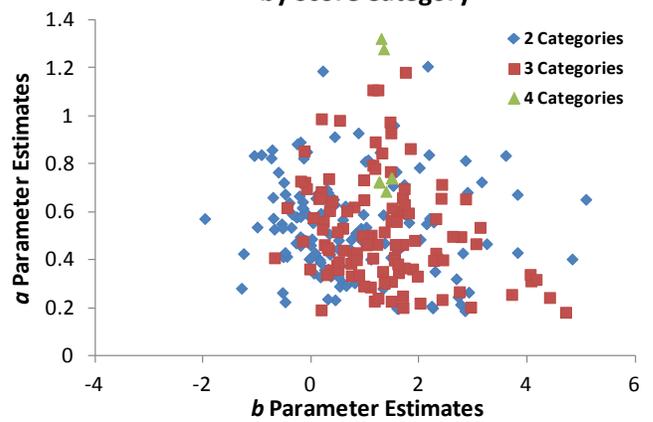
ELA G10 Scatter Plot of 2PL/GPC a and b by Claim



ELA G11 Scatter Plot of 2PL/GPC a and b
by Item Type



ELA G11 Scatter Plot of 2PL/GPC a and b
by Score Category



ELA G11 Scatter Plot of 2PL/GPC a and b
by Claim

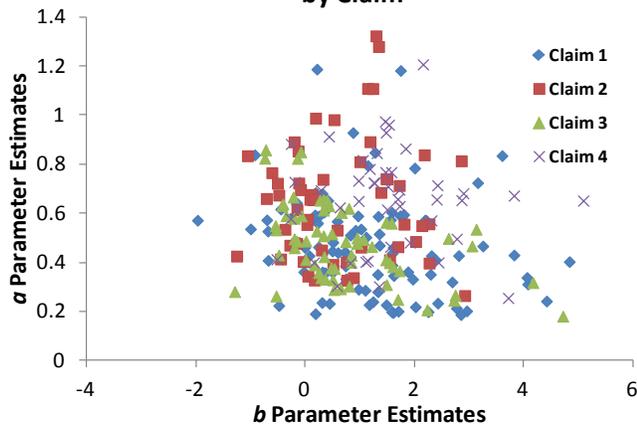
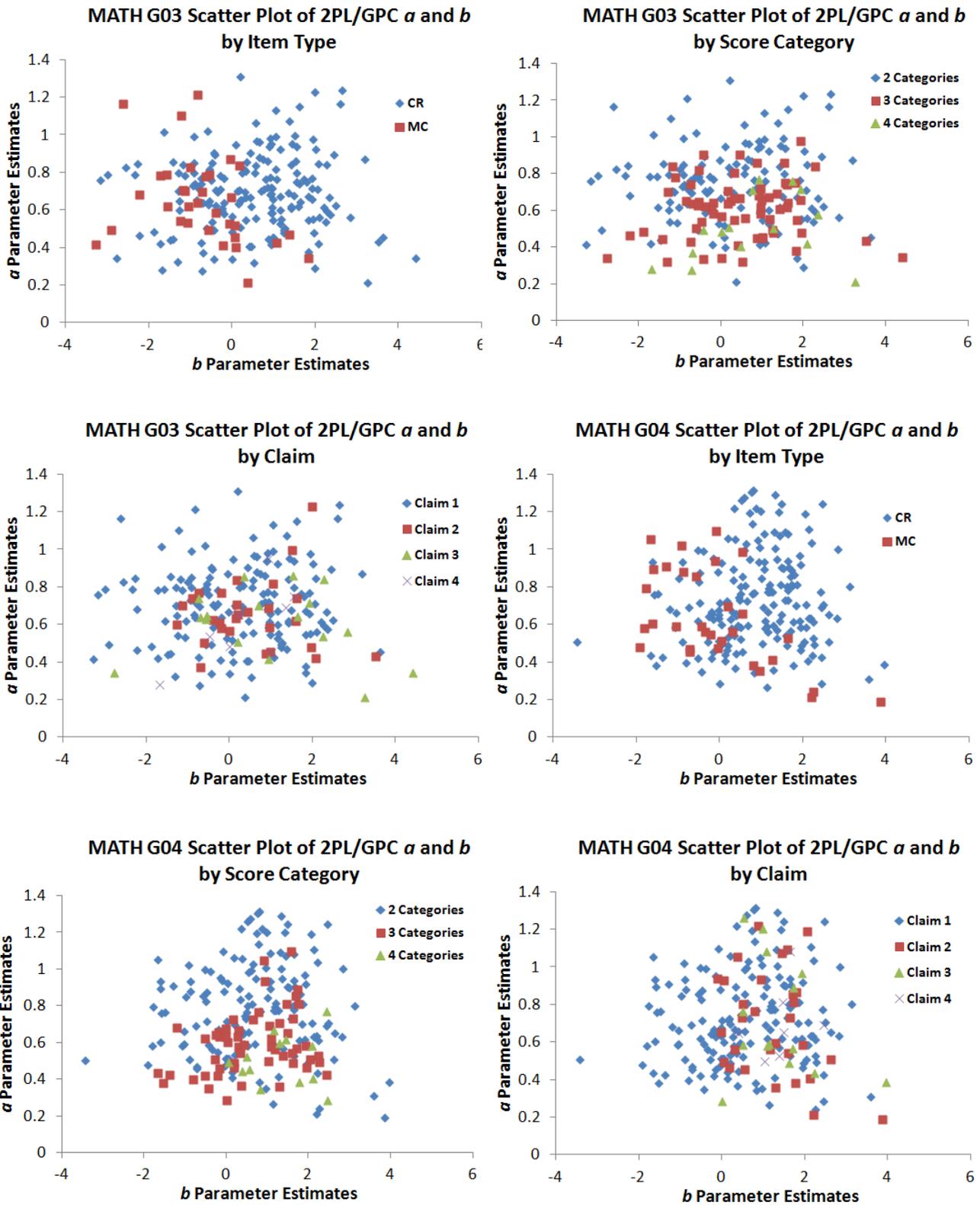
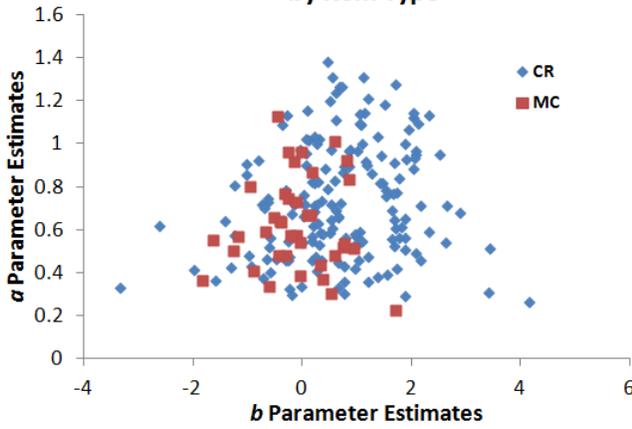


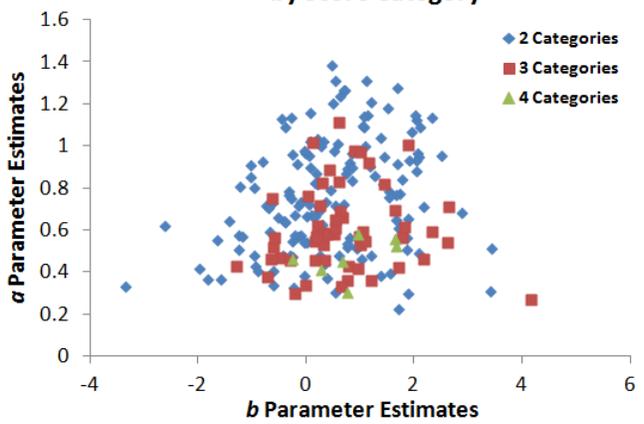
Figure 72. Scatter Plot of Mathematics 2PL/GPC Slope and Difficulty Estimates by Item Type, Score Category, and Claim



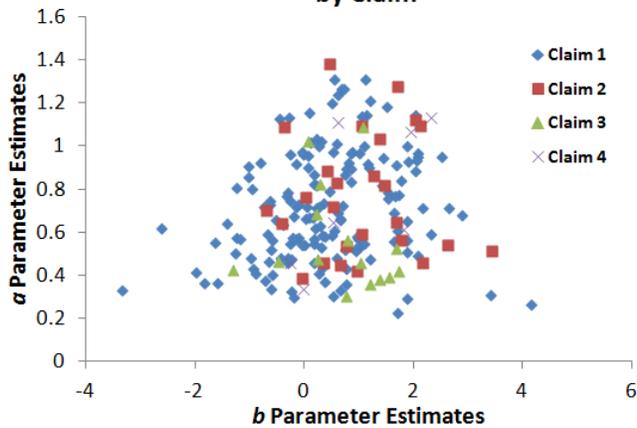
MATH G05 Scatter Plot of 2PL/GPC a and b by Item Type



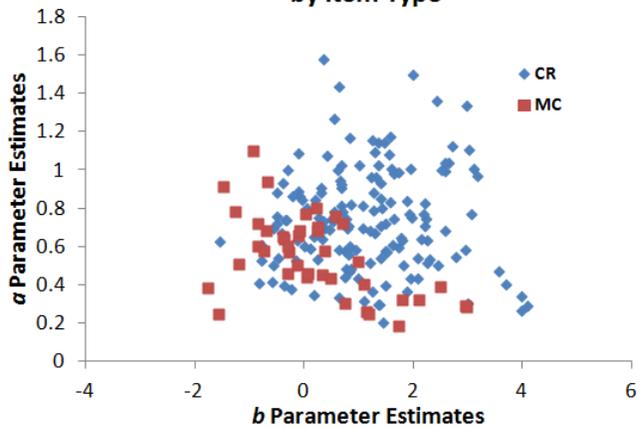
MATH G05 Scatter Plot of 2PL/GPC a and b by Score Category



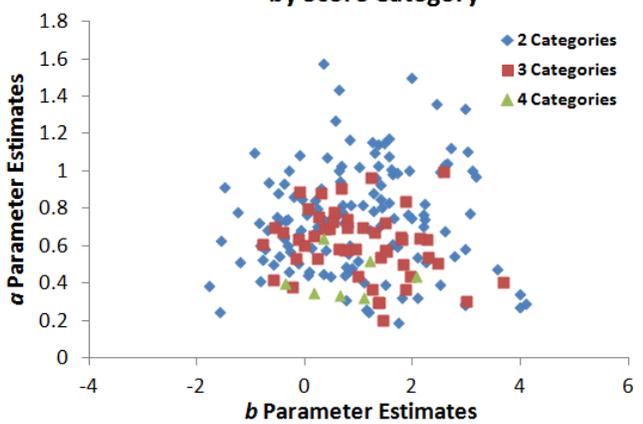
MATH G05 Scatter Plot of 2PL/GPC a and b by Claim



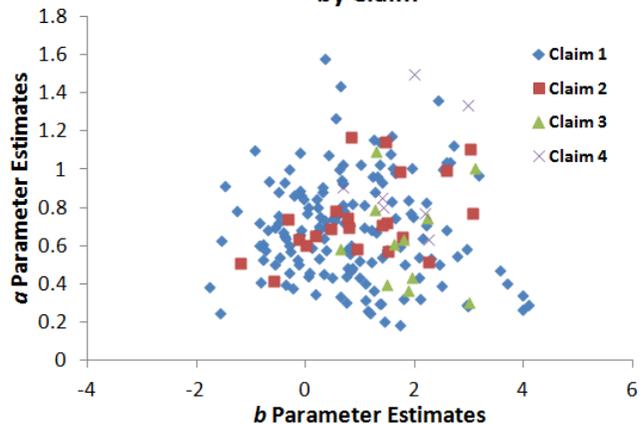
MATH G06 Scatter Plot of 2PL/GPC a and b by Item Type



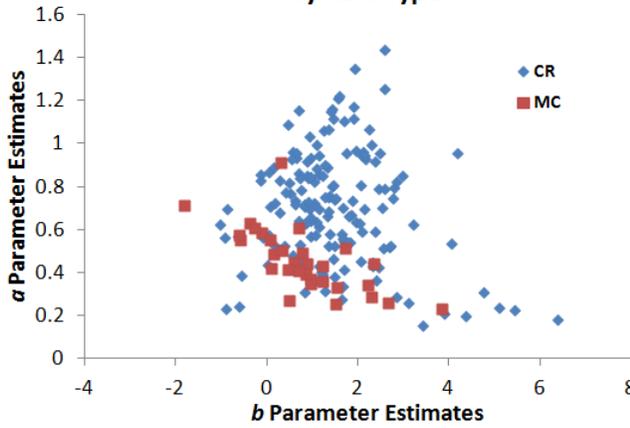
MATH G06 Scatter Plot of 2PL/GPC a and b by Score Category



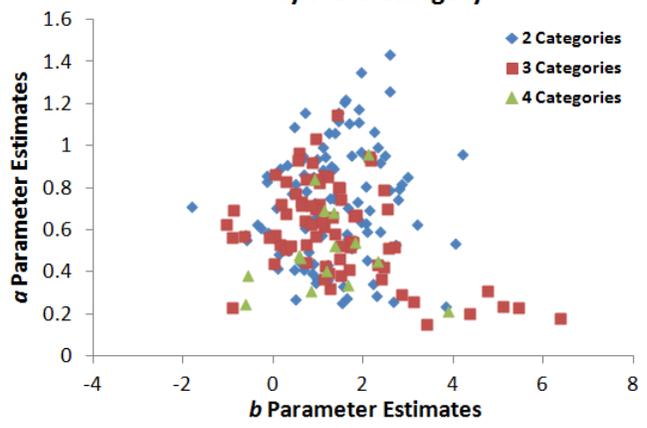
MATH G06 Scatter Plot of 2PL/GPC a and b by Claim



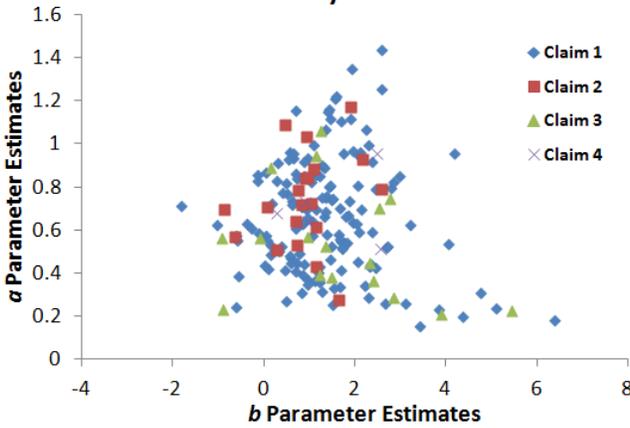
MATH G07 Scatter Plot of 2PL/GPC a and b by Item Type



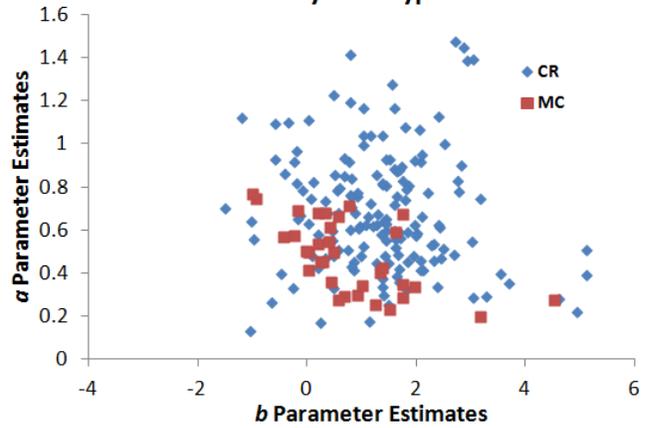
MATH G07 Scatter Plot of 2PL/GPC a and b by Score Category



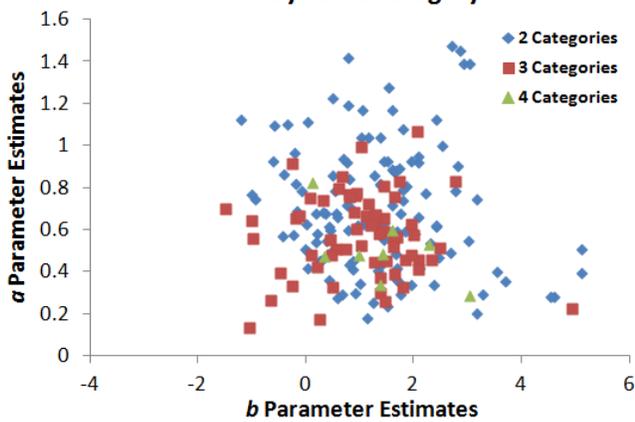
MATH G07 Scatter Plot of 2PL/GPC a and b by Claim



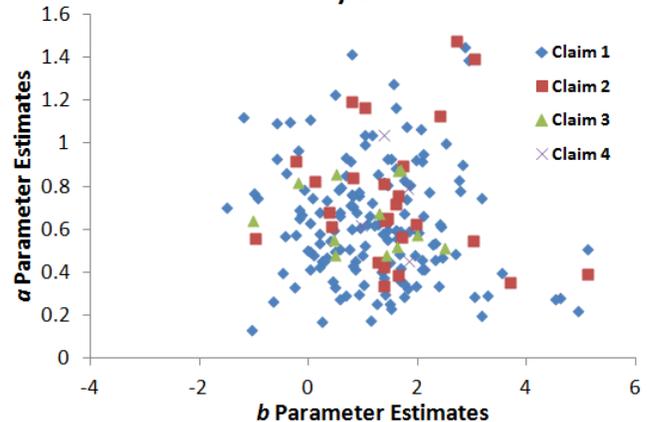
MATH G08 Scatter Plot of 2PL/GPC a and b by Item Type



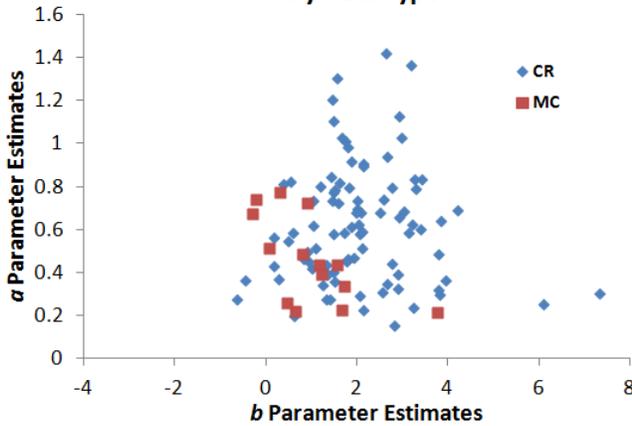
MATH G08 Scatter Plot of 2PL/GPC a and b by Score Category



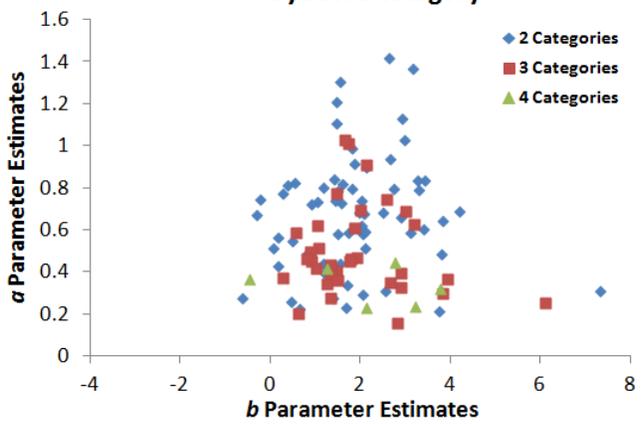
MATH G08 Scatter Plot of 2PL/GPC a and b by Claim



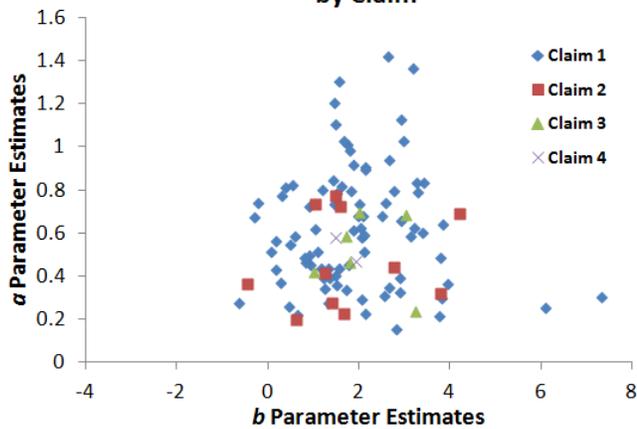
MATH G09 Scatter Plot of 2PL/GPC a and b by Item Type



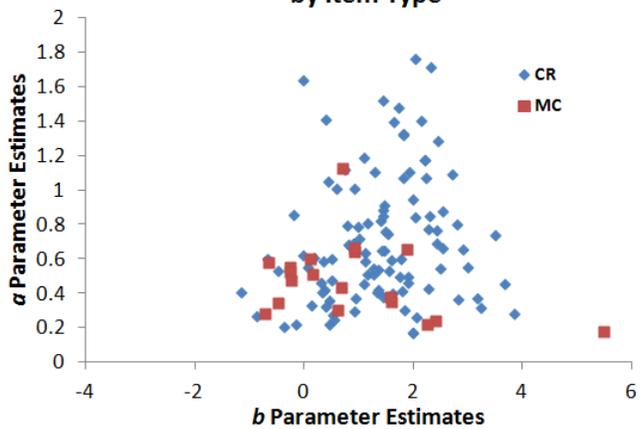
MATH G09 Scatter Plot of 2PL/GPC a and b by Score Category



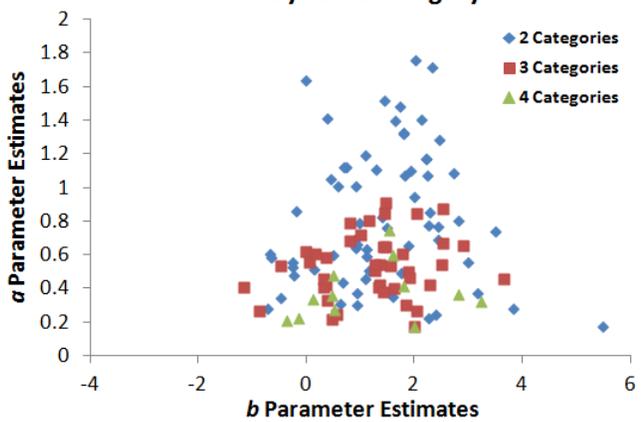
MATH G09 Scatter Plot of 2PL/GPC a and b by Claim



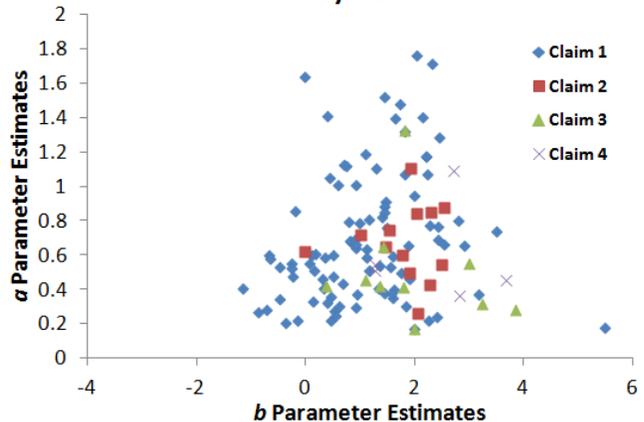
MATH G10 Scatter Plot of 2PL/GPC a and b by Item Type



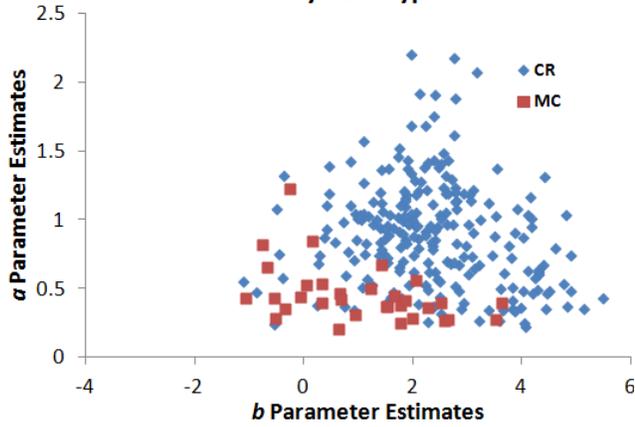
MATH G10 Scatter Plot of 2PL/GPC a and b by Score Category



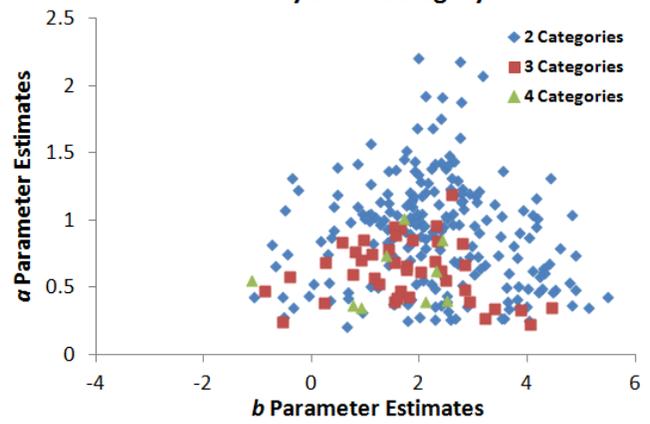
MATH G10 Scatter Plot of 2PL/GPC a and b by Claim



MATH G11 Scatter Plot of 2PL/GPC a and b by Item Type



MATH G11 Scatter Plot of 2PL/GPC a and b by Score Category



MATH G11 Scatter Plot of 2PL/GPC a and b by Claim

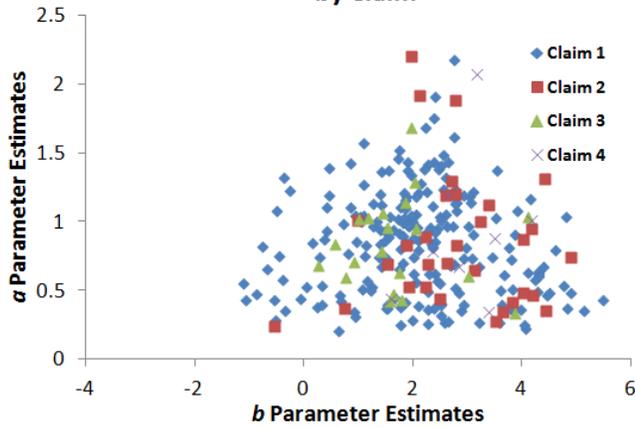
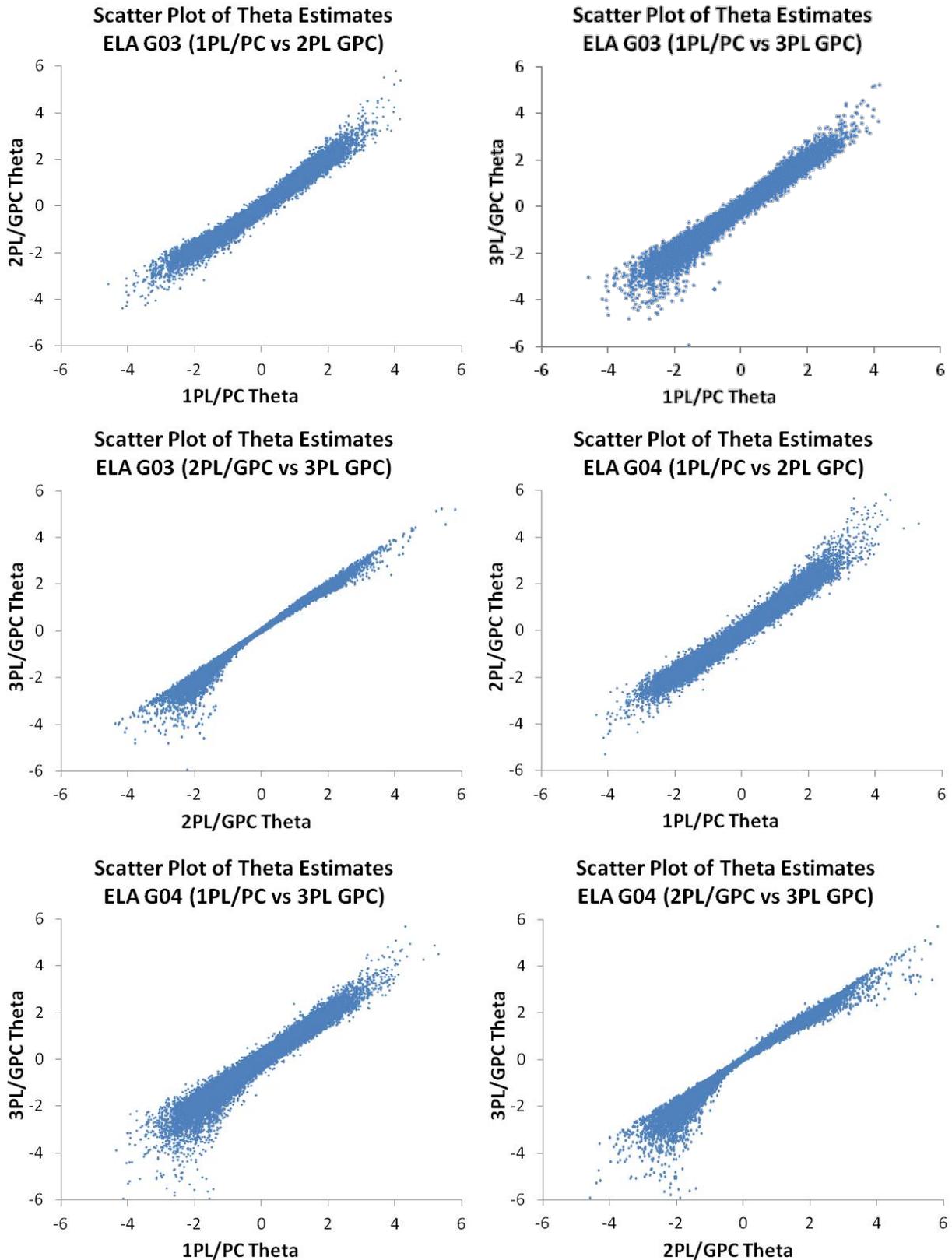
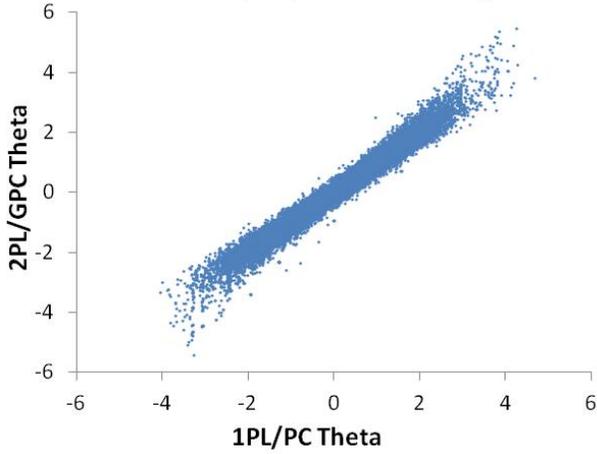


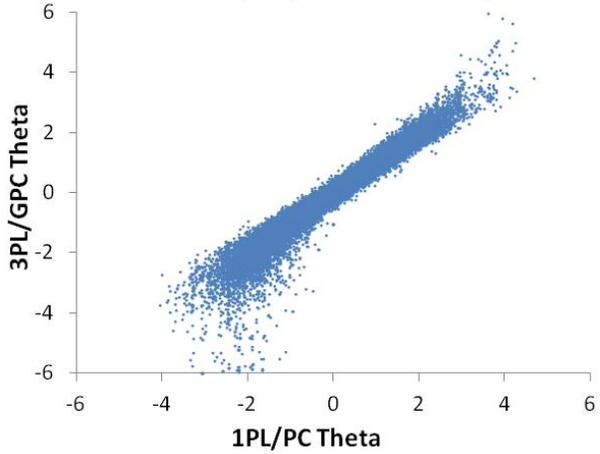
Figure 73. ELA/literacy Scatter Plots of Theta Estimates across Different Model Combinations



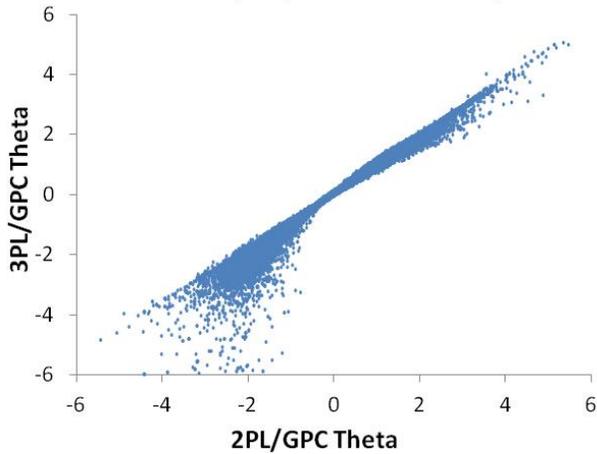
**Scatter Plot of Theta Estimates
ELA G05 (1PL/PC vs 2PL GPC)**



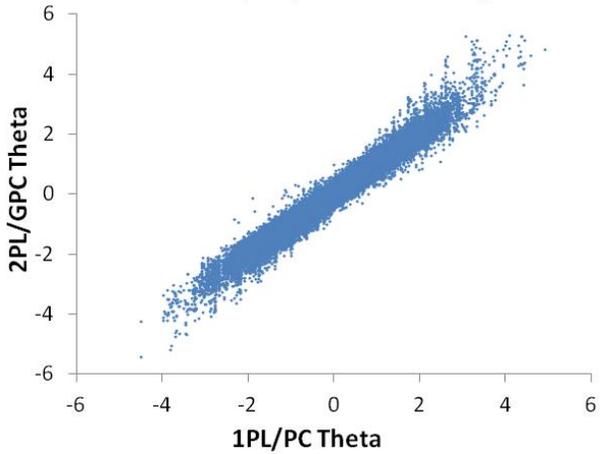
**Scatter Plot of Theta Estimates
ELA G05 (1PL/PC vs 3PL GPC)**



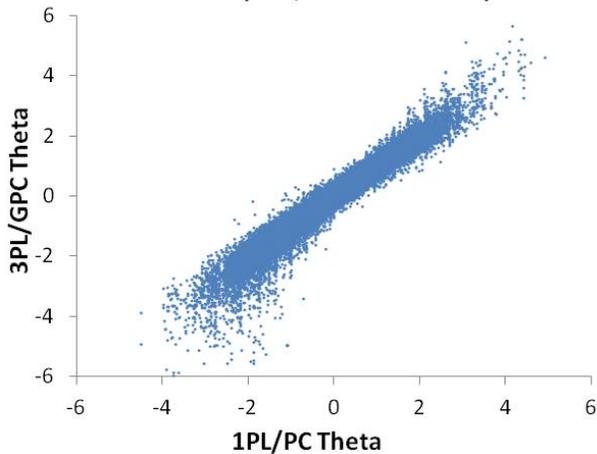
**Scatter Plot of Theta Estimates
ELA G05 (2PL/GPC vs 3PL GPC)**



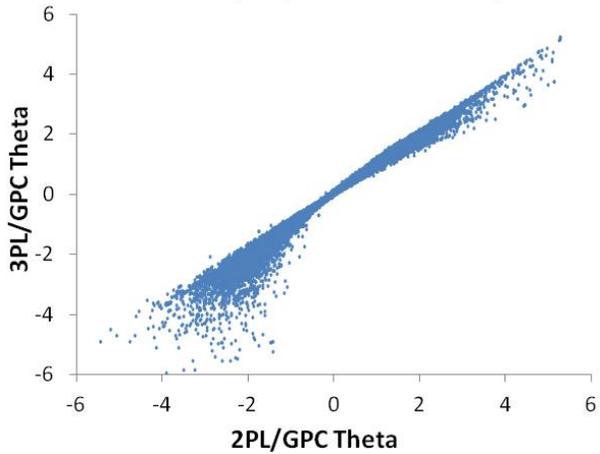
**Scatter Plot of Theta Estimates
ELA G06 (1PL/PC vs 2PL GPC)**



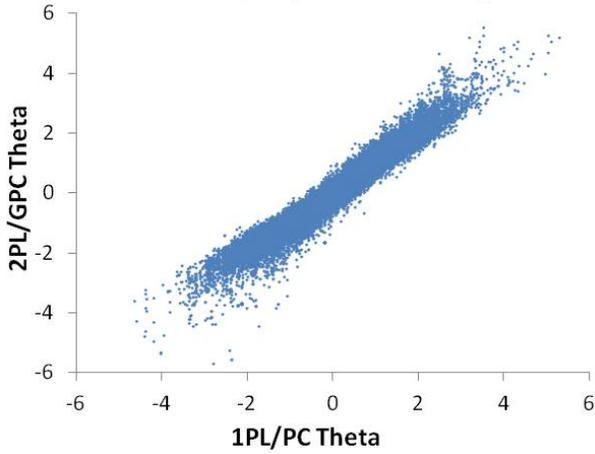
**Scatter Plot of Theta Estimates
ELA G06 (1PL/PC vs 3PL GPC)**



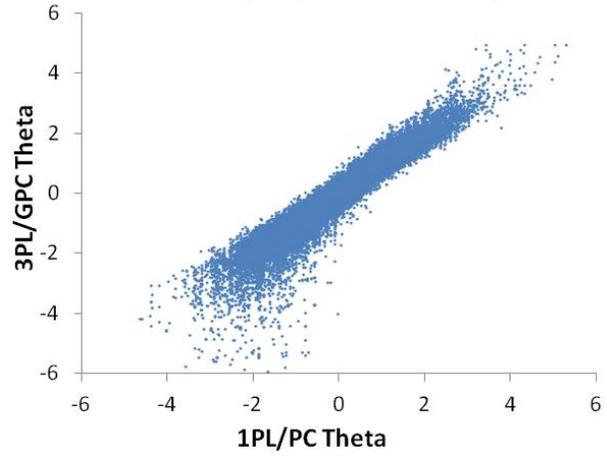
**Scatter Plot of Theta Estimates
ELA G06 (2PL/GPC vs 3PL GPC)**



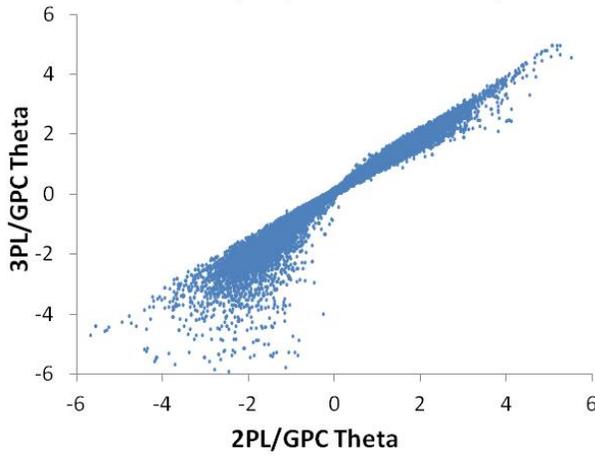
**Scatter Plot of Theta Estimates
ELA G07 (1PL/PC vs 2PL GPC)**



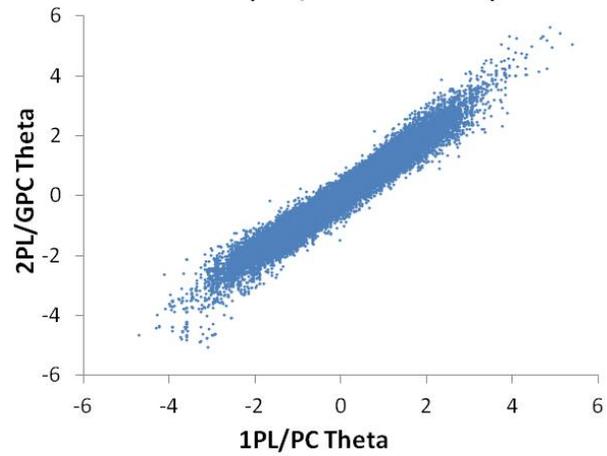
**Scatter Plot of Theta Estimates
ELA G07 (1PL/PC vs 3PL GPC)**



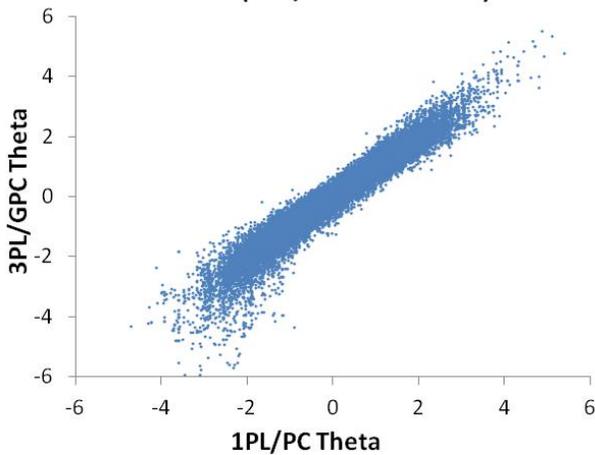
**Scatter Plot of Theta Estimates
ELA G07 (2PL/GPC vs 3PL GPC)**



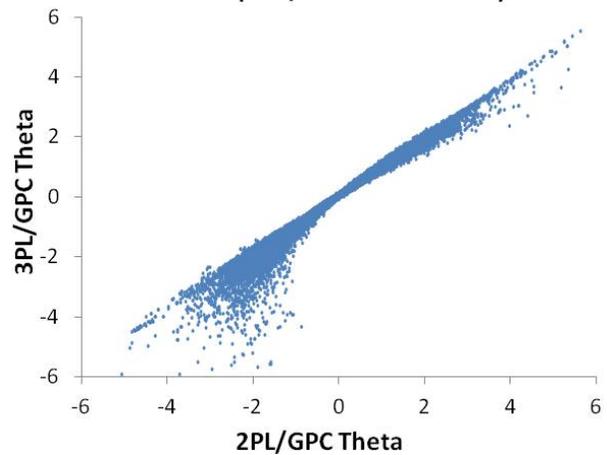
**Scatter Plot of Theta Estimates
ELA G08 (1PL/PC vs 2PL GPC)**



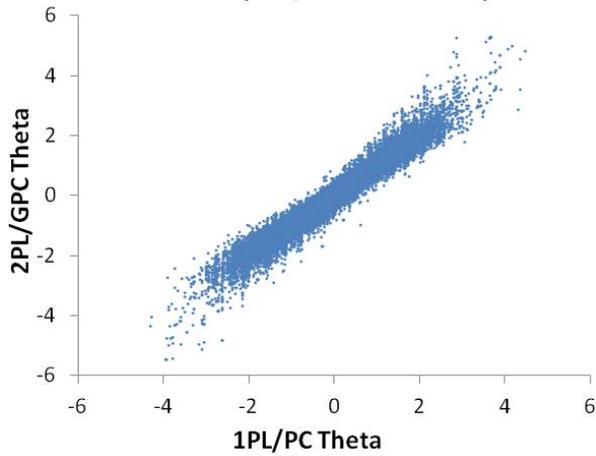
**Scatter Plot of Theta Estimates
ELA G08 (1PL/PC vs 3PL GPC)**



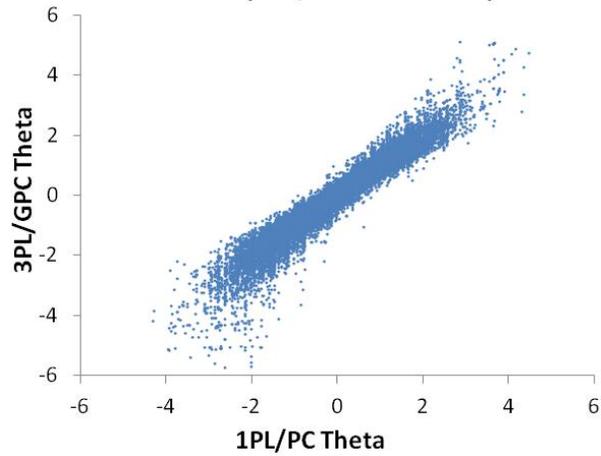
**Scatter Plot of Theta Estimates
ELA G08 (2PL/GPC vs 3PL GPC)**



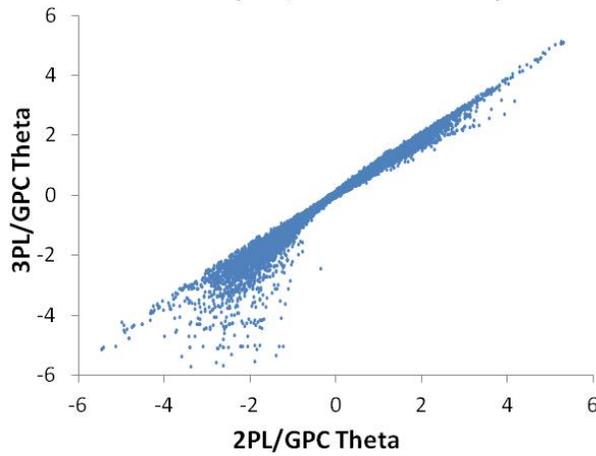
**Scatter Plot of Theta Estimates
ELA G09 (1PL/PC vs 2PL GPC)**



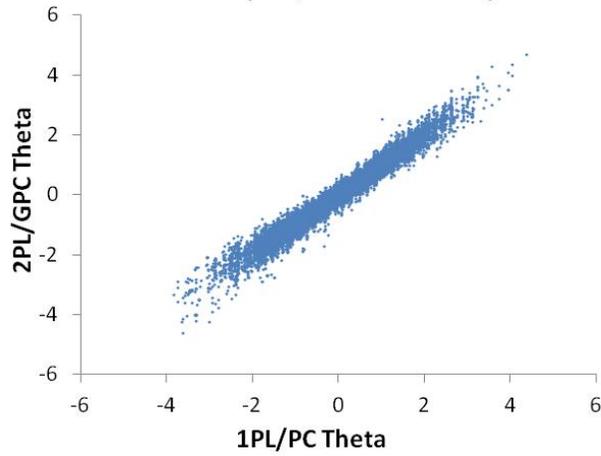
**Scatter Plot of Theta Estimates
ELA G09 (1PL/PC vs 3PL GPC)**



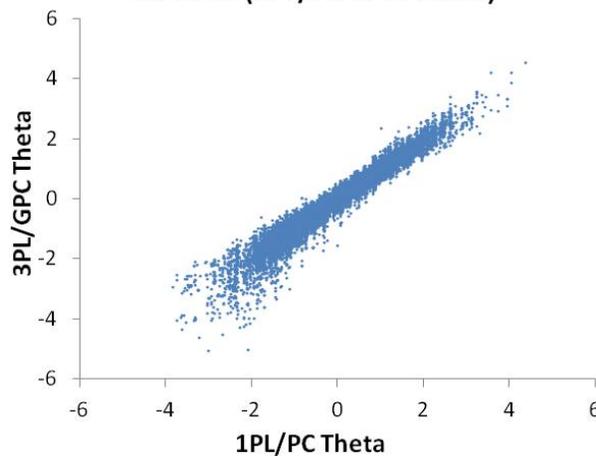
**Scatter Plot of Theta Estimates
ELA G09 (2PL/GPC vs 3PL GPC)**



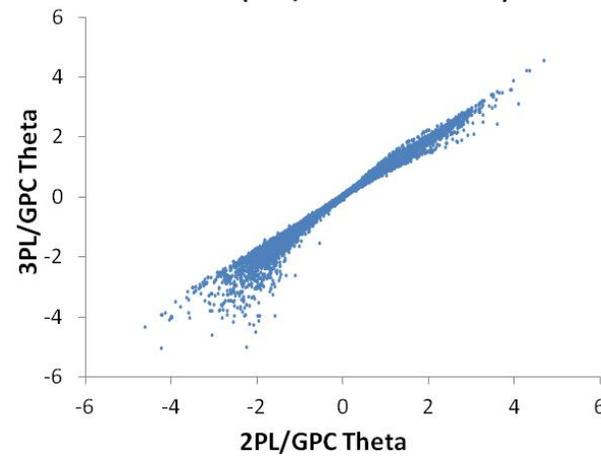
**Scatter Plot of Theta Estimates
ELA G10 (1PL/PC vs 2PL GPC)**



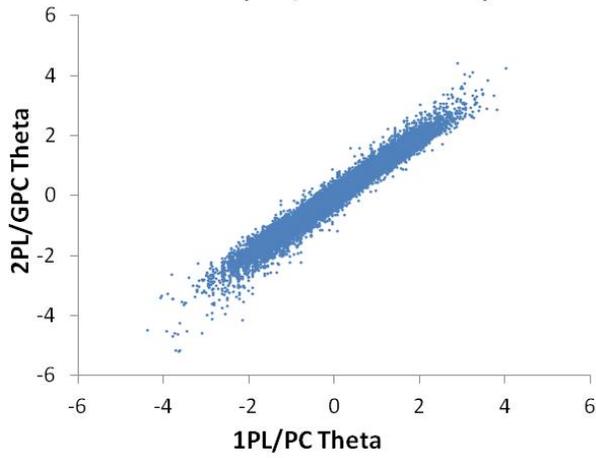
**Scatter Plot of Theta Estimates
ELA G10 (1PL/PC vs 3PL GPC)**



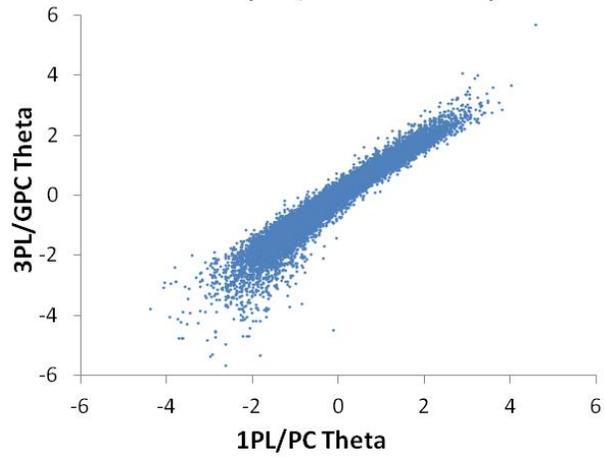
**Scatter Plot of Theta Estimates
ELA G10 (2PL/GPC vs 3PL GPC)**



**Scatter Plot of Theta Estimates
ELA G11 (1PL/PC vs 2PL GPC)**



**Scatter Plot of Theta Estimates
ELA G11 (1PL/PC vs 3PL GPC)**



**Scatter Plot of Theta Estimates
ELA G11 (2PL/GPC vs 3PL GPC)**

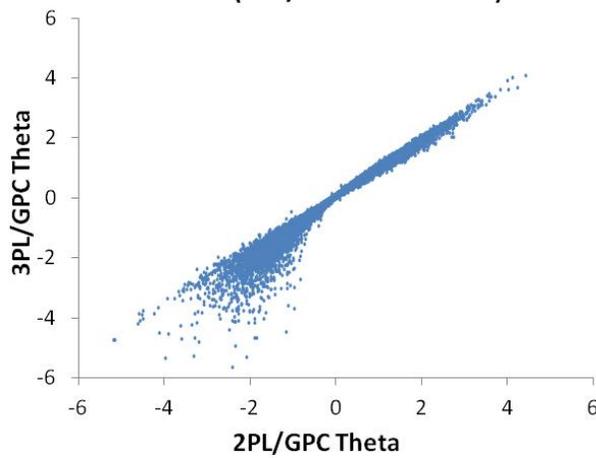
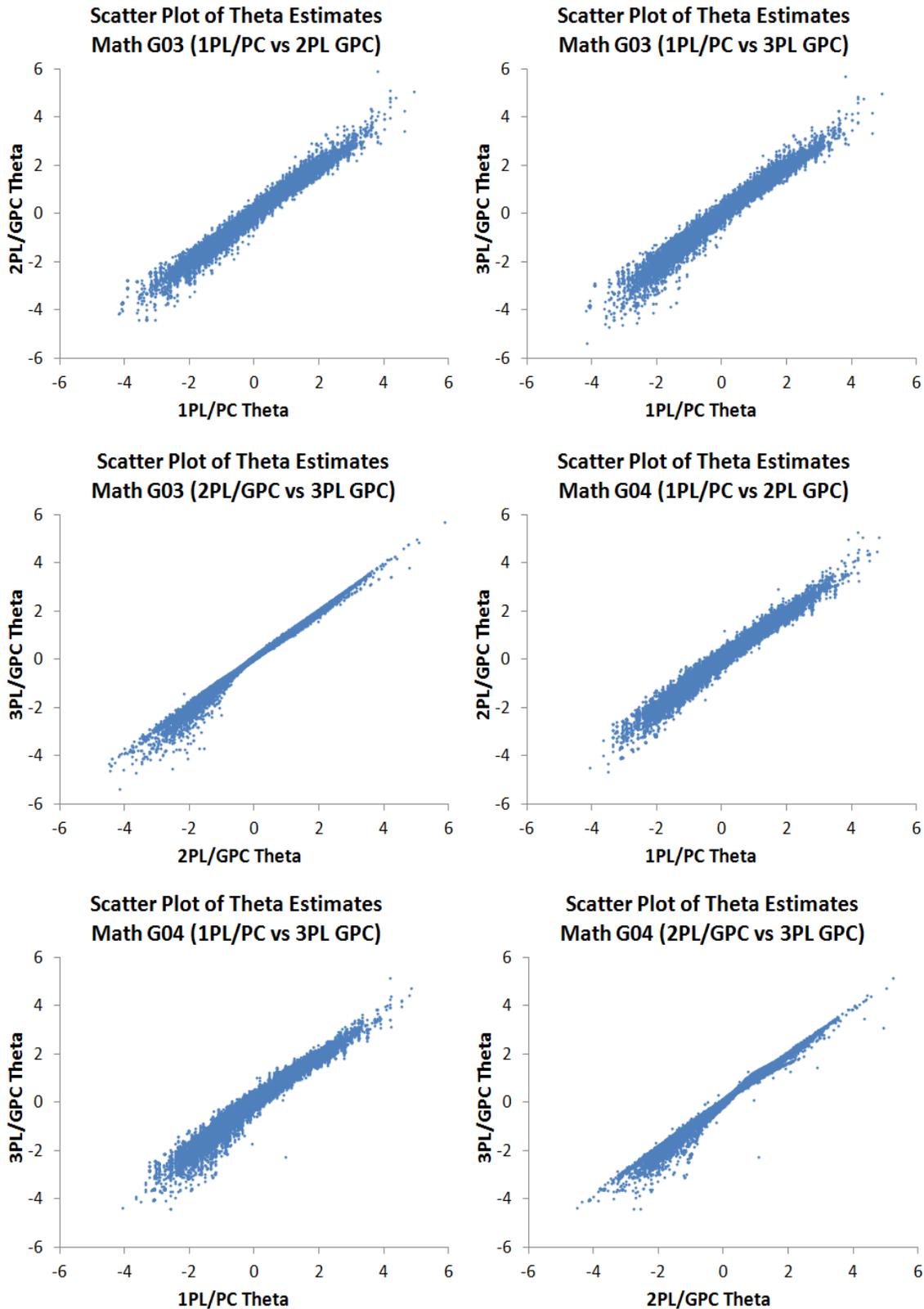
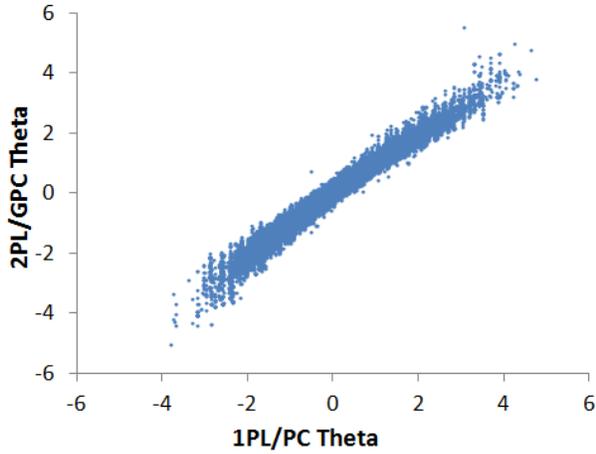


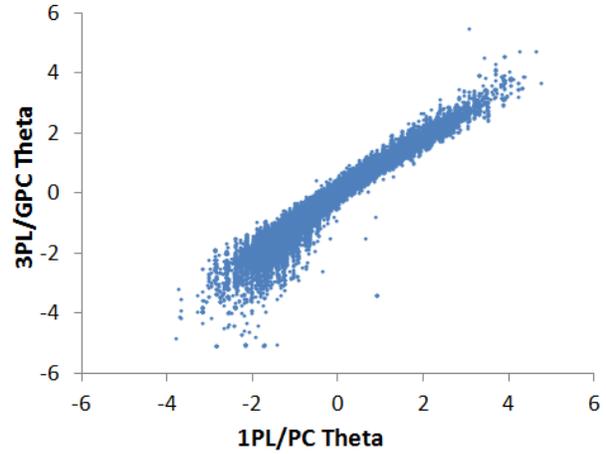
Figure 74. Mathematics Scatter Plots of Theta Estimates Across Different Model Combinations



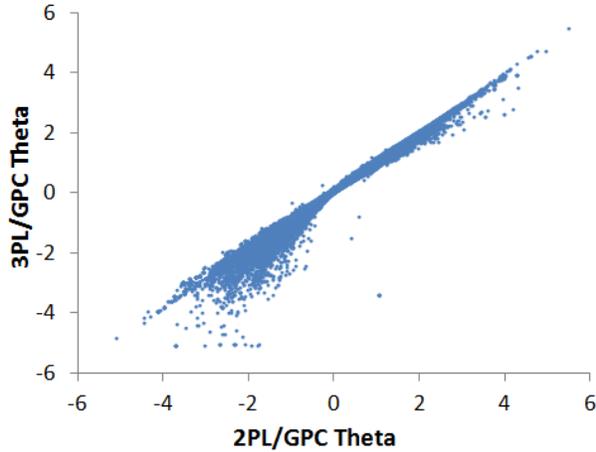
**Scatter Plot of Theta Estimates
Math G05 (1PL/PC vs 2PL GPC)**



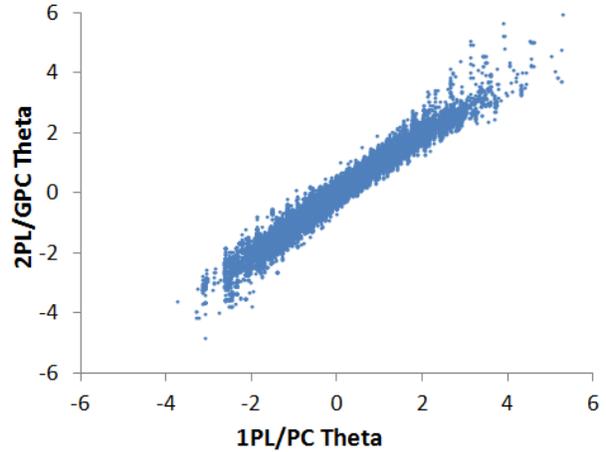
**Scatter Plot of Theta Estimates
Math G05 (1PL/PC vs 3PL GPC)**



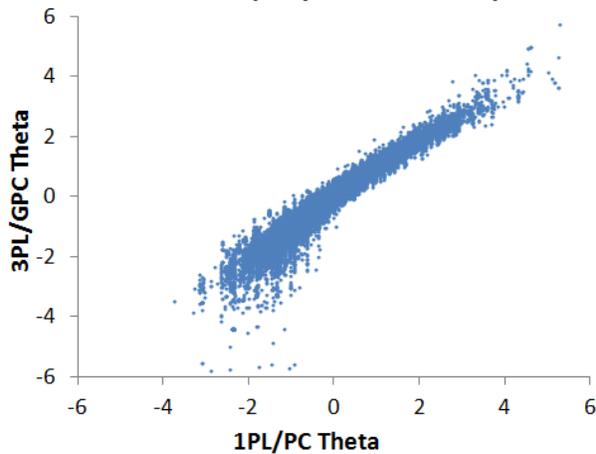
**Scatter Plot of Theta Estimates
Math G05 (2PL/GPC vs 3PL GPC)**



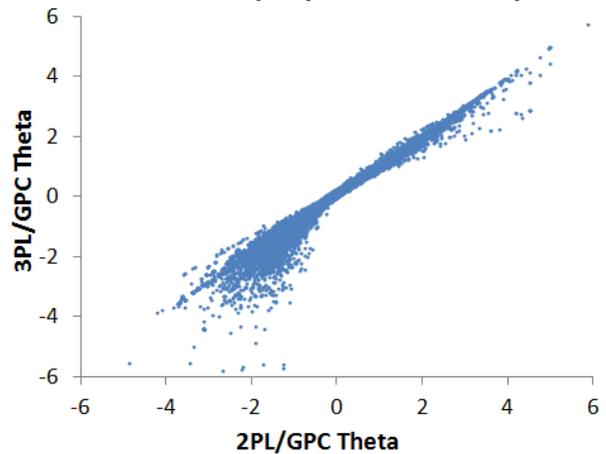
**Scatter Plot of Theta Estimates
Math G06 (1PL/PC vs 2PL GPC)**



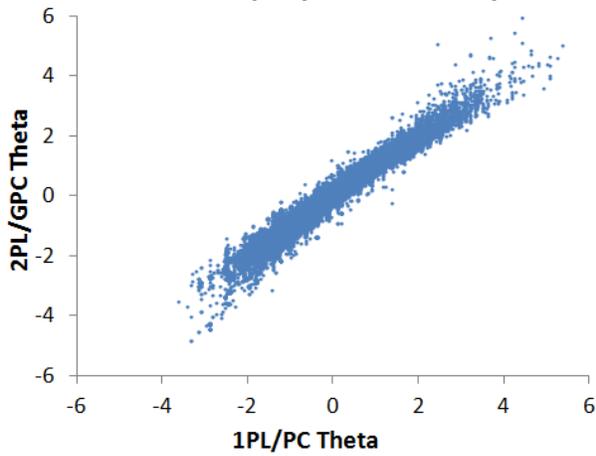
**Scatter Plot of Theta Estimates
Math G06 (1PL/PC vs 3PL GPC)**



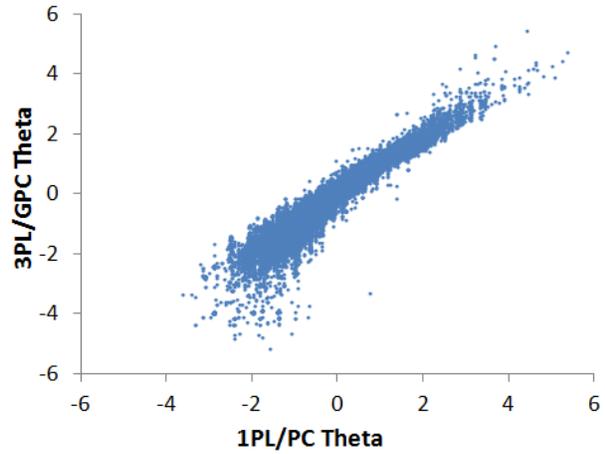
**Scatter Plot of Theta Estimates
Math G06 (2PL/GPC vs 3PL GPC)**



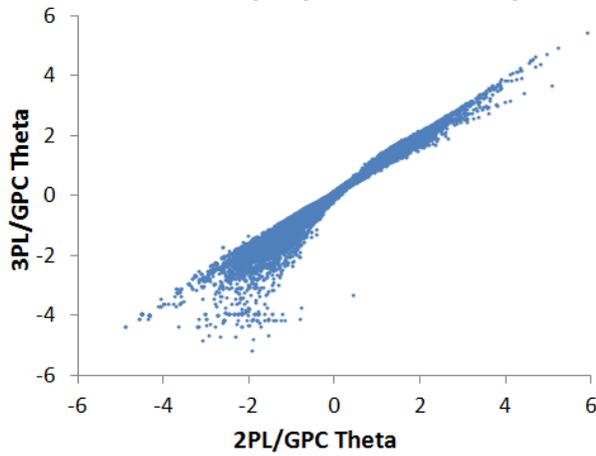
**Scatter Plot of Theta Estimates
Math G07 (1PL/PC vs 2PL GPC)**



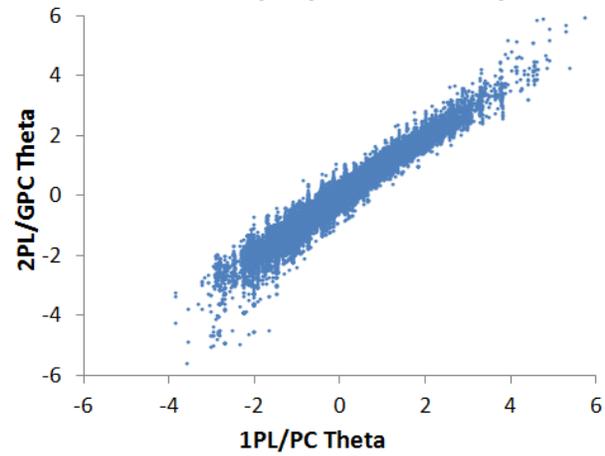
**Scatter Plot of Theta Estimates
Math G07 (1PL/PC vs 3PL GPC)**



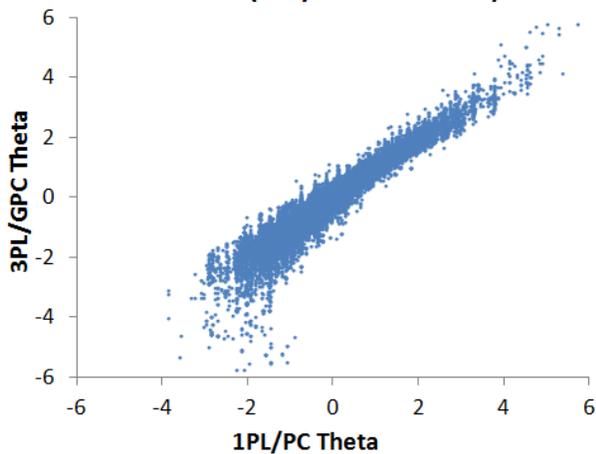
**Scatter Plot of Theta Estimates
Math G07 (2PL/GPC vs 3PL GPC)**



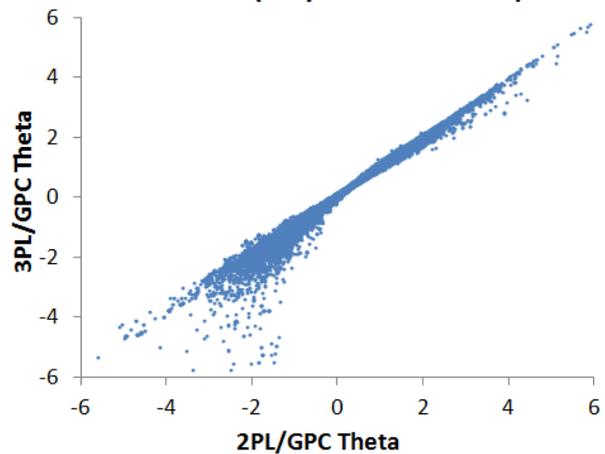
**Scatter Plot of Theta Estimates
Math G08 (1PL/PC vs 2PL GPC)**



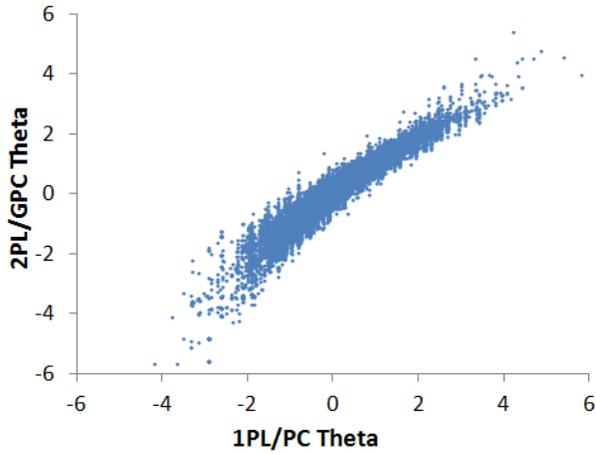
**Scatter Plot of Theta Estimates
Math G08 (1PL/PC vs 3PL GPC)**



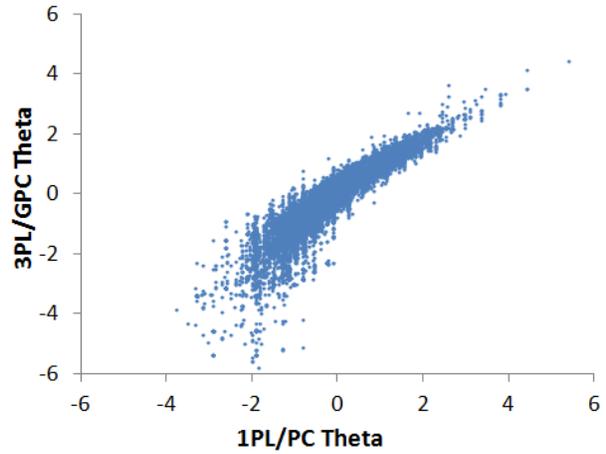
**Scatter Plot of Theta Estimates
Math G08 (2PL/GPC vs 3PL GPC)**



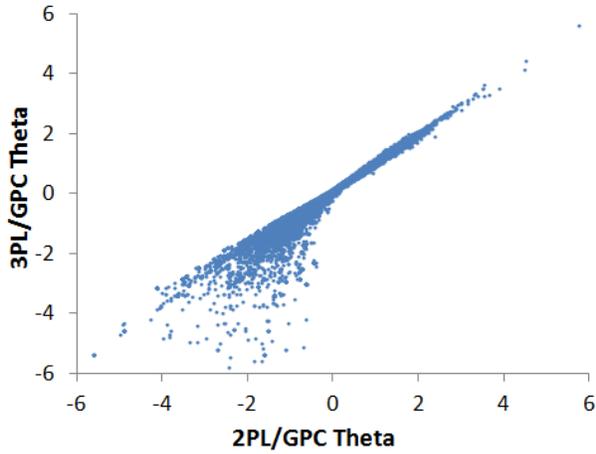
**Scatter Plot of Theta Estimates
Math G09 (1PL/PC vs 2PL GPC)**



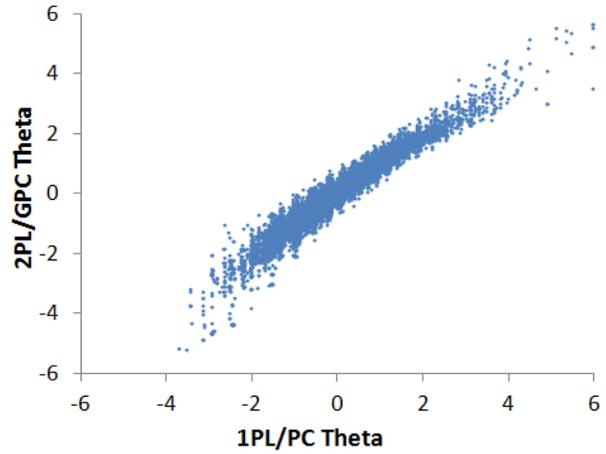
**Scatter Plot of Theta Estimates
Math G09 (1PL/PC vs 3PL GPC)**



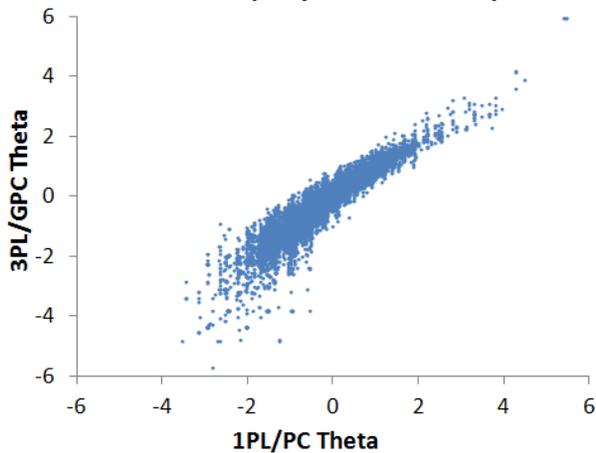
**Scatter Plot of Theta Estimates
Math G09 (2PL/GPC vs 3PL GPC)**



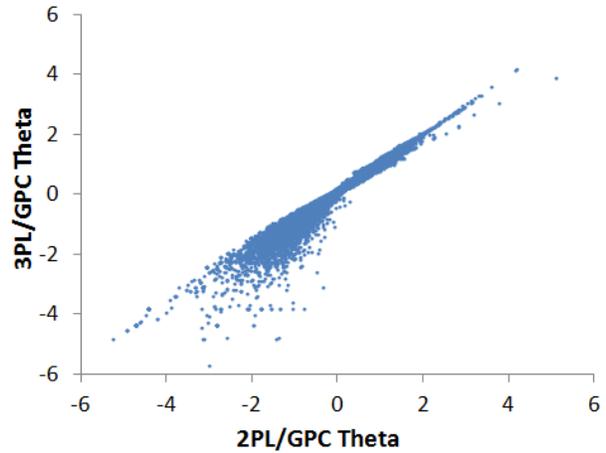
**Scatter Plot of Theta Estimates
Math G10 (1PL/PC vs 2PL GPC)**

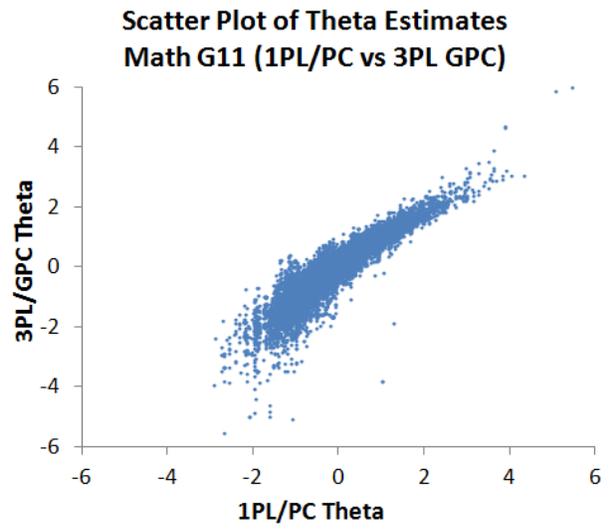
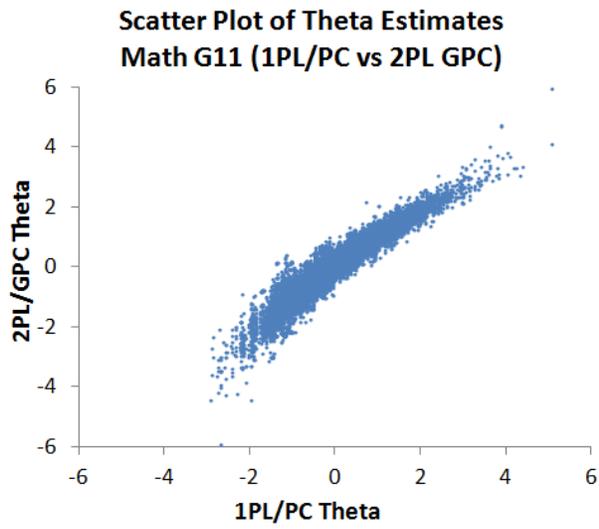


**Scatter Plot of Theta Estimates
Math G10 (1PL/PC vs 3PL GPC)**



**Scatter Plot of Theta Estimates
Math G10 (2PL/GPC vs 3PL GPC)**





References

- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and non-compensatory multidimensional items. *Applied Psychological Measurement*, 13(2), 113–127.
- Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 20, 309–310.
- Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random Coefficients Multinomial Logic Model. *Applied Psychological Measurement*, 21, 1–23.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In (Eds.) Petrov, B. N. & Csaki, F. *Proceedings 2nd International Symposium Information Theory*, 267–281, Budapest, Hungary: Akademia Kiado.
- Bock, R. D. & Zimowski, M. F. (1997). Multiple Group IRT. In W.J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, 433–448. New York: Springer-Verlag.
- Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 27(6), 395–414.
- Briggs, D. C. & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues & Practice*, 28(4), 3-14.
- Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*, 4, 384–414.
- Cochran, W. G. (1977). *Sampling Techniques*, 3rd Edition, New York: John Wiley & Sons.
- Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (RR-83-9). Princeton, NJ: Educational Testing Service.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355–368.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (RR-91-47). Princeton, NJ: Educational Testing Service.
- Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–165). Hillsdale NJ: Erlbaum.
- Ercikan, K., Schwarz, R., Julian, M., Burket, G., Weber, M., & Link, V. (1998). Calibration and Scoring of Tests with Multiple-choice and Constructed-response Item Types. *Journal of Educational Measurement*, 35, 137-155.
- Fitzpatrick, A. R., Link, V., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter Partial Credit Models. *Journal of Educational Measurement*, 33, 291–314.
- Frankel, M. (1983). Sampling theory. In Rossi, Wright & Anderson (Eds.) *Handbook of Survey Research*, 21-67.
- Haebera, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research*, 22(3), 144-149.

- Hanson, B. A., & Beguin, A. A. (1999). *Separate versus concurrent estimation of IRT parameters in the common item equating design*. ACT Research Report 99-8. Iowa City, IA: ACT.
- Hanson, B. A., & Beguin, 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.
- Henson, R. K. & Roberts, J. K. (2006). Exploratory factor analysis reporting practices in published psychological research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3), 393-416.
- Holland, P. W., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41-54.
- Ito, K. Sykes, R.C., & Yao, L. (2008). Concurrent and Separate Grade-Group Linking procedures for vertical Scaling. *Applied Measurement in Education*, 21, 187-206.
- Karkee, T., Lewis, D. M., Hoskins, 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.
- Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22, 131-143.
- Kolen, M. J. (2011) *Issues Associated with Vertical Scales for PARCC Assessments*. White paper written for PARCC. <http://www.parcconline.org/technical-advisory-committee>.
- Kolen, M. J., & Brennan, R. L. (2004). *Test Equating: Methods and Practices*. (2nd ed.). New York, NY: Springer-Verlag.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- McKinley, R. L., & Reckase, M. D. (1983). An application of a multidimensional extension of the two-parameter logistic latent trait model (ONR-83-3). (ERIC Document Reproduction Service No. ED 240 168).
- Mislevy, R.J. (1987). Recent developments in item response theory. *Review of Research in Education*, 15, 239-275.
- Mislevy, R. J., & Bock, R. J. (1990). *BILOG3: Item analysis and test scoring with binary logistic model (2nd ed.) [Computer program]*. Mooresville, IN: Scientific Software.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Orlando, M., & Thissen, D. (2003) Further examination of the performance of S-X2, an item fit index for dichotomous item response theory models. *Applied Psychological Measurement*, 27(4), 289-98.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York, NY: Macmillan.

- Quality Education Data. School Year 2011-2012. MCH. Sweet Springs: MO.
- Reckase, M. D., Martineau, J. A., & Kim, J. P. (2000, July). A vector approach to determining the number of dimensions needed to represent a set of variables. Paper presented at the annual meeting of the Psychometric Society, Vancouver, Canada.
- Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building unidimensional tests using multidimensional items. *Journal of Educational Measurement*, 25, 193–203.
- Reckase, M. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement*, 9(5), 401–412.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer-Verlag.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: a comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40, 331-352.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617.
- Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item-format tests with the one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*, 37, 221–244.
- Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, No. 1.
- Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Weeks, J. P. (2010). **Plink**: An R Package for Linking Mixed-Format Tests Using IRT-Based Methods. *Journal of Statistical Software*, 35, 1–33. URL <http://www.jstatsoft.org/v35/i12/>.
- Williams, V. S., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement*, 35, 93–107.
- Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Yao, L. (2003). **BMIRT**: Bayesian multivariate item response theory. [Computer software]. Monterey, CA: Defense Manpower Data Center.
- Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Applied Psychological Measurement*, 30, 469–492.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245–262.
- Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, 23, 299–325.
- Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–214.

- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (eds.), *Educational Measurement (Fourth Edition)*, Westport, CT: American Council on Education and Praeger Publishing.
- Zeng, J. (2010). *Development of a hybrid method for dimensionality identification incorporating an angle-based approach*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.
- Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. D. (1997). *BILOG-MG: Multiple group item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.
- Zwick, R.; Donoghue, J. R.; & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30, 233–251.

Chapter 7 Field Test Design, Sampling, and Administration

Introduction

A major goal of the Field Test Administration was to provide validity evidence in support of the Smarter Balanced summative and interim assessment purposes. The final Smarter Balanced scales and supporting elements were established in the Field Test using Smarter Balanced Consortium schools, districts, and states that were engaged in the process of implementing the Common Core State Standards (CCSS). The design, while complex, was efficient both in the testing time projected and in the number of items necessary to meet the program requirements. The test design called for each student to be exposed to the full test blueprint and mix of item types that included the less familiar performance task (PT) component. A targeted “Standard Setting Sample” was used to establish the final horizontal and vertical scales and provide the information used to set achievement levels. In a second step, a larger item-pool calibration sample was used for scaling and horizontally linking a robust set of items onto the Smarter Balanced scale established in the previous step. This second calibration step represents the entire item pool at the conclusion of the Field Test. More detail concerning the scaling and linking designs can be found in Chapter 9 Field Test IRT Scaling and Linking Analyses. The items intended for the operational CAT (Computer Adaptive Testing) administration were delivered using Linear-on-the-Fly-Testing administrations. This was advantageous since a content-balanced test blueprint can be delivered to each student using a delivery mode closer to the operational CAT. To the extent possible, samples were selected to represent the demographic characteristics of the Smarter Balanced Governing States. The primary purpose of this chapter is to describe the purposes, design principles, and implementation requirements for the Smarter Balanced Assessment Consortium Field Test administration and its results.

Smarter Balanced conducted a Pilot Test administration in 2013 to inform some aspects of the Smarter Balanced assessments. The Pilot further informed the item types to retain or the ones with entirely new formats to develop as well as the revisions to the test blueprints. Essential elements of the design were changed, such as the inclusion of Classroom Activities in the PTs. Another outcome from the Pilot was the selection of the Item Response Theory (IRT) scaling models. Based on the Pilot analysis, the unidimensional two-parameter model and the generalized partial-credit model for mixed-format items were chosen. While some expected Pilot Test outcomes needed to be revisited in the Field Test, there were somewhat different goals targeted for the Field Test. The major purposes of the Field Test administration were

- to administer and calibrate a sufficient large number of items to ensure a successful operational launch of summative and interim assessments;
- to obtain classical statistics and produce Differential Item Functioning (DIF) analyses to inform item data reviews;
- to establish the final operational horizontal and vertical scales;
- to set the achievement level standards;
- to evaluate the protocols for the test administration and computer delivery system (technology infrastructure); and
- to implement targeted test accommodations and elements of universal design.

The Field Test administration window extended from March 18 to June 6, 2014, for all participating states. In order to achieve these varied purposes, 15,673 items resulted from the Field Test for ELA/literacy (ELA) and mathematics across all grade levels. These items and tasks were delivered to 1,742,208 students from the Smarter Balanced Governing States. To support IRT calibrations, the

number of responses for each item was targeted at 1,200 observations. In many instances, items with fewer than 1,200 observations were calibrated if 500 cases were available. The student samples will be drawn from Smarter Balanced Governing States according to the same two stage, sampling strategy used for the Pilot Test. Some additional items for special study were also included, from the National Assessment of Educational Progress (NAEP) and Program for International Student Assessment (PISA) items at selected grades. These items were placed onto the Smarter Balanced vertical scale and were used in the achievement level setting (standard setting) in the fall of 2014. External items from NAEP and PISA were linked onto the Smarter Balanced scale horizontally using on-grade common items. The relative difficulty of these items was compared with Smarter Balanced items to obtain external measures of performance that could inform the achievement-level setting process.

Field Test Data Collection Design

There were two overall basic steps to the Field Test Design. The first step in the analysis was to establish the vertical and horizontal scales using a robust sample. Items were designated either on-grade or off-grade for vertical linking purposes. Vertical linking items were given across two grade levels (i.e., common items) using content from the lower adjacent grade (e.g., fifth-grade items given to sixth grade). Each student also took a performance task in order to conform to the test blueprint that could be on-grade or off-grade. These items and samples were also used in the achievement-level setting. A representative sample of Smarter Balanced test content and students were needed in order to construct the ordered-item booklets and impact data for the achievement-level setting. The external items from NAEP and PISA were also targeted at the standard settings that were given in selected grades. PISA items were administered in grade 10 for Smarter Balanced. NAEP was given in grades 4, 8, and 11 (in lieu of grade 12). The second step was used to calibrate large numbers of items horizontally in a grade to populate the main IRT item pool. All items administered in the vertical linking step were also administered on-grade to the calibration sample and served as the “common items” for linking. The calibration sample was then linked back to the vertical scale established in the first step using the common/“anchor” items in a grade. The CAT items were administered using Linear-on-the-Fly-Testing, while the performance tasks were fixed forms (not computer adaptive) administered online. The vertical scaling and item pool calibration were separate student samples.

To administer items for vertical scaling, four delivery conditions (summarized in Table 1) were employed in the data collection.

- Condition 1: Tests delivered in this condition include approximately 50 content-representative CAT items and an on-grade or off-grade performance task.
- Condition 2: Tests administered here include approximately 25 content-representative on-grade CAT items, approximately 25 content-representative upper grade CAT items, and an on-grade or off-grade performance task.
- Condition 3: Tests delivered under this condition include approximately 25 content-representative on-grade CAT items, approximately 25 content-representative lower-grade CAT items, and an on-grade or off-grade performance task.
- Condition 4: In this condition, approximately 25 content-representative on-grade CAT items, approximately 25 content-representative NAEP or PISA items, and an on-grade or off-grade performance task were included.

Condition 1 was used to calibrate a large number of items on-grade. Conditions 2 and 3 were targeted at the vertical scaling. Condition 4 was used to calibrate NAEP and PISA items onto the Smarter Balanced scale.

Table 1. Field Test Data Collection Design for ELA/literacy and Mathematics.

Condition	CAT Items	CAT Assignment	P T Assignment	External Items
Vertical Scaling Step				
1	~50 CAT	On-grade only	1 on/off grade	
2	~50 CAT	On-grade/Upper Grade	1 on/off grade	
3	~50 Cat	On-grade/Lower Grade	1 on/off grade	
4	~25 CAT	On-grade only	1 on/off grade	25NAEP/PISA
Calibration Step				
Item Pool Calibration	~50 CAT	On-grade only	1 on-grade	

Field Test Design Principles. The following design principles underpinned the data collected to ensure the best outcomes for the item analysis, calibration, and construction of the vertical scales.

- The tests presented to (and scored for) each student conform to specified test blueprint content requirements. Analyses and resulting scores are more interpretable when the test form administered to each student is appropriately and consistently content balanced. The analyses assumed that students were presented interchangeable test forms measuring essentially the same construct.
- The second principle concerned the linking for the vertical scale being based on substantial item collections administered to representative student samples across grades.
- Thirdly, items should be administered at approximately uniform rates. Assembly specifications for Linear-on-the-Fly-Testing should be detailed and firm enough to ensure consistency of the trait(s) measured but not so rigid that they force distinctly unbalanced rates of item use. Finally, items should be administered to substantial student samples.

In addition, the Field Test had the following characteristics and specifications. All content strata and item types were available to represent the construct in a grade and were administered throughout the entire testing window. The design incorporated two overlapping item and student samples, which consisted of the CAT and performance tasks that were separately delivered events. There was no distinction between summative and interim item pools in these calibration steps. After the calibration was completed and all IRT statistics were available, the items were partitioned into the summative and interim pools. Psychometric characteristics such as item response time, item exposure rates, and specifications for CAT algorithms were not examined due to the information not being available or occurred in other later phases of the program.

Field Test Delivery Modes

For the Field Test, the test delivery modes corresponded to the two separately delivered events. The performance tasks were delivered using computerized fixed forms/linear administrations. For a given performance task, students saw the same items in the same order of presentation and

associated test length. Since performance tasks had a classroom-based activity and were organized thematically, they were randomly assigned at the school level in the Field Test.

CAT (LOFT) Administration. For the CAT pool in the Field Test, Linear-on-the-fly testing (LOFT) was used to administer items to students (Gibson & Weiner, 1998; Folk & Smith, 2002). Note that the LOFT is similar to a CAT in applying content constraints to fulfill the test blueprint. LOFT delivered tests that were assembled dynamically to obtain a unique test for each student from a defined item pool where each student obtains a unique content-conforming test form. The major differences between LOFT and item-level adaptive testing are that no IRT item statistics are used in the administration, and no adaptation based on student responding/ability is incorporated into the delivery algorithm. For dynamic real-time LOFT, item exposure control (e.g., Hetter & Sympson, 1985) can be used to ensure that uniform rates of item administration are achieved. That is, it is not desirable to have some items with many observations and others with correspondingly few in comparison. The LOFT administration is closer to the operational CAT than fixed forms. This permits the scaling to reflect the operational CAT deployment. The major advantage of using LOFT was that delivering parallel fixed test forms with thousands of items in a pool in a given grade and content area was not possible. The disadvantage is that some measures of test functioning are not directly available using LOFT. Observed score (i.e., classical) statistics such as observed test reliability cannot be computed since every student essentially takes a unique test form. Even the definition of a criterion for item-test correlation and for DIF must rely on IRT methods for computing these statistics.

Performance Task Administration. In the case of performance tasks, a Classroom Activity was assigned by school and grade. Four to six separate performance tasks were associated with each Classroom Activity and were to be spiraled to all students at a grade level within a school. Smarter Balanced item and task specifications assumed computer delivery of the items and tasks. Most tasks were long enough to warrant several administration sessions. Such sessions could be same-day, back-to-back sessions with short breaks between sessions. All tasks were administered in controlled classroom settings. Expected time requirements for completing tasks and administration time were provided in subject-specific specifications. Student directions for all tasks began with an overview of the entire task, briefly describing the necessary steps. The overview gives students advanced knowledge of the scorable products or performances to be created (Khattari, Reeve & Kane, 1998). Allowable teacher-student interactions for a task were standardized (i.e., carefully scripted or described in task directions for purposes of comparability, fairness, and security). Teachers were directed not to assist students in the production of their scorable products or presentations.

The group work and teacher directions on how to form and monitor groups for the classroom component of the PTs ensured that no students are disadvantaged simply because of the group to which they are assigned. Group work was not scored but was designed to accomplish such things as the generation of data, the discussion and sharing of provided information, or role-playing for the purposes of the task. If small-group discussions could potentially advantage some groups, the teacher directions required them to use standardized scripts to summarize key points that should have come out of the group discussions. Procedures for standardizing the group-work component will vary depending on the task type. Some task steps will require teachers to play more than a monitoring role and/or students to do small-group work. Teachers and peers were directed not to assist students as they produced their scorable products. The permitted types of teacher and peer interactions for a task were standardized (i.e., carefully scripted and explicitly described in task directions) for purposes of both fairness and security. Although small-group work may be involved in some part of a task, this part was not scored. Students were informed about the nature of the final product(s) at the beginning of the task. The task directions included information for the students on what parts of their work would be scored. All scorable products or performances reflected individual

student products. Every task had multiple scorable products or performances. With responses such as essays, students were informed about which attributes of their work would be scored.

Field Test Design

To achieve the larger Smarter Balanced goals for summative and interim operational assessments and conduct the Field Test, both items and accompanying student samples were targeted to fulfill the scaling and calibration requirements. Table 2 indicates the number of CAT items and performance tasks targeted to support both summative and interim test purposes. To calculate the number of tasks, the assumptions were that each ELA/literacy task would have four items (i.e., scoreable units) and each mathematics performance task would contain six items. Some important scaling-design decisions entailed in Table 2 are listed below. For example, 1,290 items in total were targeted for administration in grade 3. Approximately 300 items were delivered to the vertical scaling sample and 990 items to the item-pool calibration sample. Similar sorts of things pertained to the performance tasks—53 items were targeted for development in total, with 47 in the item pool calibration and 6 administered in the vertical scaling step.

- Students were assigned either ELA/literacy or mathematics in order to minimize the burden of testing time on schools.
- The number of CAT items necessary was estimated using the ratio of a test blueprint (approximately 50 items) to the entire pool as 10:1, which is consistent with judgments for CAT pool size. This was used as a rule-of-thumb to project the number of CAT items needed in a grade/content area to support the Field Test purposes. For example if the blueprint required 50 items for an individual test administration, 500 items collectively would then be needed using this rule.
- The number of performance tasks was determined by their anticipated exposure rates and attrition from the summative pool. Additional items were developed for ELA/literacy due to reading passages and listening in which items are clustered and more flexibility is desired in item selection.
- To achieve these numbers in the operational tests, a 10% overage for CAT item development and a 20% overage for performance tasks development were used to account for expected item attrition during development. For example, in the case of a target of 300 CAT items, 330 were developed but 10% might not be expected to survive content or bias and sensitivity reviews.
- Under this plan, fewer than half the items were targeted for the interim tests compared with the summative tests, with approximately one-third the number of performance tasks used for the interim tests.
- Items were written for grade 11 using the Common Core State Standards were also administered at grades 9 and 10 for vertical scaling.
- The number of CAT items in the interim system was estimated to be 50% of those in the summative system. The number of performance tasks in the interim system was targeted to be 25% of those contained in the summative system.
- In operational settings, the interim pool will be “refreshed” using items that are retired from the summative tests.
- These numbers reflect grade-specific deployment. With a vertical scale, items are used across several grades in order to perform the linking. In this case, items were administered

to the upper-adjacent grade (e.g., grade 5 items given to grade 6 students), expanding the number of available items in a grade.

- Each item/task will have at least 1,200 valid student responses entering the analyses, assuming that uniform exposure control and item pools are proportional to LOFT blueprints. A sample of 1,200 observations was sufficient to obtain reasonably accurate statistics for the 2PL and GPCM IRT scaling models (Stone & Zhang, 2003). A minimum of 500 cases was necessary for inclusion in the calibrations.
- All items in the vertical scaling pool were also administered on-grade in the item-pool calibration step for IRT horizontal linking purposes.

Table 2. Field Test Design for the Item and Performance Task Pools.

Grade	Total		Vertical Scaling		Item Calibration Pool	
	CAT	PT	CAT	PT	CAT	PT
ELA/literacy						
3	1,290	53	300	6	990	47
4	1,290	53	300	6	990	47
5	1,290	53	300	6	990	47
6	1,290	53	300	6	990	47
7	1,290	53	300	6	990	47
8	1,290	53	300	6	990	47
HS* (9,10,11)	3,765	144	300	6	3,465	138
Mathematics						
3	1,125	54	300	6	825	48
4	1,125	54	300	6	825	48
5	1,125	54	300	6	825	48
6	1,125	54	300	6	825	48
7	1,125	54	300	6	825	48
8	1,125	54	300	6	825	48
HS (9,10,11)	3,435	150	300	6	3,135	144

*Note: HS refers to High School.

Numbers and Characteristics of Items and Students Obtained in the Field Test

The sample size for the Field Test is determined by the total number of items to be field tested, the sample size required for each item, and the specific field-testing design. A targeted sample size of 1,200 valid cases for each item was needed to support the production of classical statistics and IRT calibrations. This placed a premium on the item-exposure rates being relatively uniformly distributed since an established sample size was targeted.

When an item or a task was used off-grade for vertical scaling, the effective number of observations for the item/task roughly doubled. Items given on-grade for the standard setting sample were administered as common items in the calibration sample, effectively doubling the observations collectively for item collections. In addition, some statistics will require designated samples, such as students with disabilities and English Language Learners (ELLs). The sizes for the special samples will need to permit differential item functioning (DIF) comparative analysis. The Mantel-Haenszel (MH) and Standardized Mean Difference (SMD) procedures were implemented for DIF studies with IRT ability (θ) as the matching criterion. The minimal sample size for the focal or reference group is 100 and for the total (focus plus reference) group is 400.

Tables 3 and 4 show summaries of the subset of items and tasks that were used for vertical scaling and NAEP/PISA that resulted after test delivery and item exclusions. The distribution of the items according to the claims is also presented. In most cases, the number of linking items was robust. In high school mathematics, there was additional attrition of vertical linking items. A smaller set of items was subsequently eliminated in the IRT scaling step not reflected here. Tables 5 and 6 show the items obtained after the item pool calibration step by claim.

Table 3. Number of Field Test Vertical Scaling Items Obtained by Type for ELA/literacy.

	Grade						
	3	4	5	6	7	8	HS
	ELA/literacy						
Total	261	362	389	363	345	366	517
Claim 1	94	112	145	124	116	132	229
Claim 2	70	101	99	98	99	100	151
Claim 3	50	77	71	70	64	73	60
Claim 4	47	72	74	71	66	61	77
Claim Unknown							
NAEP		28				30	27
PISA							33
Off-grade		120	133	131	107	123	107
On-grade	261	242	256	232	238	243	410

Table 4. Number of Field Test Vertical Scaling Items Obtained by Type for Mathematics.

	Grade						
	3	4	5	6	7	8	HS
	Mathematics						
Total	304	440	401	324	310	336	502
Claim 1	184	240	237	167	162	166	237
Claim 2	17	21	20	24	14	18	27
Claim 3	47	67	67	53	53	44	56
Claim 4	19	30	26	26	27	24	39
Unclassified	37	82	51	54	54	84	143
NAEP		30				33	28
PISA							74
Off-grade		104	95	102	71	73	81
On-grade	304	306	306	222	239	230	319

Table 5. Number of Field Test Item Pool Calibration (Overall Total) Obtained by Claim for ELA/literacy.

	Grade						
	3	4	5	6	7	8	HS
	ELA/literacy						
Total	896	856	823	849	875	836	2,371
Claim 1	317	259	265	274	299	258	867
Claim 2	243	248	241	257	262	241	729
Claim 3	163	157	142	147	152	174	383
Claim 4	173	192	175	171	162	163	392

Table 6. Number of Field Test Item Pool Calibration (Overall Total) Obtained by Claim for Mathematics.

	Grade						
	3	4	5	6	7	8	HS
	Mathematics						
Total	1,114	1,130	1,043	1,018	942	894	2,026
Claim 1	672	677	613	576	519	493	1,123
Claim 2	55	68	55	77	71	59	147
Claim 3	166	145	168	132	120	134	433
Claim 4	68	77	84	68	67	64	185
Unclassified	153	163	123	165	165	144	138

Figures 1 and 2 on the following pages show the frequency distributions for the number of items delivered to students (i.e., test length) for ELA/literacy and mathematics for the vertical scaling. Grades are shown both together and individually either in ELA/literacy or mathematics. These item

counts per student included both the CAT and PT components. In ELA/literacy, bimodal distributions in test length are evident for all grade levels. Many students received approximately a 25-item CAT along with a PT. This was due in part to the inclusion of California, where two half-tests were administered, being resampled to maximize item exposure rates in the delivery system. In mathematics grades 6, 7, and 8, the resulting test length was comparatively short and bimodal only in selected grades. Figures 3 and 4 show similar sorts of information for the item pool calibration.

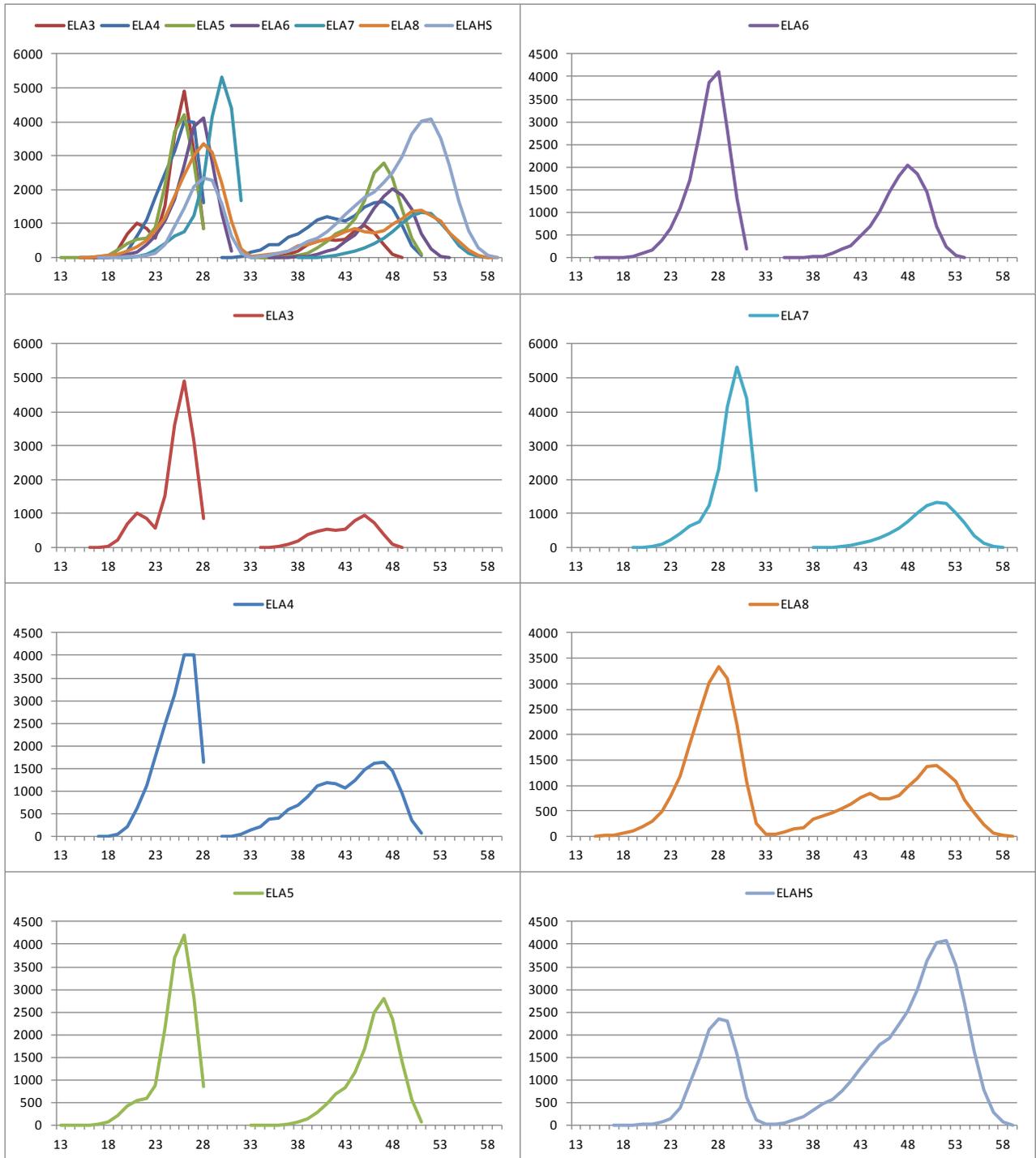


Figure 1. Distributions of Number of Items per Student in the Vertical Scaling (ELA/literacy)

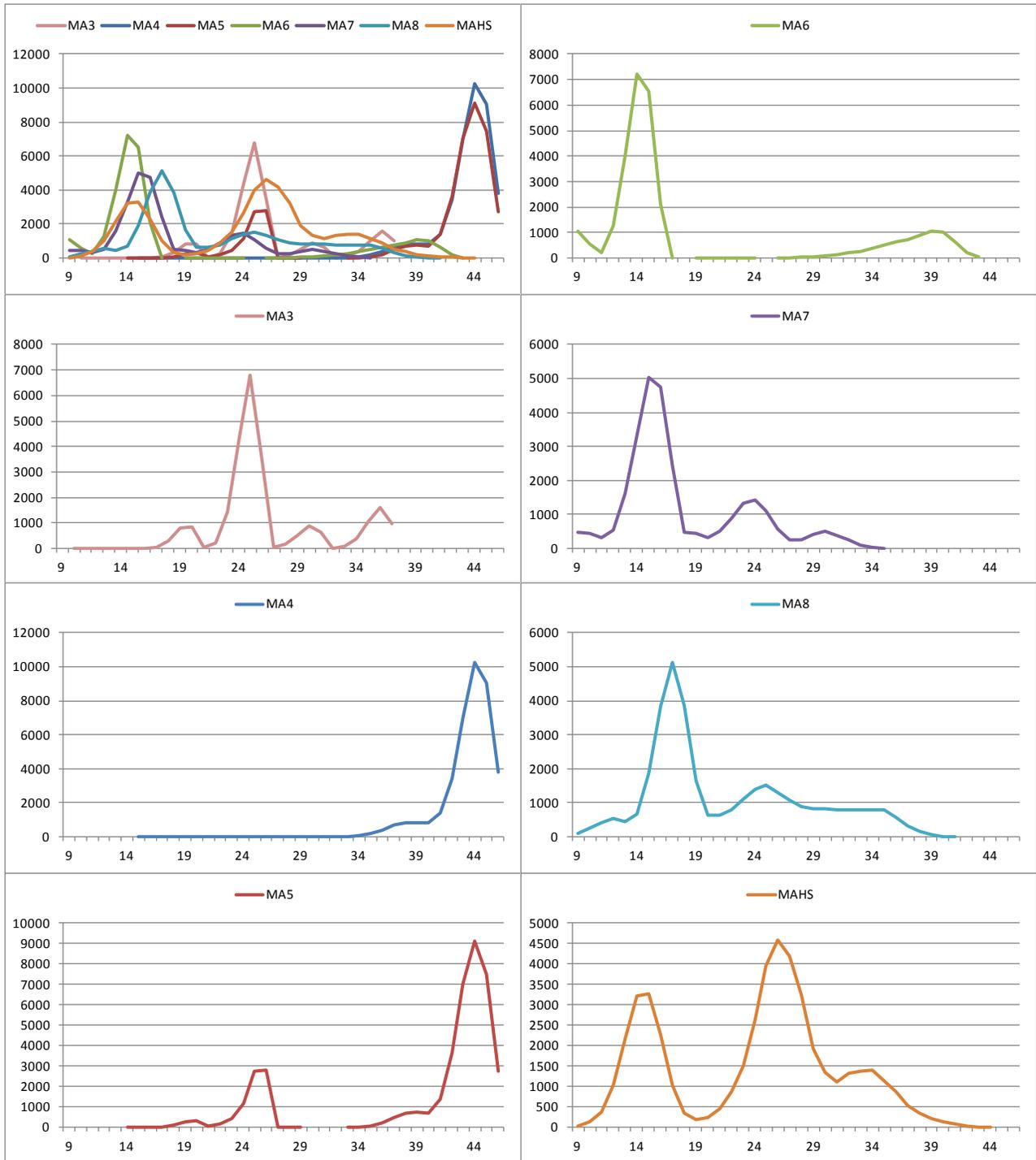


Figure 2. Distributions of Number of Items per Student in the Vertical Scaling (Mathematics)

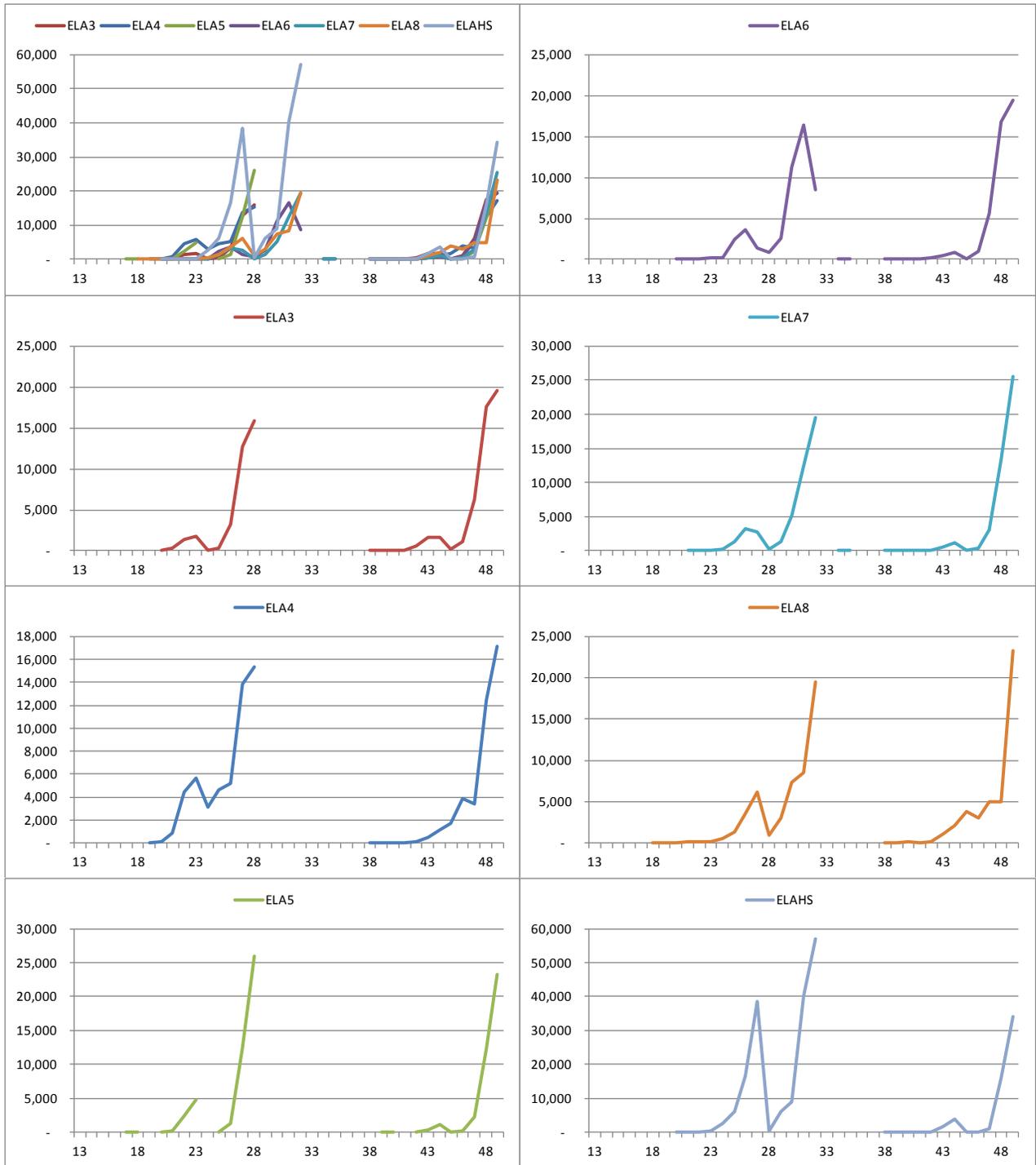


Figure 3. Distributions of Number of Items per Student in the Item Pool Calibration Sample (ELA/literacy)

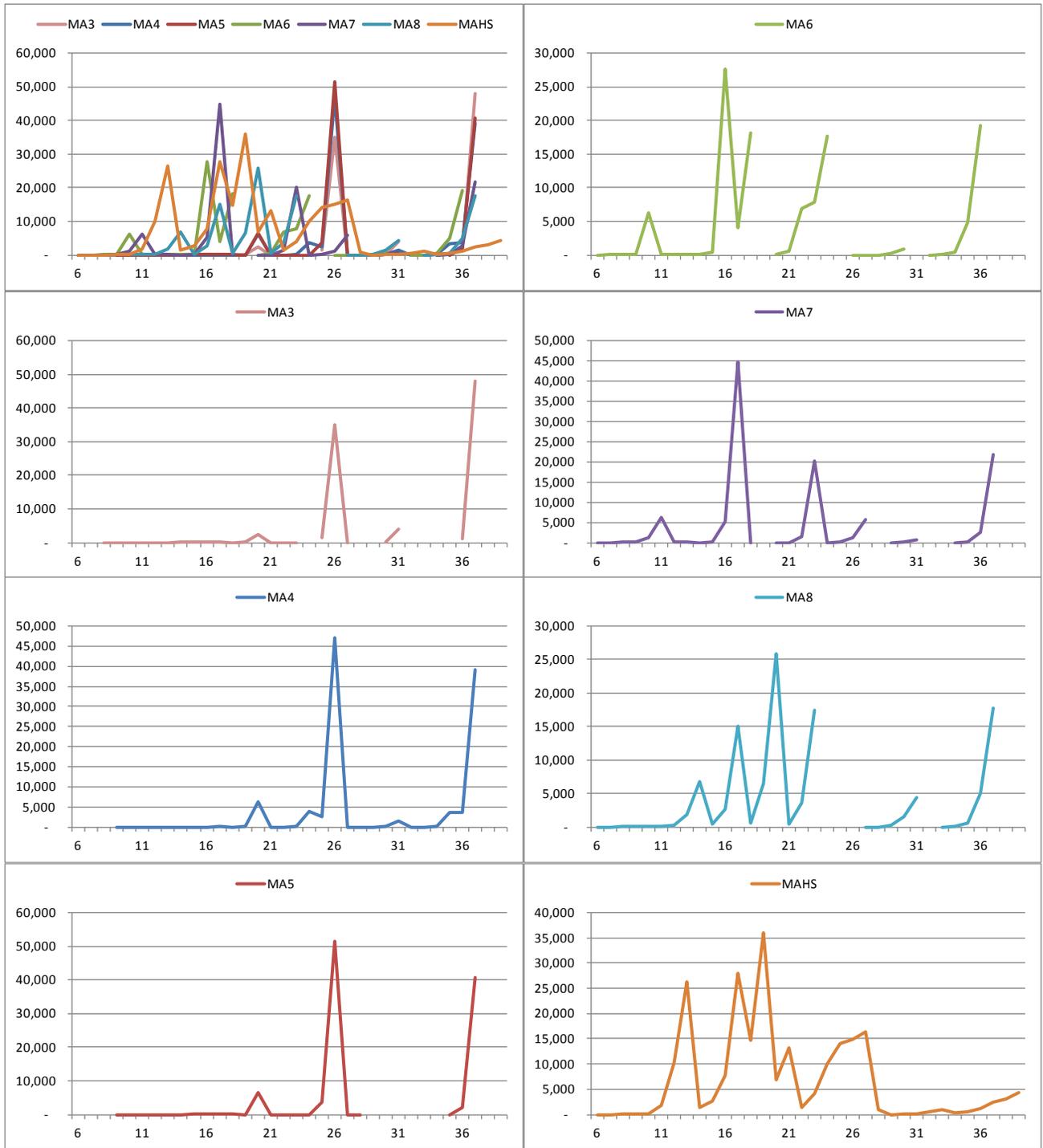


Figure 4. Distributions of Number of Items per Student in the Item Pool Calibration Sample (Mathematics)

Linking PISA and NAEP Items onto the Smarter Balanced Assessments

In the Smarter Balanced Theory of Action, a goal was to establish clear, internationally benchmarked performance expectations. To inform achievement-level setting for Smarter Balanced inferences concerning national and international performance, item collections were obtained from the NAEP and PISA programs. In the United States, national-level data on student achievement stems primarily from two sources: the National Assessment of Educational Progress (NAEP)—also known as the “Nation’s Report Card”—and participation in the Program for International Student Assessment (PISA). NAEP measures fourth-, eighth-, and twelfth-grade students’ performances, most frequently in reading, mathematics, and science, with assessments designed specifically for national and state information needs. Alternatively, the international assessments allow the United States to benchmark its performance to that of other countries in 15-year-olds’ reading, mathematical, and scientific literacy with PISA. These assessments are conducted regularly to allow the monitoring of student outcomes over time. While these assessments appear to have some general similarities, such as the age or grade of students or content areas studied, each program was designed to serve a different purpose and each is based on a separate and unique content framework and set of items. The major features of Smarter Balanced, NAEP, and PISA assessment programs are compared in Table 7.

The Field Test provided the initial opportunity to link selected external items onto Smarter Balanced assessments. A special Field Test data collection condition was required to support this goal. In a secondary step after the vertical scaling calibration of Smarter Balanced items, these external items were calibrated and linked to obtain IRT item parameters on Smarter Balanced scales. This required a content-representative collection of Smarter Balanced Field Test items in ELA/literacy or mathematics and a collection of NAEP and PISA items to be administered to designated students. These external items replaced the off-grade CAT items. Smarter Balanced used released PISA and NAEP items for this purpose. NAEP items were embedded in Smarter Balanced grades 4, 8, and 11 assessments in both content areas. After calibration of Smarter Balanced items, PISA and NAEP were calibrated onto Smarter Balanced scale(s) using randomly equivalent samples and common items. Recognizing these differences in the nature of the construct and test purposes between these programs, the resulting item parameters on the Smarter Balanced scale were used to inform inferences concerning relative performance in the Smarter Balanced achievement-level setting.

Table 7. Comparison of Features across the Smarter Balanced, NAEP, and PISA Assessment Programs.

Design Feature	Smarter Balanced	NAEP	PISA
Construct Definition	ELA/literacy Claims—Reading, Writing, Listening, & Research Text Types: Literary & Information	Reading Frameworks— (writing is separate) Text Types: Literary & Information	Reading Aspects— Text Types: Exposition, Argumentation Instruction, Transaction, & Description
	Math Claims—Concepts and Procedures, Problem solving, Model and Data Analysis, Communicating Reasoning	Math Frameworks— Number Properties and Operations, Measurement, Geometry, Data Analysis, Statistics and Probability, and Algebra	Math Aspects— Quantity, Uncertainty, Space & Shape, Change & Relationships
Item Context Effects and Test Administration Rules	The look and feel of NAEP and PISA items will likely be different from Smarter Balanced items. The provision of glossaries, other test manipulatives, and accommodation rules will differ across programs. Smarter Balanced uses technology-enhanced items, while PISA and NAEP do not.	The basic context will be maintained for NAEP items since they are administered as a set(s).	The basic context will be maintained for PISA items since they are administered as a set(s).
Testing Mode	LOFT delivery on computer and PTs online	Paper 2015: paper scale and computer-based testing scale study	Paper 2015: computer-based testing scale
Testing Window	March–June 2014	February 2013	PISA in April/May
Untimed/Timed	Untimed	Timed	Timed
Delivery Design	Smarter Balanced Field Test LOFT blueprint(s) that took into consideration the embedded set(s) properties, such as their testing length, reading load, and associated number of items	Linear Administration	Linear Administration
Constructed-Response Scoring		Approximately 30–40 % of the NAEP items require rater scoring. Scoring protocols such as training and	Approximately 30% of the PISA items associated with set(s) require rater scoring. Scoring protocols such as

Design Feature	Smarter Balanced	NAEP	PISA
		<p>qualification will need to be followed.</p> <p>Handwritten responses would need to be transcribed for anchors, training and qualification, and calibration papers.</p>	<p>training and qualification will need to be followed.</p> <p>Handwritten responses would need to be transcribed for anchors, training and qualification, and calibration papers.</p>
Cohort/ Population	Sample of 2014 Smarter Balanced Governing States	Based on 2013 US national sample with state-level comparisons	Based on 2012 US sample: 5,000 15-year-old students from 150 schools
Criterion-Referenced Inferences	Designated achievement-level scores in 2014	Proficiency cut scores exist.	Proficiency cut scores do not exist.
Anticipated Program Changes	No change after 2014 in content; schools still transitioning to the CCSS	Transitioning to computer based administration in 2015	Computer based in 2015 and assessment framework will change
IRT Model and Scaling Procedures	Scaling is at the overall content area level using the two-parameter logistic (2-PL)/generalized partial credit model (GPCM).	3-PL and GPCM in reading and math: The main scales are weighted composites of subscales, and calibration is done at the subscale level.	Rasch (calibrated separately with relation to major domain and minor domain)

Field Test Student Sampling Design

Given the purposes and the nature for Smarter Balanced assessments, it is important that the resulting test scales and associated achievement levels represent the performance characteristics of the participating Smarter Balanced states. The characteristics of the Field Test sample will ultimately be reflected in the item statistics and the scales developed.

The Field Test study targeted a representative sample as opposed to a convenience sample of volunteering (self-selected) schools. A multiple-stage stratified sampling with nested cluster sampling was used as the primary approach. The selected samples were intended to be representative of the intended population, which consists of all students from Smarter Balanced member states. The same sampling procedure will be used to recruit samples for nine grades (3–11) and two content areas (mathematics and ELA/literacy), totaling 18 separate samples. The Field Test mirrored the operational test to the extent that every student was designated to take both a CAT and a PT component. An exception was in California where students took both ELA/literacy and mathematics (half-length tests) and a single performance task in the item pool calibration. In the context of vertical scaling and the standard setting sample, some students were assigned both off-grade and on-grade test content configurations. Some states elected to use their own procedures to select representative student samples for the Smarter Balanced Field Test. A volunteer sample was also

collected but was not included in the formal sampling and test analysis. There was no oversampling of any particular subgroup.

Using the test designs given in Table 1, three different Field Test conditions and associated student samples were used that corresponded to assignment to the vertical scaling, item pool calibration, and volunteer conditions. Using school demographic characteristics, such as ethnicity or percentage proficient, a representative sample was selected separately in each state for the vertical scaling and calibration samples. The volunteer schools were used as replacement schools when necessary.

1. The vertical scaling sample (First Sampling Priority) took a content-representative sample of CAT items and performance tasks sufficient to implement vertical and horizontal scaling and construct ordered-item booklets for standard setting. It was essential that the CAT items and their content characteristics closely follow the desired operational pool and the test blueprints. For vertical linking items, there were additional students at two grade levels. Grades 9, 10, and 11 students included under “high school” were used for vertical linking of grades 8 to 11. A subset of students took NAEP/PISA items sufficient for scaling purposes. The sample size targeted for these items was the same as the Smarter Balanced ones. The items in this pool were targeted for delivery to a representative sample from participating Smarter Balanced Governing States. Since this group was used to determine the Smarter Balanced vertical scale, considerable effort was directed at identifying obtaining a representative sample.
2. Calibration Sample (Second Sampling Priority) consisted of students taking all items and tasks on-grade level. The goal was to calibrate a very large number of items in the remaining pool. This pool also included common items from the high-priority vertical scaling pool used to link them onto the final scale. Administration of vertical linking items was not necessary here since this was accomplished in the vertical scaling condition.
3. Volunteer Sample. The remaining participating students were volunteers. More students participated in the Field Test than were needed for scoring and scaling. Since the characteristics of these schools were known, they could be used as replacement schools for the vertical scaling and calibration sample when necessary. These students took both CAT items and a PT, and they may have been required to take tests in one or both content areas.

Defining the Target Population. The defining of the target population provides characteristics for evaluating the representativeness of the resulting sample and the sampling strategies used to obtain it. There were several factors considered in defining the characteristics of the target population for the Field Test, including the model for representing state participation, transition to the Common Core State Standards, and technology infrastructure available for testing.

To be representative of the target population, Field Test samples were recruited to have state representation that was proportional to the size of the state’s student enrollment (“House of Representatives” model). The percentages constitute an implicit sampling weight for each state that is reflected in the vertical scaling item- pool calibration samples. The other model not adopted was the “Senate,” where an equal number of students are contributed by each state. Per Smarter Balanced recommendations, Advisory and Governing States both participated in the Field Test where proportional representation was implemented without considering a state’s “governing” or “advisory” status. A two-state stratified random sampling was used in each member Governing State; the first stage was the state, with schools within the state as the second stage. States were sampled in proportion to their student enrollment size. It was not possible to completely control for situations where states either dropped out of the Consortium or were added after the Field Test.

The second factor considered in defining the target population is the level of Common Core State Standards implementation. Among Smarter Balanced states, the extent to which the CCSS was

implemented at the time of the Field Test administration was likely to vary considerably and no accurate information was available. The target population consisted of students from all Smarter Balanced member states and schools, regardless of the Common Core State Standards implementation level. It is likely that some scale drift will occur over time as the Common Core State Standards are more fully implemented.

The final factor considered in the definition of the target population is the capacity for online testing. While some states are currently administering online state assessments, other states may have districts and schools with varying capacity for online testing. Schools needed to have the specified level of technology infrastructure in order to participate in the Field Test. The necessary technology specifications were communicated to schools. The selected samples include students from schools with varying capacity for online testing. To some extent, the level of technology infrastructure drove the decision-making concerning the number of students that can be selected and reasonably tested online in a given school.

State Participation Conditions. Smarter Balanced states used four types of state participation models for the Field Test data collection. In August 2013, states were asked to provide their anticipated participation model. Table 8 shows the states that were initially planned to be part of the Smarter Balanced 2014 Field Test, as well as whether they recruited/selected their own sample. States could either work with the Smarter Balanced test administration workgroup or implement their own sampling that had to adhere to established criteria for representativeness. Note that some of the state participating models may have changed status due to waiver requests and other state policy decisions. The state participation models were the following:

- Early Adopter states required full participation in both content areas of the Smarter Balanced Field Test in the 2013-14 school year in place of the state's accountability test.
- Blended Basic states constitute states that committed to the number of schools minimally necessary to fulfill the prescribed Field Test sample. These schools did not take the state's accountability test in the 2013-14 school year.
- Blended Enhanced states are states that committed to more than the prescribed Field Test sample to participate in the Field Test (i.e., allowing more students than the prescribed sample but less than 100%). These schools do not take the state's accountability test in the 2013-14 school year.
- Traditional states are ones that require all schools to administer the existing state's accountability test (traditional administration) and committed schools to participate in the Smarter Balanced Field Test in the 2013-14 school year minimally necessary to fulfill the state's prescribed Field Test sample.
- Affiliate or Advisory States are any Smarter Balanced Affiliate or Advisory State member that elects to participate in the Field Test and participated in a strictly voluntary mode.

Throughout the recruiting cycle (September to February), state-participation-status changes occurred, such as Kansas withdrawing from the Consortium, Wisconsin opting not to test high school, North Carolina opting not to require field testing, California choosing to be a modified Early Adopter state, and Missouri opting not to recruit for high school. Adjustments were made to the sample where possible by proportionally allocating these cases to other states.

Table 8. State Participation and Sample Acquisition Conditions.

Condition	State-Led Sampling	Smarter Balanced-Led Sampling
Early Adopter	South Dakota	Idaho Montana
Blended Basic	Nevada Vermont	Kansas Michigan Oregon
Blended Enhanced	Washington Connecticut	California
Traditional	Missouri North Carolina West Virginia Wyoming (Plus) Iowa	Delaware Hawaii (Plus) Maine (Plus) New Hampshire North Dakota South Carolina Wisconsin (no HS)
Affiliate/Advisory	Virgin Islands (VS)	

Technical Sampling Characteristics. In the Smarter Balanced sampling design, the impact of different sampling conditions or procedures was considered. The sampling factors considered were the smallest unit of sampling, the use of sample weights, nonresponse, and sampling from voluntary districts/schools. These are discussed below.

- **Smallest sampling unit.** Whereas stratification generally increases precision when compared with simple random sampling, cluster sampling generally decreases precision. Simple random sampling at the student level cannot be conducted in educational settings since students usually reside within classrooms. In practice, cluster sampling is often used out of convenience or for other considerations. If clusters have to be used, it is usually desirable to have small clusters instead of large ones. The recommendation is to have an entire grade level from a school as the smallest sampling unit; that is, all classrooms from a participating grade level within a selected school would participate. This recommendation was made primarily due to lack of information at the classroom level. Schools were selected with probability proportional to size (PPS) from each stratum. Within a stratum, when the number of students sampled from each school is approximately equal, the students will be selected with approximately equal probability.
- **Use of sampling weights.** Sampling weights can be applied to adjust stratum cells for under- or over-representation. In general, the use of sampling weights, when needed and appropriately assigned, can reduce bias in estimation. Another alternative is to create a self-weighted sample, in which case, every observation in the sample gets the same weight. In other words, the probability of selection is the same for every unit of observation. To achieve this, the sampling plan needs to be carefully designed. In the design, a self-weighted sample results if the following criteria are met:
 - consistent state representation in the target population and Field Test sample,
 - proportional allocation for the second-stage stratified sampling at the school grouping level,

- under each stratum, simple random sampling (SRS) in one-stage cluster sampling if a grade within a school is the smallest sampling unit.
- **Nonresponse/nonparticipation.** The sampling needed to be designed to minimize nonresponse. A typical procedure to handle nonresponse is to act as if the characteristics of the nonrespondents within a stratum/cluster are the same as those of the respondents within the same stratum/cluster. Since a self-weighted sample with a defined sample size is intended, a replacement procedure may be implemented to adjust for nonresponse using specified replacement procedures. Using this procedure, replacement schools were selected within the same stratum to ensure that the schools declining to participate are replaced by schools with comparable characteristics (i.e., the same stratum). Alternatively, to avoid tedious replacement after sampling due to nonresponse, stratified sampling may be conducted based on the pool of Local Educational Agencies (LEAs) that indicated interest in the Field Test.

To minimize this bias, it was of critical importance to ensure that the selected samples for replacement were representative of the Field Test populations, both in terms of performance on state-level achievement tests and demographic characteristics. Once the samples were selected with replacements, their representativeness can be evaluated using state assessment score distributions and demographic summaries comparing samples against the state-level distributions.

Detailed Sampling Procedures. The states that make up the Smarter Balanced Consortium were the primary sampling units (PSUs). PSUs generally consist of large geographic areas that are used for the sampling frame in the first stage of the multistage sample design. Stratification permits a population to be subdivided into mutually exclusive and exhaustive subpopulations. In proportionate stratified sampling, allocation of the sample is assigned to various strata that are made proportionate to the number of population elements that comprise that stratum. Within each PSU (state), additional strata were defined to increase sampling efficiency. The appropriate use of stratification can increase sample efficiency (Frankel, 1983). Stratification is most efficient when the stratum means differ widely and stratification cells are homogeneous. Within strata, homogeneity may result in significant decreases in sampling variance relative to equal-size simple random sampling. In general, it is preferable to define more strata to improve precision if the requisite background information is available and resources permit. Stratification variables were defined as ones that are related to the variable of interest, which is academic achievement.

In this complex sampling design, cluster sampling was used within strata due to administrative constraints and cost-efficiency reasons. Cluster sampling permits the selection of sample cases in groups such as schools as opposed to individuals. Although cluster sampling normally results in less information per observation than a simple random sample, its inefficiency can usually be compensated by a corresponding increase in sample size. A random sample of schools will be selected as clusters within each stratum.

A sampling frame contains the defined population necessary to implement the design, which in this case, includes students from all K-12 public schools from Smarter Balanced member states. The Quality Education Database (QED, 2012) from the MCH Corporation is a commercially available source. This database was used for sampling. One drawback is that the QED did not contain an explicit school performance variable (e.g., percent proficient). The representativeness of the resulting samples were evaluated using state demographic data. The sampling procedures for a given grade level and content area within a given Smarter Balanced member state is briefly described below.

- **Step 1:** For a given grade level and subject area, the number of students sampled from a given state, proportional to its size, was derived from the QED database.

- Step 2: The stratification variables used to combine schools into subgroups within each state were selected. School characteristics that are expected to relate to student performance are preferred. Ideally, state-specific achievement data, which are expected to correlate highly with performance on the Smarter Balanced assessments, was preferred. If this information was not readily available, other stratification variables of interest were considered, such as economic status (percentage of Title I students). For instance, a stratum could be defined to include schools that have a high percentage of students receiving free or reduced-price lunch. It was also necessary to limit the number of stratification variables to one or two and associated number of stratification cells from two to six in order for the sampling plan to be manageable across so many states. It was not necessary for participating states to use the exact same set of stratification variables for some subgroups, since the labels may have varied locally.
- Step 3: Classify all eligible schools within each state into two to six strata based on the stratification variable(s), and determine Field Test sample size per stratum within each state through proportional allocation. This is calculated by multiplying the number of students allocated to each state in step 1 by the percentage of students represented by the specific stratum among all strata within the state. Ideally, student population size was expected to be roughly the same across different strata.
- Step 4: For nonresponse after LEAs were initially selected for Field Test participation, a list of replacement schools was constructed. The replacement schools corresponded to a single stratum cell and were evaluated to ensure a sufficient sample was available. If not, an effort was made to recruit more schools. Selecting Field Test replacements from a list of voluntary schools avoided the potential for extensive rounds of replacements. A separate list of voluntary schools was constructed for each grade from each state.
- Step 5: Field Test participants were selected from the list of voluntary schools, if available, or from the list of all eligible schools within each stratum based on the smallest sampling unit. If the smallest sampling unit is a grade within school, a simple random sample of schools will be picked from each stratum until the overall number of students from selected schools at the grade of interest reaches or approximates the predetermined number. Multiple grade levels from participating schools might have been selected. For example, a school may be selected for grade four participation because it was also selected through stratified sampling for grade 3 participation. Some schools were selected with unique characteristics if the presence of the school in the sample was necessary to ensure sample representativeness. To ensure that Field Test candidates who were excluded from participation were not accidentally included in the sample, the selection of Field Test participants from a stratum took place after removing the excluded candidates from the eligible pool of candidates within a stratum.
- Step 6: The extent to which the selected sample is representative of the target population was evaluated for a grade and content area. Specifically, within a state the variables evaluated for representativeness might have included the following demographics:
 - performance on the last state assessment taken by the students
 - gender
 - ethnicity
 - disability
 - English-proficiency status

- proxy for social economic status (SES)
- Step 7: Replacement schools were selected as needed. Extensive replacement was expected when the sample was selected from all eligible schools and if the state for which sampling is conducted follows the “traditional” participation model. For the particular grade of interest, the stratum to which the school that needs to be replaced belongs was identified. A school was selected from the list of all candidate schools belonging to the same stratum that has the most similar grade size consistent with the replacement school.

Sampling Results

Table 9 shows the expected sampling percentage as the target for each member state and by grade and content area for the vertical scaling. The targeted state participation was the first stage of sampling, which was intended to be proportional to state enrollment size. These results were affected to some extent by late state withdrawal from the Field Test. Due to the need to optimize the item exposure rates in test delivery, the targeted state participation rates (percent of Consortium) were not met. Tables 10 and 11 show the percentage participating for various subgroups for the Smarter Balanced Consortium (i.e., population) and the vertical scaling sample for ELA/literacy and mathematics. Overall, the student characteristics mostly matched the Smarter Balanced population characteristics. However, one of the most notable demographic differences between the target and sample was Hispanics for grade 7 mathematics. The overall number of students obtained was sufficient for conducting observed and IRT analyses. In some cases, items were not calibrated due to an insufficient number of observations per item (i.e., < 500) or score level (< 50). Tables 12, 13, and 14 show the same information for the item pool calibration sample.

Table 9. Sample Size (Percent) for ELA/literacy and Mathematics by Grade and Smarter Balanced State for Vertical Scaling.

State	Percent of Consortium	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
		ELA	Math	ELA	Math										
California	36.4	60.5	26.6	44.9	16.8	24.5	14.2	53.9	37.0	72.4	60.7	53.8	66.2	49.9	60.7
Connecticut	3.2	2.8	18.1	8.1	28.8	17.7	24.9	7.1	15.1	2.2	3.5	6.7	2.1	11.4	6.2
Delaware	0.7	0.0	0.3	0.3	0.1	0.0	0.8	0.1	0.0	0.0	0.0	0.0	0.7	0.1	0.3
Hawaii	1.0	6.5	3.0	2.8	0.2	2.1	0.8	1.1	0.6	1.4	1.6	0.6	1.0	0.7	0.6
Idaho	1.6	1.6	5.5	3.1	8.9	6.9	11.3	2.3	5.7	0.9	2.0	2.7	2.5	15.3	16.4
Iowa	2.9	1.1	1.8	0.0	0.1	1.2	0.3	0.9	0.2	0.6	1.5	0.0	0.0	0.2	0.2
Kansas	2.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Maine	1.1	0.5	0.3	0.7	0.5	0.5	0.5	0.2	0.2	0.1	0.0	0.2	0.1	0.2	0.3
Michigan	9.2	5.5	3.0	4.3	2.7	4.4	1.9	4.1	3.9	3.6	3.0	4.2	3.8	3.6	4.9
Missouri	5.3	1.9	2.4	3.3	1.6	1.1	1.2	2.2	1.6	0.4	1.2	2.2	2.6	3.1	0.7
Montana	0.8	0.5	3.5	1.9	7.0	4.4	6.8	1.2	2.7	0.1	0.6	1.5	0.4	3.6	0.7
Nevada	2.5	3.2	2.0	2.3	1.3	2.5	1.7	2.0	2.0	0.4	2.5	2.8	2.6	1.0	1.8
New Hampshire	1.1	0.5	0.5	0.6	0.1	0.2	0.0	0.1	0.1	0.4	0.8	0.5	0.5	0.3	0.4
North Carolina	8.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4
North Dakota	0.6	0.3	0.2	0.2	0.3	0.2	0.2	0.2	0.1	0.1	0.0	0.2	0.3	0.1	0.1
Oregon	3.3	1.0	1.4	0.5	0.5	0.7	0.6	0.9	0.4	0.4	0.5	0.8	0.6	1.2	0.6
South Carolina	4.2	0.7	0.3	0.2	0.2	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.2
South Dakota	0.7	3.5	8.3	9.7	0.2	7.2	2.2	11.5	11.4	8.6	11.6	9.6	6.1	3.1	1.7
Vermont	0.6	0.5	0.4	0.1	0.1	0.4	0.3	0.1	0.1	0.1	0.0	0.2	0.2	0.5	0.5
US Virgin Islands	0.0	0.0	0.0	0.3	1.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0
Washington	6.0	4.9	18.4	12.1	26.4	23.8	30.7	9.8	18.4	6.3	7.7	10.7	6.6	3.5	2.3
West Virginia	1.6	0.9	0.8	1.3	0.2	0.2	0.3	0.5	0.0	0.7	0.6	0.3	0.4	0.8	0.6
Wisconsin	5.0	3.7	2.7	3.0	2.0	1.9	0.8	1.0	0.4	1.2	2.0	3.0	3.1	0.0	0.0
Wyoming	0.5	0.2	0.3	0.4	0.1	0.3	0.1	0.6	0.1	0.0	0.0	0.0	0.0	0.2	0.3
Sample Size		23,223	24,799	35,689	38,925	31,594	42,380	31,535	29,946	30,913	28,271	35,913	34,880	50,657	47,608

Table 10. Student Demographic Characteristics (in Percentages) for ELA/literacy by Grade for Vertical Scaling.

Demographic Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
	Pop.	Sample	Pop.	Sample										
Male	51.0	51.0	51.1	51.5	51.0	51.1	51.1	51.1	51.1	51.2	51.1	51.0	51.3	50.0
Female	48.5	49.0	48.5	48.5	48.5	48.9	48.5	48.9	48.5	48.8	48.5	49.0	48.7	50.0
Native American or Alaska Native	1.1	1.7	1.1	3.2	1.1	2.9	1.1	2.9	1.1	1.7	1.1	2.7	1.0	2.4
Asian	6.5	8.0	6.7	7.4	6.7	7.2	6.6	7.3	6.5	9.4	6.7	7.1	6.1	7.3
Native Hawaiian or other Pacific Islander	0.8	3.0	0.8	1.5	0.8	1.6	0.8	1.2	0.7	0.8	0.7	0.9	0.7	0.8
Hispanic or Latino	28.7	31.9	28.0	26.8	27.8	28.2	27.4	28.0	26.9	42.2	26.6	27.8	26.7	32.2
Black or African American	10.7	6.9	10.6	7.0	10.8	7.9	11.1	6.8	11.4	5.1	11.4	8.0	11.8	6.9
White or Caucasian	48.7	47.0	49.4	55.8	49.6	61.5	49.9	55.0	50.3	40.6	50.6	54.1	50.2	54.3
Two or More Races	3.6	4.5	3.4	4.3	3.3	4.1	3.2	4.0	3.1	2.8	3.0	3.8	2.7	3.6
Individualized Education Program	11.4	9.4	12.3	10.5	12.5	10.8	12.1	10.4	11.7	9.3	11.5	9.3	10.4	7.1
Limited English Proficient	18.0	18.8	15.3	12.6	12.6	11.0	9.8	9.7	8.7	12.4	7.8	7.7	7.1	6.0
Economic Disadvantaged	55.4	54.1	55.3	51.6	54.6	50.1	54.2	50.7	53.1	56.1	51.9	48.8	48.6	46.6

Table 11. Student Demographic Characteristics (in Percentages) for Mathematics by Grade for Vertical Scaling.

Demographic Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
	Pop.	Sample	Pop.	Sample										
Male	51.0	50.5	51.1	51.3	51.0	51.0	51.1	50.8	51.1	50.5	51.1	50.4	51.3	49.6
Female	48.5	49.5	48.5	48.7	48.5	49.0	48.5	49.2	48.5	49.5	48.5	49.6	48.7	50.4
Native American or Alaska Native	1.1	3.2	1.1	4.0	1.1	3.3	1.1	2.9	1.1	2.3	1.1	1.7	1.0	1.5
Asian	6.5	7.6	6.7	6.3	6.7	6.6	6.6	6.4	6.5	9.6	6.7	10.2	6.1	9.9
Native Hawaiian or other Pacific Islander	0.8	1.2	0.8	0.9	0.8	0.9	0.8	0.9	0.7	1.4	0.7	0.6	0.7	0.7
Hispanic or Latino	28.7	30.6	28.0	26.9	27.8	23.7	27.4	23.4	26.9	35.7	26.6	38.9	26.7	37.1
Black or African American	10.7	8.7	10.6	10.6	10.8	8.5	11.1	6.9	11.4	5.7	11.4	5.5	11.8	5.4
White or Caucasian	48.7	58.5	49.4	66.0	49.6	69.0	49.9	63.6	50.3	45.2	50.6	42.6	50.2	47.6
Two or More Races	3.6	3.6	3.4	4.8	3.3	5.0	3.2	4.6	3.1	3.3	3.0	3.4	2.7	3.6
Individualized Education Program	11.4	9.7	12.3	10.8	12.5	11.0	12.1	9.4	11.7	8.7	11.5	8.2	10.4	6.7
Limited English Proficient	18.0	16.1	15.3	11.2	12.6	8.7	9.8	6.7	8.7	10.1	7.8	9.2	7.1	6.8
Economic Disadvantaged	55.4	52.3	55.3	50.8	54.6	48.1	54.2	45.3	53.1	51.6	51.9	50.8	48.6	48.7

Table 12. Sample Size (Percents) for ELA/literacy and Mathematics by Grade and Smarter Balanced State for the Item Pool Calibration.

State	Percent of Consortium	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
		ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math
California	36.4	38.6	30.0	36.1	25.8	27.5	24.2	36.1	53.9	41.6	60.2	36.7	62.1	62.7	61.5
Connecticut	3.2	14.7	17.1	15.0	21.5	19.7	21.4	17.4	11.6	15.6	9.6	18.1	8.4	12.1	12.1
Delaware	0.7	0.6	0.4	0.5	0.5	0.3	0.6	0.5	0.2	0.6	0.1	0.6	0.4	0.6	0.6
Hawaii	1.0	2.5	1.7	1.7	1.1	1.4	1.3	0.8	0.6	1.0	0.8	0.6	0.9	0.5	0.4
Idaho	1.6	4.8	7.1	5.2	9.3	6.8	10.7	5.5	2.7	5.2	1.4	5.9	1.7	7.1	8.4
Iowa	2.9	0.7	2.1	0.0	0.1	0.6	0.1	0.3	0.1	0.2	0.5	0.0	0.0	0.0	0.0
Kansas	2.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Maine	1.1	0.8	1.0	1.1	1.2	0.9	1.0	0.7	0.6	0.8	0.8	0.6	0.4	0.3	0.3
Michigan	9.2	5.6	5.0	4.9	3.9	5.5	3.5	4.2	3.0	3.7	2.6	3.8	2.8	3.4	3.8
Missouri	5.3	1.5	0.9	1.4	0.7	0.5	0.5	1.2	0.6	0.2	0.5	1.0	0.9	1.0	0.9
Montana	0.8	2.9	3.8	3.5	5.4	4.5	5.7	3.3	1.5	3.3	0.6	3.8	0.8	3.2	3.1
Nevada	2.5	2.3	2.3	2.2	1.7	2.2	1.9	2.3	1.1	1.9	1.5	2.1	1.6	1.4	1.9
New Hampshire	1.1	0.5	0.7	0.6	0.4	0.4	0.4	0.3	0.3	0.6	0.4	0.3	0.3	0.2	0.3
North Carolina	8.6	0.0	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.0	0.1	0.1
North Dakota	0.6	0.4	0.3	0.2	0.3	0.3	0.2	0.3	0.1	0.5	0.1	0.2	0.2	0.2	0.2
Oregon	3.3	1.8	2.0	1.7	1.6	1.8	1.7	1.7	1.7	1.9	1.5	1.4	1.7	0.5	0.6
South Carolina	4.2	0.5	0.2	0.3	0.2	0.5	0.2	0.2	0.0	0.2	0.0	0.0	0.0	0.4	0.4
South Dakota	0.7	4.0	4.2	6.0	2.8	5.1	4.0	5.9	3.4	5.4	3.5	5.7	2.2	3.2	3.1
Vermont	0.6	0.5	0.4	0.4	0.4	0.4	0.3	0.5	0.2	0.4	0.2	0.4	0.2	0.3	0.4
US Virgin Islands	0.0	0.0	0.0	0.5	0.8	0.0	0.0	0.0	0.0	0.7	0.7	0.0	0.0	0.1	0.1
Washington	6.0	12.0	16.1	14.1	18.8	17.5	18.4	14.7	16.7	12.6	12.7	15.4	13.3	1.7	0.9
West Virginia	1.6	0.9	0.8	0.8	0.6	0.3	1.0	1.0	0.5	0.6	0.4	0.5	0.5	0.6	0.7
Wisconsin	5.0	4.2	3.9	3.4	2.7	3.3	2.6	2.7	1.1	2.7	1.9	2.9	1.8	0.0	0.0
Wyoming	0.5	0.2	0.2	0.3	0.2	0.3	0.1	0.4	0.2	0.0	0.0	0.0	0.1	0.1	0.1
Sample Size		85,889	95,143	94,915	109,441	88,293	108,412	93,536	117,691	93,431	117,049	98,433	116,459	261,405	262,111

Table 13. Student Demographic Characteristics (in Percentages) for ELA/literacy by Grade for the Item Pool Calibration.

Demographic Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
	Pop.	Sample	Pop.	Sample										
Male	51.0	51.4	51.1	51.4	51.0	51.2	51.1	51.2	51.1	51.4	51.1	51.6	51.3	51.2
Female	48.5	48.6	48.5	48.6	48.5	48.8	48.5	48.8	48.5	48.6	48.5	48.4	48.7	48.8
Native American or Alaska Native	1.1	2.8	1.1	3.2	1.1	3.1	1.1	3.3	1.1	3.2	1.1	3.2	1.0	1.9
Asian	6.5	7.3	6.7	7.5	6.7	7.1	6.6	6.8	6.5	7.6	6.7	6.9	6.1	8.2
Native Hawaiian or other Pacific Islander	0.8	1.5	0.8	1.1	0.8	1.2	0.8	0.9	0.7	0.8	0.7	0.9	0.7	0.9
Hispanic or Latino	28.7	30.2	28.0	28.4	27.8	28.6	27.4	28.8	26.9	32.8	26.6	27.8	26.7	30.3
Black or African American	10.7	10.0	10.6	9.3	10.8	10.2	11.1	10.4	11.4	9.9	11.4	10.6	11.8	9.9
White or Caucasian	48.7	54.1	49.4	56.6	49.6	58.8	49.9	57.3	50.3	52.3	50.6	57.3	50.2	50.3
Two or More Races	3.6	4.1	3.4	4.2	3.3	4.0	3.2	3.9	3.1	3.5	3.0	3.4	2.7	3.2
Individualized Education Program	11.4	10.3	12.3	10.9	12.5	11.5	12.1	11.1	11.7	10.4	11.5	10.4	10.4	8.1
Limited English Proficient	18.0	16.6	15.3	13.6	12.6	11.1	9.8	9.7	8.7	9.7	7.8	7.4	7.1	6.2
Economic Disadvantaged	55.4	53.4	55.3	51.9	54.6	50.6	54.2	51.1	53.1	52.8	51.9	48.6	48.6	46.2

Table 14. Student Demographic Characteristics (in Percentages) for Mathematics by Grade for the Item Pool Calibration.

Demographic Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
	Pop.	Sample	Pop.	Sample										
Male	51.0	51.2	51.1	51.3	51.0	51.4	51.1	51.1	51.1	51.5	51.1	51.2	51.3	50.9
Female	48.5	48.8	48.5	48.7	48.5	48.6	48.5	48.9	48.5	48.5	48.5	48.8	48.7	49.1
Native American or Alaska Native	1.1	3.0	1.1	4.4	1.1	4.3	1.1	2.1	1.1	1.9	1.1	2.0	1.0	2.0
Asian	6.5	7.0	6.7	7.0	6.7	6.9	6.6	8.2	6.5	8.3	6.7	8.2	6.1	8.5
Native Hawaiian or other Pacific Islander	0.8	1.4	0.8	1.1	0.8	1.1	0.8	0.9	0.7	1.0	0.7	0.9	0.7	0.9
Hispanic or Latino	28.7	31.3	28.0	29.0	27.8	27.7	27.4	33.7	26.9	36.2	26.6	33.0	26.7	31.5
Black or African American	10.7	9.8	10.6	9.7	10.8	8.5	11.1	7.9	11.4	8.3	11.4	8.0	11.8	9.9
White or Caucasian	48.7	56.2	49.4	59.3	49.6	61.4	49.9	50.7	50.3	46.9	50.6	49.7	50.2	49.0
Two or More Races	3.6	3.8	3.4	4.4	3.3	4.4	3.2	4.4	3.1	3.7	3.0	3.9	2.7	3.2
Individualized Education Program	11.4	10.4	12.3	11.1	12.5	11.3	12.1	11.1	11.7	10.5	11.5	10.2	10.4	8.0
Limited English Proficient	18.0	17.6	15.3	13.3	12.6	10.5	9.8	11.5	8.7	11.0	7.8	9.4	7.1	7.2
Economic Disadvantaged	55.4	53.1	55.3	52.1	54.6	50.0	54.2	52.2	53.1	53.3	51.9	49.9	48.6	46.4

Field Test Administration and Security

Student Participation. All students in the specified grade levels were eligible to participate in the Smarter Balanced Field Test unless they received a special exemption. In general, if a student participated in the Consortium state's general education accountability assessment or took the Alternate Assessment based on Modified Achievement Standards (AA-MAS) and attended a school participating in the Field Test, the student was eligible to participate. Consistent with the Smarter Balanced field-testing plan, all students, including students with disabilities, English language learners (ELLs), and ELLs with disabilities, had an equal opportunity to participate in the Smarter Balanced Field Test. All students enrolled in grades 3–11 selected to participate in the Smarter Balanced ELA/literacy or mathematics assessment are required to participate except

- students with the most significant cognitive disabilities who meet the criteria for the mathematics alternate assessment based on alternate achievement standards (approximately 1% or fewer of the student population);
- students with the most significant cognitive disabilities who meet the criteria for the English language/literacy alternate assessment based on alternate achievement standards (approximately 1% or fewer of the student population); and
- ELLs who enrolled within the last 12 months prior to the beginning of testing in a US school and have a one-time exemption. These students may instead participate in their state's English language proficiency assessment consistent with state and federal policy.

Practice and Training Tests. To expose students to various items types and other features of the Field Test in ELA/literacy and mathematics, it was highly recommended that all students complete the Practice Test and/or the Training Test for the Field Test. Each resource offered students a unique opportunity to experience the testing situation in a manner similar to what was experienced on the Field Test. Practice tests were grade-specific (3–8 and 11) and included a range of item types, grade-level content, and difficulty. There were approximately 30 items on a Practice Test in each content area for the Field Test. In addition, the Practice Tests included an initial set of accessibility features that were available to all students such as highlighting text, zooming in and out, marking items for review, and the digital notepad. A user guide provided direct guidance on accessing the Practice Tests, as well as frequently asked questions. The Training Tests were not grade specific and provided students and teachers with an opportunity to become familiar with the software and all interface features and functionalities that were used in the Smarter Balanced Field Test. They were available by grade bands (3–5, 6–8, and high school) and had six items in ELA/literacy and eight to nine in mathematics. The Training Tests did not include performance tasks. Table 15 summarizes the features of the Training and Practice Tests.

Table 15. Comparison of Features for the Training and Practice Tests.

Feature	Practice Test	Training Test
Purpose	Provide students the opportunity to experience a range of grade-specific item types (as well as performance tasks) similar in format and structure to the Smarter Balanced assessments.	Provide students with an opportunity to become familiar with the software and interface features used in the Smarter Balanced assessments.
Grade Level	Individual assessments at each grade <ul style="list-style-type: none"> • Grades 3–8 and 11 	Three assessments by grade band: <ul style="list-style-type: none"> • Grades 3–5 • Grades 6–8 • High School
Type of Items	Approximately 30 items in ELA/literacy and 30 items in mathematics per grade level One ELA/literacy and one mathematics performance task available per grade level	Approximately 14–15 items per grade band (6 in ELA/literacy and 8–9 in mathematics) No performance tasks Included new item types not present in the practice test (matching tables, table fill-in, & evidence-based selected response)
Available Embedded Universal Tools, Designated Supports, and Accommodations	All universal tools Most designated supports, including: <ul style="list-style-type: none"> • Color contrast • Masking • Text-to-speech items • Translations (glossary): Spanish Most accommodations, including: <ul style="list-style-type: none"> • American Sign Language for all mathematics items and ELA/literacy listening stimuli and items • Braille • Streamlining 	All universal tools All designated supports, including: <ul style="list-style-type: none"> • Color contrast • Masking • Text-to-speech items • Translated test directions: Spanish • Translations (glossary): Spanish, Arabic, Cantonese, Filipino, Korean, Mandarin, Punjabi, Russian, Ukrainian, Vietnamese • English glossary • Full translation: Spanish All accommodations, including: <ul style="list-style-type: none"> • American Sign Language for all mathematics items and ELA/literacy listening stimuli and items • Braille • Streamlining • Text-to-speech for reading passages in grades 6 to high school

General Field Test Administration Procedures. A brief overview of the general test administration rules are provided as well as information about various test tools and accommodations.

- CAT items (i.e., non-performance tasks) and performance tasks were presented in the Field Test administration as separate tests. All students participating in the Field Test, regardless of content area (ELA/literacy or mathematics) received CAT items, a classroom activity, and a performance task. In some cases, schools choose to administer both content areas to either the same or the different groups of students.
- The number of items in the CAT portion of the Field Test varied.
- The tests were not timed, so all time estimates were approximate. Students were allowed extra time if needed.
- The Field Test could be spread out over multiple days as needed.
- The Classroom Activity had to be completed prior to administration of the performance task.
- Students were not permitted to return to a test once it had been completed and submitted.
- Within each test, there may be several segments. A student was not permitted to return to a segment once it had been completed and submitted as complete.
- Students were instructed to answer all test items on a page before going to the next one. Some pages (i.e., screens) contained multiple test items. Students used a vertical scroll bar to view all items on a page.
- Students were required to answer all test items before submission for final processing.
- Students could mark items for review and use the Past/Marked drop-down list to return to those items.

The recommended order for test administration was to implement the CAT followed by the performance task assessment. For the performance task, the Classroom Activity was conducted, followed by the individually administered, online performance task. The recommendation was to administer the performance task portion of the assessment on a separate day from the CAT. For the performance tasks, an additional recommendation was that students might be best served by sequential, uninterrupted time that may exceed the time allotted in a student's regular classroom schedule.

During the CAT portion of the test, if a test was paused for more than 20 minutes the student was

- required to log back into the student interface;
- presented with the test page containing the test item(s) he or she was working on when the test was paused (if the page contains at least one unanswered item) or with the next test page (if all items on the previous test page were all answered); and,
- not permitted to review or change any previously answered items (with the exception of items on a page that contains at least one item that was not answered yet).

During the performance task portion of the test, there were no pause restrictions. If a test was paused for 20 minutes or more, the student could return to the section and continue typing his or her responses. Any highlighted text, notes on the digital notepad, or items marked for review were not saved when a test was paused. In the event of a technical issue (e.g., power outage or network failure), students were logged out and the test was automatically paused. Students needed to log in again when resuming the test.

As a security measure, students were automatically logged out of the test after 20 minutes of test inactivity. Activity was defined as selecting an answer or navigation option in the test (e.g., clicking [Next] or [Back] or using the Past/Marked Questions drop-down list to navigate to another item). Moving the mouse or clicking on an empty space on the screen was not considered test-taking activity. Before the system logged the student out of the test, a warning message was displayed on the screen. If the student did not click [Ok] within 30 seconds after the message appeared, he or she was logged out. Clicking [Ok] restarted the 20-minute inactivity timer.

A student's CAT administration remained active until the student completed and submitted the test or 45 calendar days elapsed after the student had initiated testing, whichever occurred sooner. A second recommendation was to minimize the amount of time between beginning and completing each test within a content area. Smarter Balanced suggest that students complete the CAT portion of the test within five days of starting the designated content area. The performance task was a separate test that remained active only for ten calendar days after the student began the performance task. However, Smarter Balanced recommended that students complete the PT within three days of starting.

Test Windows, and Testing Time. The Field Test was administered March 18–June 6, 2014. For the Field Test, schools were asked to select an anticipated testing window or were provided a testing window by their state. Smarter Balanced used this information to ensure that there was sufficient server capacity for all scheduled students to test.

Table 16 contains the estimated time required for most students to complete the Smarter Balanced Field Test based on the Pilot Test. Classroom Activities were designed to fit into a 30-minute window and will vary due to the complexity of the topic and individual student needs. These estimates did not account for any time needed to start computers, load secure browsers, and log in students. Note that the duration, timing, break/pause rules, and session recommendations varied in each content area and component.

Table 16. Expected Testing Times for Smarter Balanced Field Tests.

Content Area	Grades	CAT	Performance Task	Total	Classroom Activity	Overall Total
ELA/literacy	3–5	1:30	2:00	3:30	0:30	4:00
	6–8	1:30	2:00	3:30	0:30	4:00
	HS	2:00	2:00	4:00	0:30	4:30
Mathematics	3–5	1:30	1:00	2:30	0:30	3:00
	6–8	2:00	1:00	3:00	0:30	3:30
	HS	2:00	1:30	3:30	0:30	4:00
Combined	3–5	3:00	3:00	6:00	1:00	7:00
	6–8	3:30	3:00	6:30	1:00	7:30
	HS	4:00	3:30	7:30	1:00	8:30

Test Duration (Testing Time)

The Smarter Balanced tests were untimed. For test administration planning purposes, some indication of testing time is necessary. The delivery system was not able to give a per item student response time that could be accumulated accurately corresponding to test time for a student. A rough estimate was constructed that corresponded to test duration. Test duration was defined here as when the student entered the administration until the “submit” button was pressed that ended the assessment component. Since tests are administered as separate components, test duration is computed for ELA/literacy and mathematics and for both CAT and performance tasks. The resulting test durations are shown in Tables 17 to 20, which show the number of students by duration range and the corresponding cumulative percentage. Test duration is given in minutes. For ELA/literacy, most students required more than 90 minutes to complete the CAT and performance task components. The mathematics performance tasks were for the most part completed in 90 minutes. Note that longer test duration could result from test sessions that occurred over several days.

Table 17. Distribution of Test Duration in Minutes for the ELA/literacy CAT for the Item Pool Calibration Administration.

Range	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	No.	Percent	No.	Percent										
>= 90	49,830	100.0	58,815	100.0	58,502	100.0	62,155	100.0	54,138	100.0	56,252	100.0	71,053	100.0
85 - 90	1,729	40.7	2,156	37.1	2,117	32.6	2,393	32.3	2,528	40.2	2,587	41.0	6,094	69.6
80 - 85	1,913	38.7	2,305	34.8	2,279	30.2	2,770	29.6	2,722	37.4	2,970	38.2	7,160	67.0
75 - 80	2,195	36.4	2,547	32.3	2,352	27.5	2,902	26.6	3,162	34.4	3,223	35.1	8,766	64.0
70 - 75	2,449	33.8	2,724	29.6	2,534	24.8	3,074	23.5	3,357	30.9	3,452	31.7	10,187	60.2
65 - 70	2,782	30.9	2,843	26.7	2,638	21.9	3,047	20.1	3,694	27.2	3,642	28.1	12,059	55.9
60 - 65	3,069	27.6	3,082	23.7	2,794	18.9	3,097	16.8	3,793	23.1	3,815	24.3	13,574	50.7
55 - 60	3,198	23.9	3,309	20.4	2,634	15.6	2,861	13.4	3,575	18.9	3,746	20.3	14,645	44.9
50 - 55	3,355	20.1	3,275	16.8	2,663	12.6	2,483	10.3	3,401	15.0	3,456	16.4	15,254	38.6
45 - 50	3,291	16.1	3,260	13.3	2,393	9.5	2,136	7.6	2,877	11.2	3,181	12.7	15,270	32.1
40 - 45	3,208	12.2	3,010	9.8	2,077	6.8	1,762	5.3	2,438	8.1	2,765	9.4	14,582	25.6
35 - 40	2,686	8.4	2,509	6.6	1,613	4.4	1,222	3.3	1,885	5.4	2,297	6.5	13,049	19.4
30 - 35	2,015	5.2	1,743	3.9	1,041	2.5	831	2.0	1,296	3.3	1,644	4.1	10,836	13.8
25 - 30	1,226	2.8	1,061	2.1	633	1.3	498	1.1	824	1.8	1,144	2.4	8,156	9.2

Table 17. Distribution of Test Duration in Minutes for the ELA/literacy CAT for the Item Pool Calibration Administration continued.

Range	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	No.	Percent	No.	Percent										
20 - 25	672	1.3	556	0.9	331	0.6	316	0.6	479	0.9	623	1.2	6,162	5.7
15 - 20	351	0.5	250	0.3	149	0.2	157	0.2	278	0.4	346	0.5	4,519	3.0
10 - 15	81	0.1	64	0.1	33	0.0	31	0.0	75	0.1	93	0.1	1,941	1.1
5 - 10	9	0.0	13	0.0	7.0	0.0	11	0.0	14	0.0	34	0.0	636	0.3

Table 18. Distribution of Test Duration in Minutes for the ELA/literacy Performance Task for the Item Pool Calibration.

Range	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	No.	Percent	No.	Percent										
>= 90	45,051	100.0	54,619	100.0	49,075	100.0	48,593	100.0	42,699	100.0	48,834	100.0	41,013	100.0
85 - 90	1,170	44.8	1,680	40.5	1,693	42.2	2,034	45.8	1,752	51.4	2,039	47.4	3,479	75.2
80 - 85	1,513	43.4	1,880	38.6	1,967	40.2	2,235	43.5	2,080	49.4	2,360	45.2	4,120	73.1
75 - 80	1,697	41.6	2,151	36.6	2,196	37.9	2,448	41.0	2,392	47.1	2,491	42.7	4,816	70.6
70 - 75	1,937	39.5	2,343	34.3	2,458	35.3	2,760	38.3	2,700	44.3	2,764	40.0	5,613	67.7
65 - 70	2,180	37.1	2,603	31.7	2,539	32.4	2,994	35.2	2,903	41.3	3,082	37.0	6,347	64.3

Table 18. Distribution of Test Duration in Minutes for the ELA/literacy Performance Task for the Item Pool Calibration continued.

Range	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	No.	Percent	No.	Percent										
60 - 65	2,479	34.4	2,759	28.9	2,718	29.4	3,261	31.9	3,294	38.0	3,249	33.7	7,170	60.4
55 - 60	2,679	31.4	3,069	25.9	2,794	26.2	3,397	28.2	3,463	34.2	3,335	30.2	7,843	56.1
50 - 55	3,149	28.1	3,243	22.5	3,104	22.9	3,359	24.4	3,620	30.3	3,571	26.6	9,031	51.4
45 - 50	3,284	24.3	3,339	19.0	3,253	19.3	3,338	20.7	3,822	26.2	3,726	22.8	10,008	45.9
40 - 45	3,481	20.2	3,255	15.3	3,075	15.4	3,364	17.0	4,010	21.8	3,722	18.8	11,008	39.8
35 - 40	3,427	16.0	2,981	11.8	2,834	11.8	3,169	13.2	3,890	17.3	3,529	14.8	11,457	33.2
30 - 35	3,099	11.8	2,588	8.5	2,404	8.5	2,893	9.7	3,591	12.8	3,207	11.0	11,508	26.3
25 - 30	2,702	8.0	2,201	5.7	2,079	5.6	2,456	6.5	3,289	8.7	2,920	7.5	11,400	19.3
20 - 25	2,192	4.7	1,737	3.3	1,589	3.2	1,909	3.7	2,501	5.0	2,304	4.4	10,893	12.4
15 - 20	1,631	2.0	1,315	1.4	1,108	1.3	1,427	1.6	1,900	2.2	1,775	1.9	9,608	5.8

Table 19. Distribution of Test Duration in Minutes for the Mathematics CAT for the Item Pool Calibration.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent										
>= 90	24,813	100.0	36,112	100.0	43,479	100.0	34,510	100.0	37,517	100.0	38,256	100.0	37,933	100.0
85 - 90	1,085	73.2	1,984	66.5	2,367	59.5	2,180	70.1	1,575	66.4	1,504	65.5	2,249	83.1
80 - 85	1,423	72.0	2,577	64.7	3,021	57.2	2,495	68.2	1,869	65.0	1,883	64.1	2,919	82.1
75 - 80	1,784	70.5	2,887	62.3	3,537	54.4	3,152	66.1	2,429	63.3	2,363	62.4	3,867	80.8
70 - 75	2,292	68.5	3,549	59.6	4,047	51.1	3,938	63.4	2,887	61.2	2,983	60.3	5,009	79.1
65 - 70	2,936	66.1	4,291	56.3	4,808	47.4	4,811	59.9	3,526	58.6	3,564	57.6	6,225	76.9
60 - 65	3,612	62.9	5,233	52.3	5,461	42.9	5,636	55.8	4,121	55.4	4,328	54.4	7,917	74.1
55 - 60	4,547	59.0	6,100	47.5	6,171	37.8	6,469	50.9	5,062	51.7	4,891	50.5	9,887	70.6
50 - 55	5,632	54.1	7,203	41.8	6,621	32	7,432	45.3	5,808	47.2	5,746	46.1	12,076	66.2
45 - 50	6,924	48.0	7,888	35.1	7,049	25.8	8,103	38.9	6,958	42.0	6,513	40.9	15,206	60.8
40 - 45	7,849	40.5	8,112	27.8	6,540	19.3	8,356	31.9	7,739	35.8	7,224	35.0	18,451	54.0
35 - 40	8,437	32.0	7,678	20.3	5,679	13.2	8,517	24.6	8,366	28.8	8,269	28.5	21,056	45.8
30 - 35	8,041	22.9	6,046	13.2	4,016	7.9	7,526	17.2	8,001	21.4	7,905	21.0	22,147	36.4
25 - 30	6,516	14.2	4,332	7.5	2,476	4.1	5,906	10.7	6,920	14.2	6,511	13.9	20,762	26.6

Table 19. Distribution of Test Duration in Minutes for the Mathematics CAT for the Item Pool Calibration continued.

Range	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	No.	Percent	No.	Percent										
20 - 25	4,274	7.2	2,553	3.5	1,297	1.8	3,736	5.6	4,743	8.0	4,670	8.0	17,463	17.3
15 - 20	1,975	2.6	1,026	1.2	555	0.6	1,882	2.4	2,722	3.7	2,730	3.8	12,992	9.6
10 - 15	371	0.4	189	0.2	86.0	0.1	745	0.8	1,199	1.3	1,163	1.3	6,583	3.8
5 - 10	28	0.0	36	0.0	16	0.0	125	0.1	265	0.2	300	0.3	1,944	0.9

Table 20. Distribution of Test Duration in Minutes for the Mathematics Performance Tasks.

Range	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	No.	Percent	No.	Percent										
>= 90	5,514	100.0	7,396	100.0	11,447	100.0	9,067	100.0	7,722	100.0	8,935	100.0	9,188	100.0
85 - 90	452	93.9	704	93.0	1,030	89.0	671	90.3	207	91.5	302	90.4	558	94.5
80 - 85	571	93.4	896	92.3	1,159	88.0	876	89.5	287	91.3	472	90.1	733	94.1
75 - 80	779	92.7	1,121	91.5	1,626	86.9	1,135	88.6	393	91.0	625	89.6	1,031	93.7
70 - 75	991	91.9	1,481	90.4	2,152	85.3	1,346	87.4	473	90.5	902	88.9	1,408	93.1
65 - 70	1,227	90.8	1,941	89.0	2,730	83.2	1,799	85.9	676	90.0	1,142	87.9	1,888	92.2
60 - 65	1,787	89.4	2,645	87.2	3,548	80.6	2,317	84.0	917	89.3	1,597	86.7	2,557	91.1

Table 20. Distribution of Test Duration in Minutes for the Mathematics Performance Tasks continued.

Range	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
	No.	Percent	No.	Percent										
55 - 60	2,585	87.4	3,375	84.7	4,634	77.2	3,252	81.5	1,253	88.3	2,018	85.0	3,663	89.5
50 - 55	3,329	84.6	4,581	81.5	5,820	72.7	4,147	78.0	1,817	86.9	2,597	82.8	4,962	87.3
45 - 50	4,550	80.9	6,323	77.2	7,340	67.1	5,412	73.6	2,766	84.9	3,799	80.0	7,091	84.4
40 - 45	6,014	75.8	8,196	71.2	8,844	60.0	7,254	67.8	4,347	81.9	5,306	75.9	9,847	80.1
35 - 40	8,118	69.1	10,273	63.5	10,675	51.5	9,134	60.0	6,752	77.1	7,414	70.2	13,388	74.2
30 - 35	10,214	60.1	12,780	53.8	11,642	41.3	11,036	50.2	9,837	69.7	9,942	62.2	17,129	66.1
25 - 30	12,095	48.8	14,615	41.7	11,900	30.0	11,840	38.3	13,081	58.9	12,189	51.6	21,009	55.8
20 - 25	13,288	35.4	14,041	27.9	10,098	18.6	11,464	25.6	15,711	44.5	13,651	38.5	24,165	43.1
15 - 20	11,387	20.6	10,484	14.6	6,359	8.9	8,095	13.3	15,031	27.2	13,025	23.8	25,152	28.6
10 - 15	7,183	8.0	5,011	4.7	2,838	2.7	4,312	4.6	9,767	10.7	9,083	9.8	22,352	13.5

Universal Tools, Designated Supports, and Accommodations

The Smarter Balanced Assessment Consortium’s *Usability, Accessibility, and Accommodations Guidelines* are intended for school-level personnel and decision-making teams, including Individualized Education Program (IEP) teams, as they prepare for and implement the Smarter Balanced assessments. The Guidelines provide information for classroom teachers, English development educators, special education teachers, and related services personnel to use in selecting and administering universal tools, designated supports, and accommodations for those students requiring them. The Guidelines are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.

The Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* apply to all students. They emphasize an individualized approach to the implementation of assessment practices for those students who have diverse needs and participate in large-scale content assessments. The Guidelines focus on universal tools, designated supports, and accommodations for the Smarter Balanced content assessments of English language arts/literacy and mathematics. At the same time, the Guidelines support important instructional decisions about accessibility and accommodations for students who participate in the Smarter Balanced assessments. The Guidelines recognize the critical connection between accessibility and accommodations in instruction and accessibility and accommodations during assessment. The Field Test and Training Tests contained embedded universal tools, designated supports, and accommodations and are defined in Table 21. Embedded resources are those that are part of the computer administration system, whereas non-embedded resources are provided outside of that system. Chapter 5 on Test Fairness presents a more comprehensive discussion of these issues.

Table 21. Definitions for Universal Tools, Designated Supports, and Accommodations.

Type	Definition
Universal Tools	Access features of the assessment that either are provided as digitally delivered components of the test administration system or separate from it. Universal tools are available to all students based on student preference and selection.
Designated Supports	Access features of the assessment available for use by any student for whom the need has been indicated by an educator (or team of educators working with the parent/guardian and student). They either are provided as digitally delivered components of the test administration system or separate from it.
Accommodations	Accommodations are changes in procedures or materials that increase equitable access during the Smarter Balanced assessments. Assessment accommodations generate valid assessment results for students who need them; they allow these students to show what they know and can do. Accommodations are available for students with documented IEPs or Section 504 Plans. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standards, or intended outcome of the assessment.

Test Security

The test environment refers to all aspects of the testing situation. The test environment includes what a student can see, hear, or access (including access via technology). Requirements of a secure test environment include, but are not limited to, the following:

- Providing a quiet environment, void of talking or other distractions that might interfere with a student's ability to concentrate or might compromise the testing situation.
- Actively supervising students to prevent access to unauthorized electronic devices that link to outside information, communication among students, and photographing or otherwise copying test content.
- Removing information displayed on bulletin boards, chalkboards or dry-erase boards, or charts (e.g., wall charts that contain literary definitions, maps, mathematics formulas, etc.) that might assist students in answering questions must be removed.
- Seating students so there is enough space between them to minimize opportunities to view each other's work, or providing students with tabletop partitions.
- Allowing students access to only the allowable resources identified by Smarter Balanced specific to the assessment (or that portion of an assessment).
- Allowing only students who are testing to observe assessment items. Students who are not being assessed or unauthorized staff should be removed from the testing environment.
- Administering the Smarter Balanced Field Test only through the Student Interface via a secure browser.

Item security rules included, but were not limited to, the following:

- Unless assigned as an accommodation, no copies of the test items, stimuli, reading passages, performance task materials, or writing prompts could be made or otherwise retained. This rule included any digital, electronic, or manual device used to record or retain item information.
- Descriptions of test items, stimuli, printed reading passages, or writing prompts must not be retained, discussed, or released to anyone. All printed test items, stimuli, and reading passages must be securely shredded immediately following a test session.
- Test items, stimuli, reading passages, or writing prompts must never be sent by e-mail or fax or replicated/displayed electronically.
- Secure test items, stimuli, reading passages, or writing prompts must not be used for instruction.
- No review, discussion, or analysis of test items, stimuli, reading passages, or writing prompts were allowed at any time by students, staff, or teaching assistants, including before, during, or between sections of the test. Student interaction with test content during a test was limited to what was dictated for the purpose of a performance task that was standardized.
- No form or type of answer key may be developed for test items.

Test security incidents, such as improprieties, irregularities, and breaches, were behaviors prohibited during test administration, either because they lent a student a potentially unfair advantage or because they compromised the secure administration of the assessment. Whether intentional or by accident, failure to comply with security rules, either by staff or students, constituted a test security

incident. Improprieties, irregularities, and breaches were reported in accordance with each severity level. Definitions of three types of test security incidents are given in Table 22.

Table 22. Definitions for Three Levels of Test Security Incidents.

Type	Definition
Impropriety	An unusual circumstance that has a low impact on the individual or group which has a low risk of potentially affecting student performance on the test, test security, or test validity. These circumstances can be corrected and contained at the local level. An example of an impropriety might include posting a practice item to a social media site by a student.
Irregularity	An unusual circumstance that affects an individual or group of students who are testing and may potentially influence student performance on the test, test security, or test validity. These circumstances can be corrected and contained at the local level, but submitted in the online system for resolution of the appeal for testing impact.
Breach	An event that poses a threat to the validity of the test. These circumstances have external implications for the Consortium and may result in a decision to remove the test item(s) from the available secure bank. A breach incident must be reported immediately.

Test monitors were instructed to be vigilant before, during, and after testing for any situations that could lead to or be an impropriety, irregularity, or breach. The following instructions were given:

- Actively supervise students throughout the test session to ensure that students do not access unauthorized electronic devices, such as cell phones, or other unauthorized resources or tools at any time during testing.
- Make sure students clear their desks of and put away all books, backpacks, purses, cell phones, electronic devices of any kind, as well as other materials not explicitly permitted for the test.
- Make sure the physical conditions in the testing room meet the criteria for a secure test environment. Students should be seated so there is enough space between them to minimize opportunities to view another student’s work.
- Students who are not being tested and unauthorized staff must not be in the room where a test is being administered. Determine where to send these students during testing and prepare appropriate assignments for them as needed.
- Make sure no instructional materials directly related to the content of the tests are visible to students, including posters or wall charts.
- States should ensure that specific guidance is provided for districts that have minimal personnel and may experience potential conflicts of interest in the identification, investigation, and/or reporting of test security incidents.

References

- Folk, V. G. & Smith, R. L. (2002). Models for delivery of CBTs. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.). *Computer-based Testing: Building the Foundation for Future Assessments* (pp. 41-66). Mahwah, New Jersey: Lawrence Erlbaum.
- Frankel, M. (1983). Sampling theory. In *Handbook of Survey Research*, Wright, Anderson, & Rossi (Eds.). New York: Academic Press.
- Gibson, W. M. & Weiner, J. A. (1998). Generating Random Parallel Test Forms Using CTT in a Computer-based Environment. *Journal of Educational Measurement*, 35, 297-310.
- Hetter, R. D. & Sympson, J. B., (1997). Item Exposure Control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.). *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 141-144). Washington, DC: American Psychological Association.
- Khatri, N., Reve, A. L., & Kane, M. B. (1998). *Principles and Practices of Performance Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Quality Education Data. School Year 2011-2012. MCH. Sweet Springs: MO.

Chapter 8 Field Test Data Steps and Classical Test Analyses

Introduction

There were two steps or phases for the Smarter Balanced Field Test. The purpose of the first step of the Field Test was a) to include a robust set of items to establish the horizontal and vertical scales using a representative student sample, and b) to perform the achievement level setting conducted in the fall of 2014. In the vertical scaling sample, selected computer adaptive test (CAT) items and performance tasks (PTs) were administered to students at adjacent lower grades to permit vertical linking (i.e., the off-grade administration of items/tasks). A more manageable number of items and students was needed to conduct the vertical scaling (i.e., standard setting) within the short time confines of the Field Test in order to permit the 2014 achievement level setting. The purpose of the second step of the Field Test was to calibrate a large, robust pool of items onto the scale established in the vertical scaling analysis. All the items in the vertical scaling were also administered on-grade to students in the calibration study for linking purposes. In the calibration step, students were administered items/tasks that were targeted for that grade (i.e., on-grade administration only). This horizontal calibration (item-pool calibration) and linking resulted in the final entire parameter estimates for the Smarter Balanced item pool. The test windows (i.e., administration) of the vertical scaling and item pool calibration steps overlapped in the spring of 2014. The primary focus here is to demonstrate the properties of the Smarter Balanced items pool at the conclusion of the item-pool calibration step that reflected the items available for operational administrations. The results presented are for the item pool calibrations step unless they are also explicitly labeled as vertical scaling outcomes.

In accordance with the Field Test design, students were administered a CAT component, which was intended to conform to the Smarter Balanced test blueprint, and a selected performance task. The CAT and performance task components work in conjunction to fulfill the content requirements for the test blueprint. They consist of a variety of different item types, some of which were machine scored. All single-selection selected-response (SR) items had three to five answer choices. Multiple-selection selected-response (MSR) items provided five to eight answer choices. The performance task items had scores ranging from zero to a maximum of four score points. In the case of ELA/literacy (ELA), the extended student-writing sample was scored for three dimensions of writing (purpose/focus/organization, evidence/elaboration, and conventions). A performance task was expected to have approximately 4 to 6 scorable units (i.e., items) yielding approximately 12 to 15 score points in total. For example in grade 6 mathematics, a task could have six items with maximum score levels corresponding to 1, 2, 2, 2, 2, 3 that pertained to three short answer items and three equation items. In grade 11 ELA, there were 11 maximum raw score points associated with a task; six points for the extended writing response, two short answer items and a matching item.

The CAT items were administered in the context of linear-on-the-fly testing (LOFT), in which the test content that was sampled conformed to the Smarter Balanced test blueprint. Unlike a fixed or linear test form (e.g., a paper-and-pencil administration), there are no intact test forms common across students which are necessary to compute classical test reliability or the overall “number correct” common across substantial numbers of students. The primary advantage of the LOFT administration is its efficiency in test delivery since test forms can be constructed dynamically for each student that conform to the CAT test blueprint. Given the size of the item pool in a grade and content area, it would have been difficult to construct the necessary number of linear, blueprint conforming test forms without a LOFT administration.

Decisions concerning the data steps are included here since they had important implications for the resulting item quality and composition of the resulting item pools in ELA/literacy and mathematics. Further information can be found on vertical and horizontal scaling in Chapter 9 and on test design

in Chapter 4. An explanation of the differential item functioning (DIF) methods used is given in Chapter 6 on the Pilot Analysis.

Data Inclusion/Exclusion Rules for Items and Students

The first step was to create a sparse data matrix for analysis reflecting item scores as well as missing information by design. Each row of the matrix was a student response vector where the columns were the available items. For a given grade, the dimension of this sparse matrix is the total number of students times the total number of unique items (i.e., scorable units). Many of the cells of this matrix represent items not administered to students by design. This “missing” information was indicated in the sparse matrix as “not presented” items, which is the typical practice when multiple test forms exist. Smarter Balanced defined condition codes for various sorts of invalid responses to polytomous items that consisted of the following designations: B (Blank), U (Insufficient), F (Non-scorable language), T (Off topic), and M (Off purpose). These condition codes were ultimately resolved as scores of zero in the data matrix used in calibration. This data matrix for each grade and content area was then the focus of the subsequent analyses.

The inclusion/exclusion rules were applied prior to the classical test analysis and IRT calibrations in order to ensure that the best possible statistical outcomes resulted. Inclusion and exclusion logic for both items and students were also implemented using IRT statistics. This IRT item exclusion might include issues like non-convergence during parameter estimation or very low IRT discrimination values. They are included in this data step to help avoid confusion concerning the final number of Field Test students and items. Since there were a limited number of performance tasks, extra effort was made to preserve the associated items within a given task. First, the student exclusion rules are presented. These are followed by rules applied to both the CAT selected-response (SR), constructed-response (CR) items, and the performance tasks (PT).

Student Exclusions. The following rules were implemented in the vertical scaling step. For the item calibration step, short tests (less than 9 items) were included in the analysis.

1. A record was excluded if a student was deemed to not have made a reasonable attempt on the Smarter Balanced Field Test. Students were eliminated if their response time (i.e., test duration) was very short, which likely indicated that a reasonable effort was not made or some other anomaly occurred. Test duration was defined as the time when the student entered the test administration until the test was completed using the “submit” button.
 - a. A student record was excluded if the full-length CAT test event was completed in less than 15 minutes.
 - b. Note that in California, students took shortened ELA/literacy and mathematics CAT components expected to be approximately 25 items in each content area as opposed to 50 items in a single content area. Half-length CAT events administered to California students were eliminated if the test duration was less than eight minutes.
 - c. If the ELA/literacy performance task was completed in less than 15 minutes or the mathematics performance task was completed in less than ten minutes, the student’s score was excluded.
2. All student records with a zero on all item scores were excluded.
3. Students with scored responses to less than nine items were excluded.

The impact of applying these student exclusion rules is shown in Table 1. The number of students excluded was relatively negligible except in high school ELA/literacy and mathematics. Table 1 reflects the number of students that have valid performance task scores (all students have CAT scores), which may be lower than the student counts in other tables.

Table 1. Summary of Students Excluded and Resulting Sample Size.

Vertical Scaling			Item Pool Calibration			
Grade	No. Students	No. Excluded	No. Valid	No. Students	No. Excluded	No. Valid
ELA						
3	23,788	585	23,203	85,889	1,830	84,059
4	36,271	572	35,699	94,915	1,393	93,522
5	32,220	614	31,606	88,293	1,503	86,790
6	32,229	681	31,548	93,536	1,790	91,746
7	32,005	1,126	30,879	93,431	2,895	90,536
8	37,129	1,213	35,916	98,433	3,163	95,270
HS	57,608	7,073	50,535	261,405	27,462	233,943
Mathematics						
3	25,671	848	24,823	95,143	2,604	92,539
4	39,522	595	38,927	109,441	1,645	107,796
5	42,818	433	42,385	108,412	1,186	107,226
6	34,014	775	33,239	117,691	2,172	115,519
7	32,176	1,885	30,291	117,049	5,342	111,707
8	38,856	2,135	36,721	116,459	5,656	110,803
HS	56,658	7,375	49,283	262,111	37,425	224,686

Item Exclusions and Data Steps. Item quality was inspected using frequency distributions and classical item statistics prior to conducting the IRT calibration. After consultation with Smarter Balanced, poor-quality items were excluded by using either statistical or judgmental rules. Items were excluded based on the following rules and guidelines:

1. all selected-response items with rounded item difficulty at or below 0.10;
2. CAT (polytomous, non-selected-response items) with a rounded item difficulty at or below 0.02,

3. performance tasks (performance task non-selected-response items) with a rounded average item difficulty at or below 0.01;
4. CAT polytomous items with any score categories having 10 or fewer observations;
5. all dichotomous items that have 30 or fewer observations obtaining a score point of 1;
6. items having fewer than 500 observations; and
7. selected-response items incorporating the combined psychometric staff evaluation of the item-total correlation and empirical item plots.

For constructed-response items, score categories with fewer than 10 students at on-grade level were collapsed with neighboring categories in both on-grade and off-grade data sets. If the category that needed to be collapsed was a middle category, it was collapsed with the category with smaller number of observations.

Using IRT-derived rules, additional item exclusions were performed to ensure the most reasonable item and student estimates would result. Items were excluded based on the following IRT-derived rules.

- a. Non-convergence during Marginal Maximum Likelihood (MML) estimation
- b. Discrimination parameter estimates below 0.10
- c. The quality of additional items were evaluated based on
 - i. Selecting outliers by rank ordering the IRT discrimination parameters and classical item-total correlations
 - ii. Selecting outliers by rank ordering the IRT difficulty parameter and observed p-value
 - iii. Identifying unreasonably high chi-square by rank ordering sample size and chi-square
 - iv. Identifying large standard errors for IRT discrimination and/or difficulty parameters
 - v. Item characteristic curves with poor fit between observed and expected performance.

Tables 2 and 3 show a summary of the total item inventory—the items lost strictly to content and scoring decisions, items analysis, and IRT exclusions. They show the number of items that survived (Final Pool) after all the exclusion rules were applied in ELA and mathematics for the item pool. The subsequent IRT exclusions were included here for completeness. These tables list the original inventory of all items developed and the number of items not used or otherwise scored for content reasons. The “sample size” column shows the number of items eliminated for small sample size or fewer than 10 observations in a score category and applying the various exclusion rules. No items were dropped from the calibration analysis because of DIF. A large number of items were precluded from IRT analysis due to small sample size in high school ELA and mathematics based on classical test analysis. The final set of items was used to derive the classical item and test statistics of record and those entering into the IRT scaling labeled under the “Resulting Pool” column. A significant number of items were not calibrated due to an insufficient sample size in high school. These items can be piloted and scaled in subsequent operational administrations.

Table 2. Summary of ELA Item Exclusions (Item Pool Calibration) by Type.

Grade	Initial	Content	Small Sample Size		Poor Item Statistics		Final
	Pool	Issues	(50,300)	(300,500)	Classical	IRT	Pool
3	1,045	30	13	18	69	19	896
4	965	17	13	19	38	22	856
5	975	23	31	14	65	19	823
6	984	23	19	11	60	22	849
7	1,033	27	20	11	77	23	875
8	1,010	20	17	23	95	19	836
HS	3,371	61	272	386	248	33	2,371

Table 3. Summary of Mathematics Item Exclusions (Item Pool Calibration) by Type.

Grade	Initial	Content	Small Sample Size		Poor Item Statistics		Final
	Pool	Issues	(50,300)	(300,500)	Classical	IRT	Pool
3	1,163	1	-	-	44	4	1,114
4	1,207	9	-	-	56	12	1,130
5	1,108	2	-	-	45	18	1,043
6	1,115	8	-	-	80	9	1,018
7	1,037	5	-	-	76	14	942
8	1,036	9	-	-	103	30	894
HS	3,386	75	25	772	433	55	2,026

Item Pool Composition (Vertical Scaling and Item Pool Calibration Steps)

Since the vertical scaling item sets were used to establish the Smarter Balanced scales, it is important to delineate the composition of the items types. Tables 4 and 5 classified items by purpose and type for the vertical scaling. Items were targeted for on-grade or off-grade administration for the vertical scaling. The mixture of items types included both selected- and constructed-response items. Constructed-response items could be dichotomously (right/wrong) or polytomously (with provision for partial credit) scored. The item counts reflect both CAT and

performance task items. NAEP and PISA items were also given in selected grades. These tables also show the number of items that remained for vertical scaling after all the item exclusions were applied. Table 6 shows the distributions of the on-grade items by claim and item type. The claims in ELA pertain to reading, writing, listening/speaking, and research, respectively. In mathematics, the claims pertain to concepts/processes, problem solving, communicating reasoning, and modeling/data analysis. Table 7 shows the same types of information for the vertical linking items. Table 8 shows the distribution for all items by claim and type from the calibration item pool for ELA and mathematics. All items contained in the calibration step consisted of all available items in the pool targeted in the Field Test for on-grade administration; this was inclusive of the vertical scaling items. The readministration of “on-grade vertical scaling” items was necessary to link the item pool calibration items onto the scale.

Table 4. Summary of ELA Vertical Scaling Items by Purpose and Type.

Item Purpose	Response Type	Score Type	Grade						
			3	4	5	6	7	8	HS
On-Grade	Selected-response		115	92	95	81	77	91	142
	Other	Dichotomous	110	119	122	114	122	113	216
		Polytomous	36	31	39	37	39	39	52
Off-Grade Vertical Linking	Selected-response			57	53	46	27	38	39
	Other	Dichotomous		45	62	63	58	62	58
		Polytomous		18	18	22	22	23	10
NAEP	Selected-response			22				20	12
	Other	Dichotomous		2				2	4
		Polytomous		4				8	11
PISA	Selected-response								17
	Other	Dichotomous							12
		Polytomous							4

Table 5. Summary of Vertical Scale Mathematics Items by Purpose and Type.

Item Purpose	Response Type	Score Type	Grade						
			3	4	5	6	7	8	HS
On-Grade	Selected-response		48	65	78	21	39	41	66
	Other	Dichotomous	221	212	175	174	185	159	203
		Polytomous	35	29	53	27	15	30	50
Off-Grade Vertical Linking	Selected-response			11	12	31	9	7	18
	Other	Dichotomous		76	71	56	55	60	56
		Polytomous		17	12	15	7	6	7
NAEP	Selected-response			20				19	18
	Other	Dichotomous		2				6	4
		Polytomous		8				8	6
PISA	Selected-response								19
	Other	Dichotomous							44
		Polytomous							11

Table 6. Number of On-grade Vertical Scaling Items by Content Area and Characteristics.

Item Type	Grade						
	3	4	5	6	7	8	HS
ELA							
Total	261	242	256	232	238	243	410
Selected-response	115	92	95	81	77	91	142
Dichotomous	110	119	122	114	122	113	216
Polytomous	36	31	39	37	39	39	52
Claim 1	94	72	91	71	75	83	181
Claim 2	70	67	67	67	70	66	126
Claim 3	50	51	46	45	46	49	39
Claim 4	47	52	52	49	47	45	64
Mathematics							
Total	304	306	306	222	239	230	319
Selected-response	48	65	78	21	39	41	66
Dichotomous	221	212	175	174	185	159	203
Polytomous	35	29	53	27	15	30	50
Claim 1	184	182	182	107	134	130	191
Claim 2	17	17	17	20	10	15	22
Claim 3	47	51	49	40	38	35	44
Claim 4	19	23	20	19	21	17	33
Unclassified	37	33	38	36	36	33	29

Table 7. Number of Off-grade Vertical Linking Items by Content Area and Characteristics.

Item Type	Grade					
	4	5	6	7	8	HS
ELA						
Total	120	133	131	107	123	107
Selected-response	57	53	46	27	38	39
Dichotomous	45	62	63	58	62	58
Polytomous	18	18	22	22	23	10
Claim 1	40	54	53	41	49	48
Claim 2	34	32	31	29	34	25
Claim 3	26	25	25	18	24	21
Claim 4	20	22	22	19	16	13
Mathematics						
Total	104	95	102	71	73	81
Selected-response	11	12	31	9	7	18
Dichotomous	76	71	56	55	60	56
Polytomous	17	12	15	7	6	7
Claim 1	58	55	60	28	36	46
Claim 2	4	3	4	4	3	5
Claim 3	16	18	13	15	9	12
Claim 4	7	6	7	6	7	6
Unclassified	19	13	18	18	18	12

Table 8. Number of On-grade Calibration Items by Content Area and Characteristics.

Item Type	Grade						
	3	4	5	6	7	8	HS
ELA							
Total	896	856	823	849	875	836	2371
Selected-response	386	336	286	300	280	301	875
Dichotomous	381	376	379	413	437	372	1216
Polytomous	129	144	158	136	158	163	280
Claim 1	317	259	265	274	299	258	867
Claim 2	243	248	241	257	262	241	729
Claim 3	163	157	142	147	152	174	383
Claim 4	173	192	175	171	162	163	392
Mathematics							
Total	1114	1130	1043	1018	942	894	2026
Selected-response	166	189	239	101	109	161	530
Dichotomous	815	789	633	792	743	610	1247
Polytomous	133	152	171	125	90	123	249
Claim 1	672	677	613	576	519	493	1123
Claim 2	55	68	55	77	71	59	147
Claim 3	166	145	168	132	120	134	433
Claim 4	68	77	84	68	67	64	185
Unclassified	153	163	123	165	165	144	138

Before presenting the classical results, a discussion of processing of ELA essay scores for performance tasks is necessary. For performance tasks in ELA/literacy, students were administered a writing task (i.e., extended writing response) that is scored on three dimensions of writing that correspond to organization (0-4 points), elaboration (0-4 points), and conventions (0-2 points). That is, three separate scores (i.e., scorable units) are obtained for a single student writing sample. The correlations between the dimensions of organization and elaboration exceeded 0.95 in many instances. This high degree of dependence precluded them from being calibrated as separate items due to very high local item dependence. As a result, the two writing dimensions for organization and elaboration were averaged and rounded up if necessary. This resulted in a single 0 to 4 point score for these two dimensions along with the original conventions score (0-2) for the long writing task. These three ELA/literacy raw scores from the long writing task were then used in the IRT scaling.

Classical Item and Test Analysis

Classical (traditional or observed) item and test statistics are presented here for both items and tests. Tests are defined as the collection of items administered to students for the CAT using Linear-on-the-Fly-Testing (LOFT) administration combined with a performance task. This test definition

pertains to the items remaining after all the item exclusions were applied. Both CAT and performance task components were needed to fulfill the test blueprint. Each statistic provides some key information about the quality of each item or test based on empirical data from the Smarter Balanced assessments. Classical measures include statistics such as item difficulty, item-test correlations, and test statistics (e.g., reliability). Other descriptive measures, such as the percentage of students at each response option or score level, were used to evaluate item functioning but were not reported here. Classical test analyses were conducted in part to gain information about the quality of items, such as the following:

- Based on item difficulty, is the item appropriate for testing at a given grade level?
- How effective is the item in distinguishing students with high and low ability? Did higher ability students perform better on the item than lower ability students?
- For selected-response (SR) items, is the key the only correct choice? Are all item distractors wrong? Are distractors constructed in a way that is more attractive to low ability students compared with high ability ones? Are high ability students more likely to choose the key than the distractors?
- For constructed-response (CR) items, do high ability students tend to score in upper score categories and less able students in lower ones?
- For an item that is administered in multiple grades for the purpose of vertical scaling, do students in a higher grade level tend to perform better on the item than students in a lower grade level?
- Does the item show DIF? In other words, does the item tend to be especially difficult for a specified group of students with comparable levels of ability?
- Are scores sufficiently reliable for the intended purposes?

To address these properties, the analyses include several components: item difficulty, item discrimination, item response distribution for CR items, differential item functioning and score reliability. In the context of the Field Test, these statistics also provided information that was used to exclude poorly functioning items prior to the IRT calibration step or to inform future, item writing activities on the part of content developers.

As mentioned at the outset, the presentation of statistics is more difficult to summarize in some respects since no fixed forms containing a set number of items exist in the Field Test. There were many potential combinations of CAT test forms presented to students due to the LOFT administration and the sheer number of items in the pool in a given grade and content area. Several types of classical analysis rely on the provision of a criterion variable for computing item-test correlations or differential item functioning that is typically defined as the total raw score. Since there are many variations in the CAT items presented to students due to the LOFT administration along with performance tasks, defining a common test criterion was not possible. To circumvent these problems and provide the best available criterion score, the student ability (i.e., theta estimate) was used. As a result, item-test correlations and DIF depended on incorporating IRT ability as the criterion score. Chapter 9 on the IRT analysis provides a description of the methods used to compute theta. This modified classical test analysis was conducted to obtain additional evidence concerning item and test properties, item pool characteristics, and eventually performing the data review of items by content developers for operational administrations.

Item Difficulty. The percent of maximum possible score is computed for each item as an indicator of item difficulty with a range of 0.0 to 1.0. A relatively higher value indicates an easier item. An item difficulty of 1.0 indicates that all students received a perfect score on the item. An average item

score of 0.0 for an item indicates that no students answered the item correctly or only received partial credit for the item in the case of polytomous or CR items.

For dichotomous items and SR items, the percent of maximum possible score is simply equivalent to the percentage of students who answered the item correctly. The formula for p -value for selected response is

$$p\text{-value}_{SR} = \frac{\sum X_{ic}}{N_i},$$

where X_{ic} is the number of students that answered item i correctly, and N_i is the total number of students observed for item i .

A polytomous item is an item that is scored with more than two ordered categories, such as the scores from the ELA/literacy performance task essay. For polytomous items (i.e., constructed-response), the p -value is defined as

$$p\text{-value}_{CR} = \frac{\sum X_{ij}}{N_i \times \text{Max}(k)},$$

where X_{ij} is the score assigned for a given constructed-response item and k are the score levels associated with the item. Another interpretation is that item difficulty for constructed-response items is the mean score for the item divided by the maximum score. For example, a polytomous item had scores ranging from a low score of zero to three as the maximum and the observed mean score was 2.1. The observed percent of maximum can also be calculated as $2.1/3 = 0.70$, or 70 percent, of the maximum score was achieved by students on this hypothetical constructed-response item. In the case of a selected-response item (i.e., multiple-choice), the maximum score is one by definition and defaults to the selected-response p -value.

A wide item difficulty range is needed to measure student ability that can vary greatly particularly for operational administrations of the adaptive test. Very easy or difficult items require additional review to ensure that the items are valid and are grade appropriate. Note that some items served as anchor items in vertical scaling. These items are administered across multiple grade levels and therefore can have several sets of grade-level-specific classical item statistics. For vertical scaling, item difficulty across different grade levels was assessed to evaluate if students in the upper grade level generally performed better in comparison with a lower grade level.

Item Discrimination. Item discrimination evaluates how well an item distinguishes between low and high ability students. In classical (non-IRT) item analysis it is the correlation between the item score and total test score. The expectation is that high ability students will outperform low ability students on an item. The item discrimination statistic is calculated as the correlation coefficient between the item score and criterion score (i.e., IRT ability estimate). A relatively high item-total correlation coefficient value is desired, as it indicates that students with higher scores on the overall test tended to perform better. In general, item-total correlation ranges from -1.0 (for a perfect negative relationship) to 1.0 (for a perfect positive relationship). However, a negative item-total correlation typically signifies a problem with the item, as the higher ability students generally are getting the item wrong or a low score and the lower ability students are getting the item right or are assigned a higher score level.

Some coefficients used in computing item-total correlations are the point-biserial and polyserial correlation coefficient. The point-biserial correlation is used for dichotomous items and polyserial correlation used for polytomous items. The point-biserial correlation coefficient is a special case of

the Pearson correlation coefficient used for dichotomous items. The point-biserial correlation is computed using

$$r_{ptbis} = \frac{(\mu_+ - \mu_-)}{\sigma_{tot}} \sqrt{pq}$$

where μ_+ is the mean criterion score of examinees answering the item correctly, μ_- is the mean criterion score of the examinees answering the item incorrectly, σ_{tot} is the standard deviation of the criterion score, p is the proportion of examinees answering the item correctly, and q equals $(1 - p)$.

The polyserial correlation measures the relationship between a polytomous item and the criterion score. Polyserial correlations are based on a polyserial regression model (Olsson, 1979; Drasgow, 1988), which assumes that performance on an item is determined by the examinee's position on an underlying latent variable that is normally distributed at a given criterion score level. Based on this approach, the polyserial correlation can be estimated as

$$r_{polyreg} = \frac{\beta \sigma_{tot}}{\sqrt{\beta^2 \sigma_{tot}^2 + 1}}$$

in which β is a series of parameters estimated using maximum likelihood and σ_{tot} is the standard deviation of the criterion score. The biserial correlation could have been chosen for dichotomous items but the point-biserial and its interpretation is more familiar to many users.

Distractor Analysis. For each selected-response item, distractor analyses were conducted. The quality of distractors is an important component of an item's overall quality. Distractors should be clearly incorrect, but at the same time plausible and attractive to lower ability students. The following distractor analyses are conducted to evaluate the quality of distractors.

- The percentage of students at each response option is calculated. For the key (i.e., the correct answer), this percentage is the item difficulty value. If the percentage of students who selected a distractor is greater than the percentage of students who selected a key, the item should be examined to determine if it has been incorrectly keyed or double-keyed.
- The point-biserial correlation is calculated for each response option. While the key should have a positive point-biserial correlation with the criterion score, the distractors should exhibit negative point-biserial correlations (i.e., lower ability students would likely choose the distractors, while the higher ability students would not).
- The average ability level (measured by criterion score) is calculated for students at each response option. Students choosing the key should be of higher ability levels than students choosing distractors.
- The percentage of high ability students at each response option is calculated. High ability students were defined as the top 20 percent of students in the ability distribution (grade and content area). If the percentage of high ability students who selected a distractor is greater than the percentage of high ability students who selected a key, the item should be examined further.

For each constructed-response item, the following statistics are evaluated.

- The percentage of students at each score level is calculated. If there were very few students at certain score levels, this might suggest that some score categories need to be collapsed or that the scoring rubric needs adjustment.

- The average ability level is calculated for students at each score level. Students at a higher score level on this item should be of higher ability levels (i.e., having higher average ability estimates) than students at a lower score level on this item.
- The item-test correlation is computed using the polychoric correlation.

Reliability Analyses. The variance in the distributions of test scores, essentially the differences among individuals, is partly due to real differences in the knowledge, skills, or ability being tested (true variance) and partly due to random errors in the measurement process (error variance). Score reliability is an estimate of the proportion of the total variance that is true variance. The estimates of reliability used here are internal-consistency measures. The formula for the internal consistency reliability, as measured by Cronbach's coefficient alpha (Cronbach, 1951), is

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_x^2} \right],$$

where n is the number of items, σ_i^2 is the variance of scores on the i^{th} item, and σ_x^2 is the variance of the total score (sum of scores on the individual items).

The standard error of measurement (SEM) provides a measure of score instability in the score metric. The formula for computing the SEM is

$$\sigma_e = \sigma_x \sqrt{1 - \alpha},$$

where reliability α is Cronbach's alpha estimated above, and σ_x is the standard deviation of the scores. The SEM can be used to determine the confidence interval (CI) that captures an examinee's true score.

Item Flagging Criteria for Content Data Review. Flagging is used to identify certain statistical characteristics of items that indicate poor functioning. For example if an item is very difficult for a grade level or has very low discrimination, its properties should be reviewed further by content developers before selecting it for inclusion in the item pool. Content developers reviewed items after the Field Test analysis in conjunction with item statistics. Items with poor classical statistics were designated using various types of flags. These flags were used in conjunction with substantive attention to the item content to determine if a problem exists and if any corrective action was required. At a minimum, flagged items underwent additional scrutiny for content appropriateness, bias and sensitivity, and overall statistical performance relative to expectations. Any item with substantial changes was returned to the item bank for further pretesting prior to operational use. Items that were functioning very poorly could either be excluded from further use or be rewritten to improve their performance as new items. Table 9 lists the flagging definitions for selected- and constructed-response items. Note that items were also flagged for differential item functioning (DIF) presented later. If an item was judged to have potential flaws after reviewing the item content, it was flagged for further content review. The item flags for the vertical scaling are listed in Tables 10 and 11 for ELA and Mathematics and Tables 12 and 13 for the item pool calibration. These tables demonstrate that a significant number of items were flagged "A" indicating difficult items particularly dichotomously scored items. The pattern of flagging was not consistent across grades in the case of vertical linking items.

Table 9. Item Flagging Based on Classical Statistics and Judgmental Review.

Flag	Definition
A	High difficulty (less than 0.10)
B	CR items with percentage obtaining any score category less than three percent of total N
C	CR items with higher criterion score mean for students in a lower score-point category
D	SR items with proportionally more high-proficient students selecting a distractor over the key
F	SR items with higher criterion score mean for students choosing a distractor than the mean for those choosing the key
H	Low difficulty (greater than 0.95)
P	SR items with positive distractor biserial correlation
R	Low item-total correlation (less than 0.30)
V	Item more difficult at the higher-grade level for vertical linking items
Z	Item needs content review (judgmental decision)

Table 10. Summary of Vertical Scaling Items with Flags (ELA).

Grade	Grade Assignment	Response Type		Flags											
				A	B	C	D	F	H	P	R	V	Z		
3	3	SR					5	2			29	18	1		
		Other	Dichotomous	26	4							8	1	5	
			Polytomous	16	25	1									
4	3	SR					1	1			23	7	1		
		Other	Dichotomous	4	2							1	1	1	
			Polytomous	4	4										
	4	SR			1			9	4			16	17	1	
		Other	Dichotomous	19	2							9	1	1	
			Polytomous	8	10										
5	4	SR					1				15	5	1		
		Other	Dichotomous	4	1							2	1		
			Polytomous	1	4										
	5	SR			2			8	8			24	18	6	
		Other	Dichotomous	17	2	1						9	3	4	
			Polytomous	6	8								1	1	1
6	5	SR					4	6			23	10	6		
		Other	Dichotomous	4		1						4	3	2	
			Polytomous	2	1								1	1	1
	6	SR						11	6			22	19	5	2
		Other	Dichotomous	29	8	2						24	8	6	
			Polytomous	2	10									4	

Grade	Grade Assignment	Response Type		Flags									
				A	B	C	D	F	H	P	R	V	Z
7	6	SR					2	2		14	8	5	
		Other	Dichotomous	10	1						7	8	2
			Polytomous	1	1							4	
	7	SR					10	6		22	22	3	
		Other	Dichotomous	29	9						19	6	2
			Polytomous	2	9							1	
8	7	SR					2			10	9	3	
		Other	Dichotomous	13	3						13	6	4
			Polytomous		2							1	
	8	SR					13	9		38	36	15	
		Other	Dichotomous	31	7	1					24	13	5
			Polytomous	1	8							6	
HS	8	SR					6	4		9	15	15	1
		Other	Dichotomous	8							9	13	1
			Polytomous									6	
	HS	SR					14	9		54	37		1
		Other	Dichotomous	72	15	2					52		3
			Polytomous	2	10	1							

Table 11. Summary of Vertical Scaling Items with Flags (Mathematics).

Grade	Grade Assignment	Response Type		Flags									
				A	B	C	D	F	H	P	R	V	Z
3	3	SR					4	2		10	8		
		Other	Dichotomous	34	7						6	4	
			Polytomous	3	7						2		
4	3	SR								1	1		
		Other	Dichotomous	5							2	4	
			Polytomous		1				1				
	4	SR					4	2		19	12		
		Other	Dichotomous	31	4						5	2	
			Polytomous	3	12	1							
5	4	SR								5	1		
		Other	Dichotomous	11	1						1	2	
			Polytomous	1	4	3							
	5	SR					2			22	12	3	
		Other	Dichotomous	39	5	1					9	5	
			Polytomous	12	11						1	1	
6	5	SR								21	1	3	
		Other	Dichotomous	9	1						4	5	
			Polytomous	2	5							1	
	6	SR					4	2		12	7	2	
		Other	Dichotomous	50	13						11	15	1

Grade	Grade Assignment	Response Type		Flags										
				A	B	C	D	F	H	P	R	V	Z	
			Polytomous	4	5							5		
7	6	SR					3	1		10	3	2		
		Other	Dichotomous	13	2						1	15		
			Polytomous										5	
	7	SR			1			4	3		8	6	3	
		Other	Dichotomous	61	19							15	20	
			Polytomous	9	7								3	
8	7	SR					2	3		4	4	3		
		Other	Dichotomous	18	2						9	20	1	
			Polytomous	2	2								3	
	8	SR						11	6		24	19	3	
		Other	Dichotomous	77	32							15	11	1
			Polytomous	11	15							2	1	
HS	8	SR					2			1	3	3		
		Other	Dichotomous	19	5						1	11		
			Polytomous	2	3								1	
	HS	SR			3			29	10		59	48		
		Other	Dichotomous	99	32							28		2
			Polytomous	15	12									

Table 12. Summary of Item Flags for the Item Pool Calibration (ELA).

Grade	Response Type		Flags									
			A	B	C	D	F	H	P	R	V	Z
3	SR					42	37		116	95		58
	Other	Dichotomous	92	17						56		37
		Polytomous	54	81								6
4	SR		2			29	17		82	85		35
	Other	Dichotomous	88	18						49		33
		Polytomous	29	50						1		2
5	SR		5			43	40		92	90	1	55
	Other	Dichotomous	68	11						56	1	35
		Polytomous	15	23						1		
6	SR					33	28		109	109	3	41
	Other	Dichotomous	107	22						75	5	48
		Polytomous	13	34							5	
7	SR					35	42		100	115	2	50
	Other	Dichotomous	123	29	1					78	6	57
		Polytomous	8	32							3	
8	SR		2			45	50		118	123	12	56
	Other	Dichotomous	113	32	6					94	11	62
		Polytomous	6	27						1	7	2
HS	SR		7			151	169		422	408		156
	Other	Dichotomous	419	114	24					281		184
		Polytomous	7	38	1							2

Table 13. Summary of Items with Flags for the Item Pool Calibration (Mathematics).

Grade	Response Type		Flags									
			A	B	C	D	F	H	P	R	V	Z
3	SR					6	6	1	32	30		11
	Other	Dichotomous	115	26						38	2	35
		Polytomous	15	20						4		6
4	SR					13	13		48	38		18
	Other	Dichotomous	123	19						32	1	41
		Polytomous	30	42	3							12
5	SR		2			22	22		68	63	4	27
	Other	Dichotomous	148	37						24	4	39
		Polytomous	45	52						2		4
6	SR		1			20	12		37	37		24
	Other	Dichotomous	163	45				1		45	11	68
		Polytomous	22	17	3					2	5	7
7	SR		3			19	17		45	40	2	22
	Other	Dichotomous	236	53						54	10	68
		Polytomous	30	23							2	4
8	SR		4			45	31		73	73	1	54
	Other	Dichotomous	254	80	3					44	11	76
		Polytomous	48	56	2					7		19
HS	SR		13			164	132		328	328		156
	Other	Dichotomous	855	268	14			1		182		329
		Polytomous	165	155	5					7		58

Field Test Classical Results

The item and test analyses include the statistics for classical item difficulty (i.e., observed percent of maximum possible score), item discrimination, and reliability. Results are presented primarily for the vertical scaling sample first. This is followed by the presentation of classical item and test results for the calibration sample that represent performance with respect to the entire Smarter Balanced item pool.

Vertical Scaling Results: Classical Item and Test Statistics. The average, item difficulty, and item-total correlation or discrimination are presented in Tables 14 and 15 for the vertical scaling of ELA and mathematics. Item statistics are given for the on-grade and off-grade items sets. Overall, the average item difficulty (observed percentage of the maximum possible score) shows that the items administered were difficult for Field Test administration participants. Most items had item difficulty levels below 0.5. An average item difficulty of 0.5 would indicate that students generally obtained half of the available score points. The most difficult items were in grade 8 and high school (on-grade) in mathematics (0.24). The easiest items were off-grade in grade 4 mathematics (0.51). It also shows the average item discrimination for the on-grade and off-grade items. The average item-test correlations ranged from a low of 0.47 in high school (on-grade) in ELA to a high of 0.62 in grades 4 and 7 (off-grade) in mathematics. The NAEP and PISA items were somewhat easier compared with the Smarter Balanced items. They demonstrated high item-test correlations when the overall IRT ability was used as the criterion.

Figures 1 and 2 compare item difficulty by plotting performance on vertical linking items across grades in ELA and Mathematics. The assumption for the vertical scaling is that in general the items will be easier in the higher-grade level compared with the lower one. The figures show that the items tend to be shifted above the diagonal line indicating that they were easier in the upper grades. There tended to be greater performance differences in grades three and four and less difference in higher-grade levels such as grade 8 and high school. Some items far off the diagonal line indicating performance differences across grades, might be considered as “outliers” and eliminated from the vertical linking. In consultation with the Smarter Balanced Technical Advisory Committee, the decision was made not to eliminate vertical linking items based solely on differences in across-grade item performance. The rationale was that leaving these items in the vertical linking better reflects performance differences across grade levels and how student growth is represented.

Table 16 presents the correlations between the CAT component and the performance tasks for the vertical scaling. The percent of the maximum possible raw score was computed for both the CAT and performance task components. The percent of the maximum possible raw score range is from 0.0 to 1.0. The correlations are across all combinations of the CAT LOFT administrations and different performance tasks for the vertical scaling sample in a grade and content area.

Table 14. Number of Items, Average Item Difficulty, and Discrimination for ELA Vertical Scaling Items.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
On-Grade	No. of Items	261	242	256	232	238	243	410
	Difficulty	0.34	0.35	0.38	0.35	0.34	0.36	0.34
	Discrimination	0.51	0.50	0.52	0.49	0.49	0.49	0.47
Off-Grade Vertical Linking	No. of Items		120	133	131	107	123	107
	Difficulty		0.45	0.45	0.42	0.36	0.38	0.36
	Discrimination		0.54	0.52	0.52	0.51	0.51	0.49
NAEP	No. of Items		28				30	27
	Difficulty		0.55				0.55	0.46
	Discrimination		0.56				0.53	0.54
PISA	No. of Items							33
	Difficulty							0.61
	Discrimination							0.62

Table 15. Number of Items, Average Item Difficulty, and Discrimination for Mathematics Vertical Scaling Items.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
On-Grade	No. of Items	304	306	306	222	239	230	319
	Difficulty	0.39	0.36	0.32	0.30	0.27	0.24	0.24
	Discrimination	0.59	0.58	0.56	0.60	0.59	0.53	0.53
Off-Grade Vertical Linking	No. of Items		104	95	102	71	73	81
	Difficulty		0.51	0.40	0.37	0.32	0.31	0.32
	Discrimination		0.62	0.61	0.58	0.62	0.59	0.56
NAEP	No. of Items		30				33	28
	Difficulty		0.49				0.47	0.41
	Discrimination		0.56				0.57	0.56
PISA	No. of Items							74
	Difficulty							0.41
	Discrimination							0.59

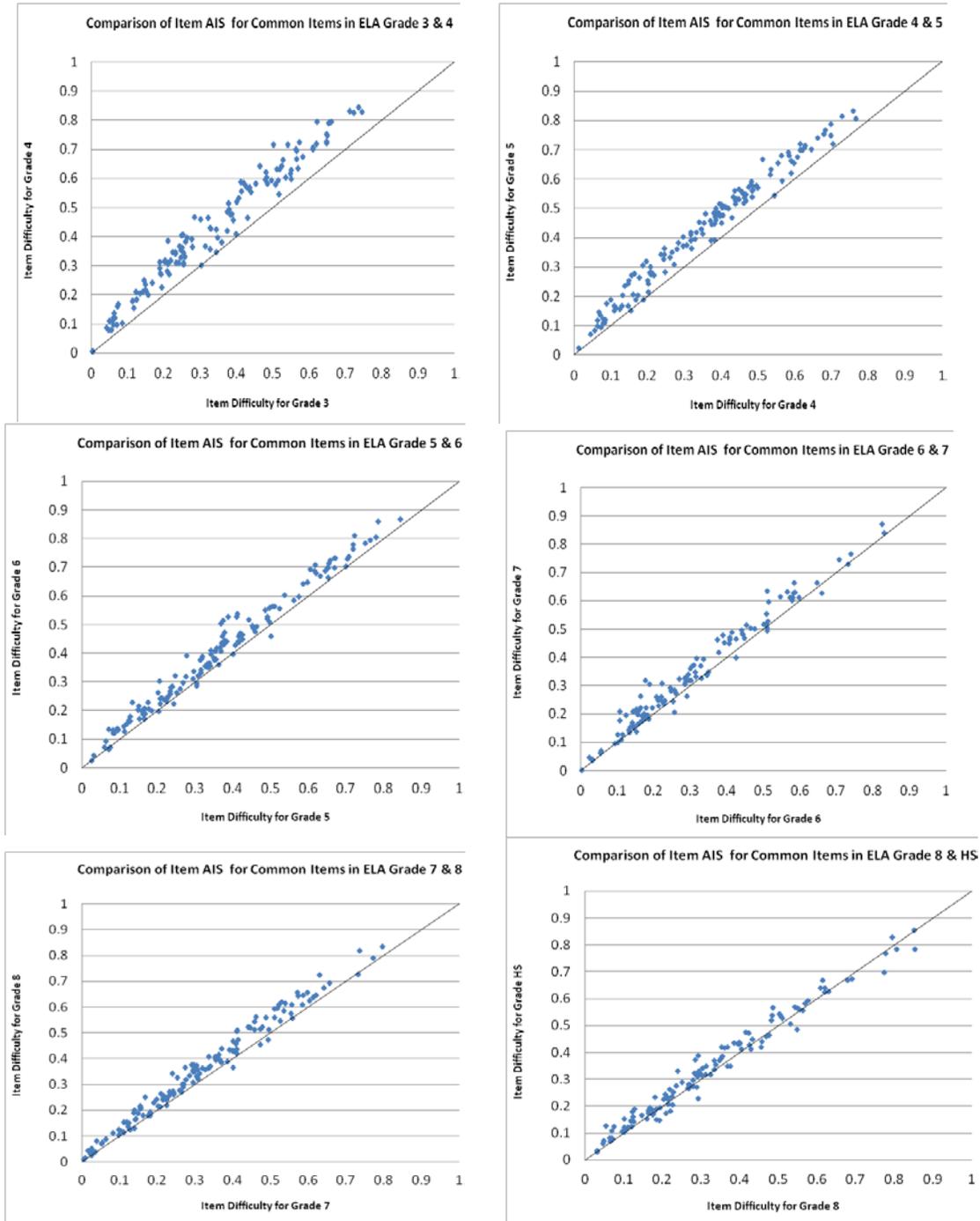


Figure 1. P-value Plots for Vertical Linking Items (ELA) (AIS is used here as association between p-values)

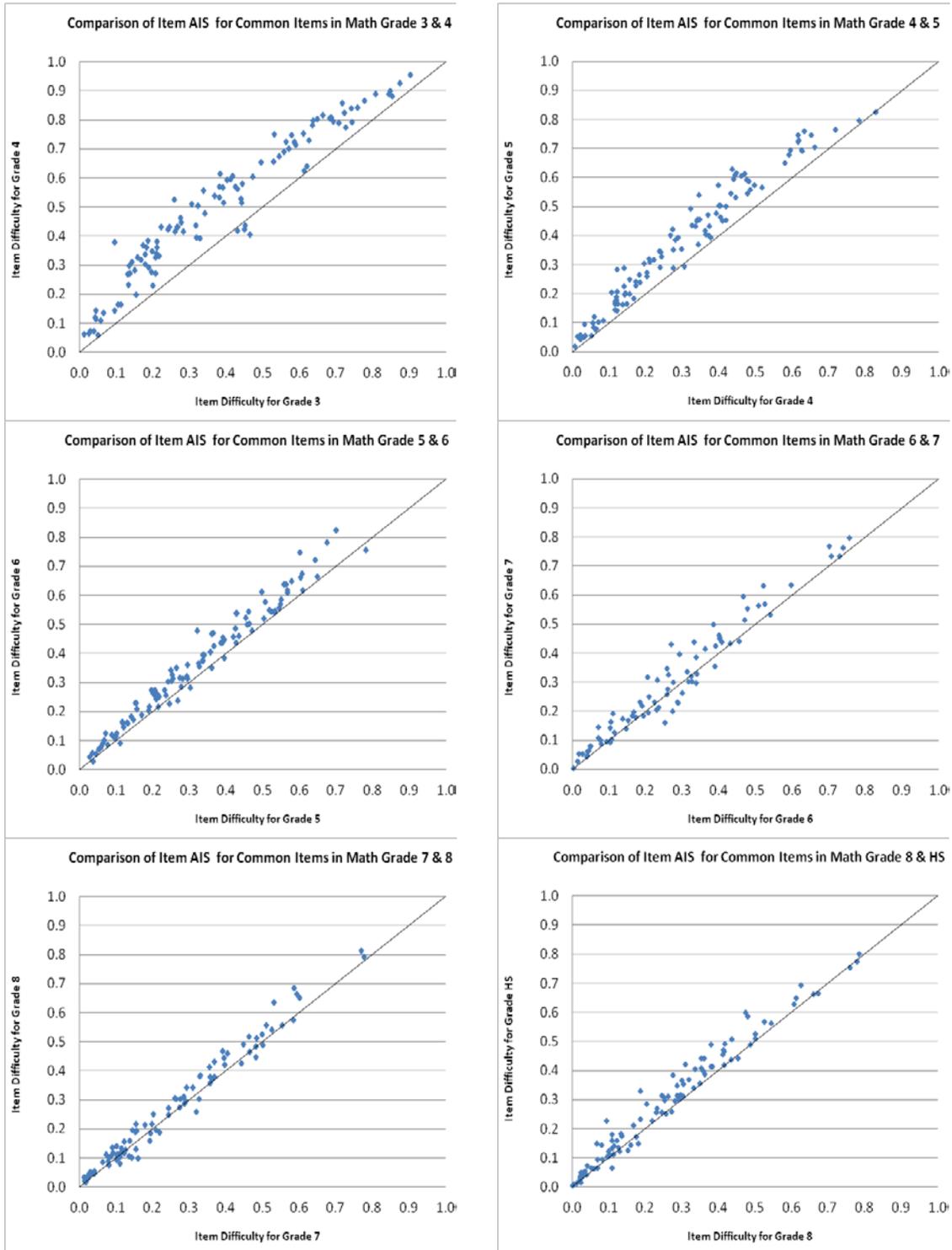


Figure 2. P-value Plots for Vertical Linking Items (Mathematics) (AIS is used here as association between p-values)

Table 16. Pearson Correlation between CAT and Performance Tasks for the Vertical Scaling.

Grade	No. of Students*	Correlation
ELA		
3	18,347	0.55
4	25,613	0.46
5	24,441	0.60
6	24,531	0.61
7	24,248	0.62
8	26,759	0.58
HS	49,392	0.55
Mathematics		
3	20,588	0.60
4	33,025	0.64
5	36,531	0.66
6	25,725	0.60
7	22,230	0.58
8	27,043	0.51
HS	46,877	0.53

Note: *No. of students refers to the number of students that have valid performance task scores (all students have CAT scores), which may be lower than the counts given in other tables.

Item Pool Calibration Results: Classical Item and Test Statistics. Table 17 shows the number of items, average item difficulty, and discrimination for the item-pool calibration sample for both ELA and mathematics. This information reflects the item pool combining performance from both performance tasks and CAT items. This Table shows that most students obtained a relatively small portion of the available score points. Items were particularly difficult in mathematics. The average discrimination was high.

Table 17. Number of Items, Average Item Difficulty, and Discrimination for the for Item Pool Calibration.

	Grade						
	3	4	5	6	7	8	HS
ELA							
No. of Items	1,015	948	952	961	1,006	990	3,310
Classical Difficulty	0.33	0.34	0.35	0.32	0.32	0.34	0.33
Classical Discrimination	0.50	0.50	0.50	0.47	0.47	0.46	0.45
Mathematics							
No. of Items	1,162	1,198	1,106	1,107	1,032	1,027	3,311
Classical Difficulty	0.39	0.35	0.28	0.29	0.24	0.22	0.22
Classical Discrimination	0.61	0.60	0.61	0.65	0.66	0.61	0.58

Table 18, which is similar to Table 16, shows the inter-correlations between the CAT and performance task for the item pool calibration. The correlations are of a similar magnitude across grades and content areas and across vertical scaling and item-pool calibration samples.

Table 18. Pearson Correlations between CAT and Performance Tasks for the Item Pool Calibration.

Grade	No. of Students	Correlation
ELA		
3	58,440	0.57
4	56,037	0.55
5	53,280	0.64
6	58,074	0.64
7	56,987	0.66
8	56,960	0.63
HS	114,621	0.58
Mathematics		
3	65,261	0.64
4	66,936	0.65
5	61,457	0.65
6	59,901	0.65
7	60,549	0.61
8	56,133	0.58
HS	116,112	0.56

Performance task reliability, as expressed by internal consistency, for performance tasks in the item pool calibration is shown in Table 19. Performance task reliability can be reported since students administered a given performance task all responded to the same set of items. The number of tasks in a grade and content area are presented and the median sample size across that set of performance tasks. The minimum, maximum, average and the standard deviations are presented for Cronbach's alpha and the standard error of measurement (SEM). Reliabilities ranged from 0.07 to 0.79 for ELA and 0.22 to 0.81 for mathematics. Note that there are multiple items or scorable units associated with a given task. In some cases, one or more items might have been dropped from a given task that resulted in very low reliability reported. Note that in the computation of an operational test score both a CAT and performance task will contribute to the overall score. There will likely be a greater number of items associated with the operational CAT compared with the performance tasks.

It is likely that the CAT combined with the performance task will result in sufficient overall levels of reliability.

Table 19. Reliability and SEM of Performance Tasks for the Item Pool Calibration.

		Reliability					SEM			
Grade	No. of PT	Median N	Min	Max	Mean	SD	Min	Max	Mean	SD
ELA										
3	19	2,392	0.58	0.72	0.66	0.04	1.04	1.42	1.25	0.11
4	24	1,542	0.07	0.74	0.65	0.13	0.44	1.58	1.36	0.23
5	25	1,401	0.64	0.74	0.69	0.03	1.20	1.51	1.36	0.07
6	20	1,961	0.62	0.79	0.70	0.04	1.15	1.53	1.30	0.11
7	25	1,328	0.64	0.78	0.72	0.03	1.15	1.46	1.26	0.10
8	27	1,251	0.62	0.76	0.70	0.04	1.01	1.63	1.32	0.14
HS	28	2,813	0.61	0.76	0.71	0.04	1.19	1.41	1.31	0.05
Mathematics										
3	24	2,004	0.55	0.79	0.69	0.07	0.93	1.46	1.24	0.14
4	28	1,731	0.53	0.77	0.65	0.06	0.96	1.41	1.18	0.13
5	20	2,234	0.57	0.76	0.66	0.05	0.96	1.61	1.19	0.18
6	30	993	0.58	0.78	0.68	0.05	0.81	1.72	1.19	0.26
7	30	956	0.51	0.78	0.65	0.08	0.53	1.20	0.96	0.15
8	28	946	0.51	0.76	0.64	0.07	0.76	1.51	1.02	0.23
HS	28	2,578	0.22	0.81	0.64	0.14	0.42	1.35	1.01	0.22

Figures 3 and 4 show the distribution of test difficulty for ELA and mathematics for the item pool calibration sample by grade level and across all grades. Using LOFT for the CAT items in a grade and content area, students were administered slightly different numbers and types of items in which the total raw score varied. Students were also administered different performance tasks. In such a design, there are many definitions of a total raw score and test difficulty. As a result, the average percent of maximum is used in a given grade and content area. Since students were administered

different items, test difficulty is the overall raw score divided by the maximum possible score for the collection administered to a given student. This corresponds to the observed percent of the maximum possible raw score including both the CAT and performance task components. A detailed example is given below.

1. A hypothetical student is administered 25 items that contain a mixture of dichotomously and polytomously scored items.
2. The item scores for the student are summed across the 25 items; a total of 45 points were obtained by this student.
3. The maximum possible raw score based on these items is 60 points.
4. The observed percent of maximum for this student is then $45/60$ or 0.75.
5. The distribution of the observed percent of maximum are plotted in the two figures. Each student would have taken essentially a unique set of items that may have varied in item difficulty.

These figures show the tests were difficult for students, which was also reflected by the average item difficulties.

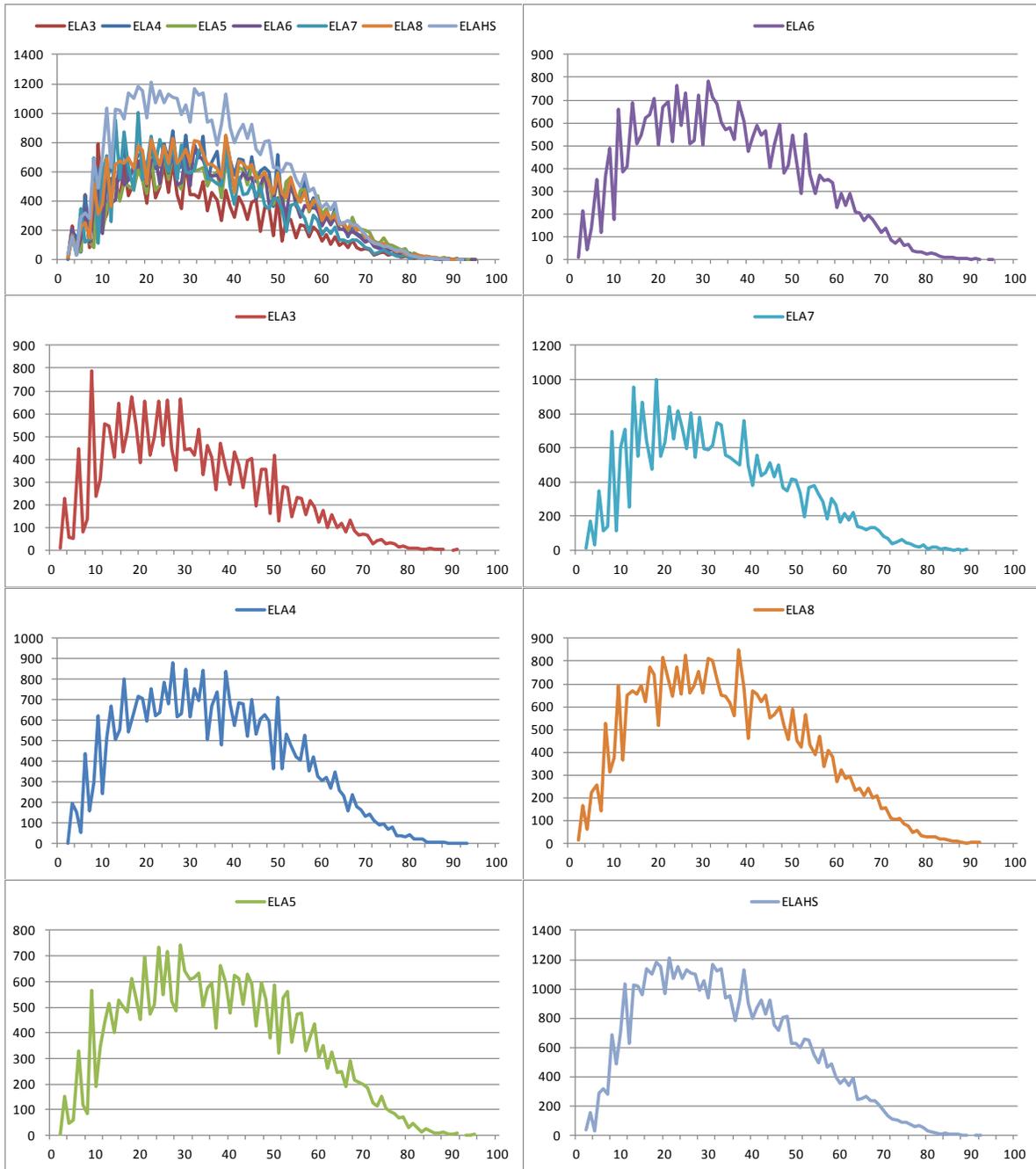


Figure 3. Distributions of Total Raw Scores as a Percentage of the Corresponding Maximum Possible Score for the Item Pool Calibration Sample (ELA)

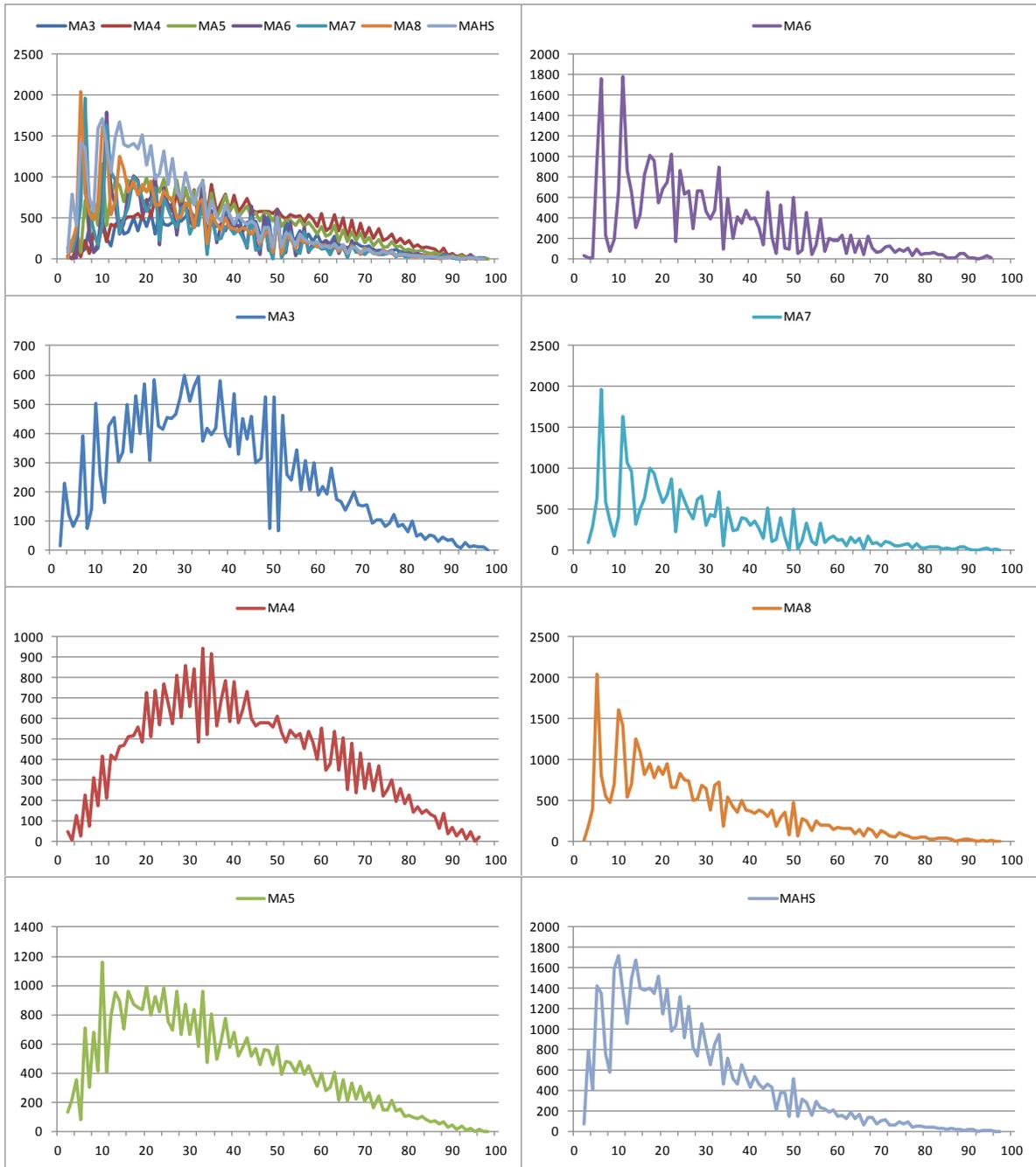


Figure 4. Distributions of Total Raw Scores as a Percentage of the Corresponding Maximum Possible Score for the Item Pool Calibration Sample (Mathematics)

Subgroup Analysis of Test Difficulty for the Item Pool Calibration. Sample size and test difficulty are reported for various subgroups. Test difficulty was defined as the observed percent of the total possible test score. This was computed by taking the overall raw score for a given student and dividing it by the maximum possible raw score. Test difficulty defined here ranges from 0.0 to 1.0. Tables 20 and 21 show average test difficulty for gender, demographic groups, limited English proficiency (LEP), accommodations (Individual Educational Plan: IEP), and Title 1 students for ELA and mathematics.

Table 20. Summary of Average Test Difficulty by Subgroup for ELA.

Grade	Subgroup	No. of Students	Average Test Difficulty	SD
3	Female	41,230	0.32	0.16
	Male	42,829	0.29	0.15
	African American	7,799	0.24	0.13
	Asian/Pacific Islander	7,106	0.35	0.17
	Native American/Alaska Native	1,662	0.23	0.12
	Hispanic	24,222	0.25	0.14
	Multiple	4,175	0.30	0.16
	White	39,095	0.34	0.16
	IEP	8,296	0.21	0.13
	LEP	13,886	0.21	0.11
	Title 1	44,640	0.25	0.13
4	Female	45,755	0.35	0.17
	Male	47,767	0.31	0.16
	African American	8,101	0.26	0.14
	Asian/Pacific Islander	7,771	0.39	0.18
	Native American/Alaska Native	2,154	0.24	0.13
	Hispanic	25,158	0.27	0.14
	Multiple	4,795	0.32	0.16
	White	45,543	0.37	0.16

Grade	Subgroup	No. of Students	Average Test Difficulty	SD
	IEP	9,818	0.22	0.14
	LEP	12,720	0.21	0.11
	Title 1	48,323	0.27	0.14
5	Female	42,623	0.38	0.17
	Male	44,167	0.33	0.16
	African American	8,218	0.29	0.15
	Asian/Pacific Islander	6,817	0.42	0.18
	Native American/Alaska Native	1,752	0.26	0.14
	Hispanic	23,262	0.30	0.15
	Multiple	4,418	0.35	0.17
	White	42,323	0.40	0.16
	IEP	9,679	0.22	0.13
	LEP	9,619	0.22	0.11
	Title 1	43,796	0.30	0.15
	6	Female	45,094	0.34
Male		46,652	0.30	0.15
African American		8,976	0.26	0.14
Asian/Pacific Islander		6,820	0.39	0.17
Native American/Alaska Native		2,138	0.24	0.13
Hispanic		25,012	0.26	0.14
Multiple		4,422	0.31	0.16
White		44,378	0.35	0.16

Grade	Subgroup	No. of Students	Average Test Difficulty	SD
	IEP	9,801	0.19	0.11
	LEP	8,808	0.18	0.10
	Title 1	46,642	0.26	0.14
7	Female	44,517	0.34	0.16
	Male	46,019	0.30	0.15
	African American	8,269	0.26	0.13
	Asian/Pacific Islander	7,397	0.39	0.17
	Native American/Alaska Native	2,068	0.24	0.12
	Hispanic	28,357	0.27	0.14
	Multiple	3,983	0.33	0.16
	White	40,462	0.36	0.16
	IEP	9,007	0.19	0.11
	LEP	8,666	0.18	0.09
	Title 1	47,399	0.27	0.14
	8	Female	46,616	0.36
Male		48,654	0.31	0.16
African American		9,630	0.28	0.14
Asian/Pacific Islander		7,284	0.40	0.18
Native American/Alaska Native		2,163	0.26	0.14
Hispanic		25,194	0.29	0.14
Multiple		4,059	0.34	0.16
White		46,940	0.37	0.16

Grade	Subgroup	No. of Students	Average Test Difficulty	SD
	IEP	9,464	0.20	0.12
	LEP	6,987	0.19	0.10
	Title 1	46,063	0.29	0.15
HS	Female	116,646	0.35	0.16
	Male	117,297	0.30	0.16
	African American	21,824	0.26	0.14
	Asian/Pacific Islander	21,973	0.39	0.18
	Native American/Alaska Native	3,443	0.28	0.14
	Hispanic	71,245	0.28	0.15
	Multiple	8,344	0.33	0.17
	White	107,114	0.35	0.16
	IEP	17,934	0.19	0.12
	LEP	14,881	0.18	0.09
	Title 1	109,507	0.28	0.15

Table 21. Summary of Average Test Difficulty by Subgroup for Mathematics

Grade	Subgroup	No. of Students	Item Difficulty	SD
3	Female	45,600	0.37	0.19
	Male	46,939	0.37	0.20
	African American	8,445	0.28	0.16
	Asian/Pacific Islander	7,348	0.46	0.21
	Native American/Alaska Native	1,904	0.28	0.16
	Hispanic	27,423	0.30	0.17
	Multiple	4,552	0.37	0.20
	White	42,867	0.42	0.19
	IEP	9,236	0.25	0.18
	LEP	16,361	0.26	0.16
	Title 1	49,037	0.31	0.17
4	Female	52,827	0.36	0.19
	Male	54,969	0.37	0.20
	African American	9,420	0.26	0.16
	Asian/Pacific Islander	8,101	0.45	0.22
	Native American/Alaska Native	3,347	0.26	0.16
	Hispanic	28,703	0.28	0.17
	Multiple	6,228	0.36	0.20
	White	51,997	0.41	0.19
	IEP	11,645	0.23	0.17
	LEP	14,337	0.23	0.14
	Title 1	56,022	0.30	0.17

Grade	Subgroup	No. of Students	Item Difficulty	SD
5	Female	52,355	0.29	0.18
	Male	54,871	0.30	0.20
	African American	8,203	0.20	0.14
	Asian/Pacific Islander	7,877	0.39	0.22
	Native American/Alaska Native	3,162	0.19	0.14
	Hispanic	27,072	0.22	0.15
	Multiple	6,164	0.30	0.19
	White	54,748	0.34	0.19
	IEP	11,851	0.17	0.15
	LEP	11,264	0.16	0.11
	Title 1	53,518	0.23	0.16
	6	Female	56,975	0.27
Male		58,624	0.27	0.19
African American		8,629	0.19	0.14
Asian/Pacific Islander		10,229	0.38	0.22
Native American/Alaska Native		1,472	0.19	0.15
Hispanic		37,395	0.21	0.15
Multiple		6,236	0.28	0.19
White		51,638	0.32	0.19
IEP		12,344	0.14	0.13
LEP		13,138	0.14	0.11
Title 1		60,158	0.21	0.15

Grade	Subgroup	No. of Students	Item Difficulty	SD
7	Female	55,007	0.23	0.17
	Male	56,723	0.23	0.18
	African American	8,577	0.15	0.13
	Asian/Pacific Islander	10,308	0.33	0.21
	Native American/Alaska Native	1,265	0.17	0.13
	Hispanic	39,425	0.17	0.13
	Multiple	5,100	0.24	0.17
	White	47,055	0.27	0.18
	IEP	11,098	0.12	0.11
	LEP	12,188	0.11	0.10
	Title 1	59,209	0.18	0.14
	8	Female	54,758	0.22
Male		56,054	0.22	0.17
African American		8,330	0.15	0.12
Asian/Pacific Islander		10,013	0.31	0.20
Native American/Alaska Native		1,316	0.17	0.13
Hispanic		35,537	0.16	0.12
Multiple		5,168	0.23	0.17
White		50,448	0.25	0.17
IEP		10,639	0.12	0.10
LEP		10,307	0.11	0.09
Title 1		55,026	0.17	0.13

Grade	Subgroup	No. of Students	Item Difficulty	SD
HS	Female	112,663	0.20	0.15
	Male	112,092	0.21	0.16
	African American	20,772	0.14	0.11
	Asian/Pacific Islander	22,132	0.31	0.21
	Native American/Alaska Native	3,370	0.16	0.12
	Hispanic	70,446	0.15	0.12
	Multiple	8,227	0.20	0.16
	White	99,808	0.23	0.16
	IEP	16,684	0.11	0.09
	LEP	16,621	0.11	0.09
	Title 1	105,246	0.16	0.12

Differential Item Functioning (DIF) Analyses for the Calibration Item Pool

In addition to classical item and test analyses, differential item functioning (DIF) analyses were also performed on the Field Test items. DIF analyses are used to identify those items that identify groups of students (e.g., males versus females) with the same underlying level of ability that have different probabilities of answering an item correctly. To perform a DIF analysis, students are separated into relevant subgroups based on ethnicity, gender, or other demographic characteristics. Students in each subgroup are then ranked relative to their total test score (conditioning on ability). Item performance from the focal group to be examined (e.g., females) is compared conditionally based on ability with the reference group (e.g., males). The definitions for the focal and reference groups used are given in Table 22. A DIF analysis asks, “If we compare focal-group and reference-group students of the same overall ability (as indicated by their performance on the full test), are any test items appreciably more difficult for one group compared with another group?” DIF in this context is viewed as a potential source of invalidity.

DIF statistics are used to identify items that are *potentially* functioning differentially. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences. If the item is differentially more difficult for an identifiable subgroup when conditioned on ability, it may be measuring something different from the intended construct to be measured. However, it is important to recognize that DIF-flagged items might be related to actual differences in relevant knowledge or statistical Type I error.

Table 22. Definition of Focal and Reference Groups.

Group Type	Focal Groups	Reference Groups
Gender	Female	Male
Ethnicity	African American	White
	Asian/Pacific Islander	
	Native American/Alaska Native	
	Hispanic	
Special Populations	Limited English Proficient (LEP)	English Proficient
	Individualized Education Program (IEP)	No IEP
	Title 1	Not Title 1

Table 23. DIF Flagging Logic for Selected-Response Items.

DIF Category	Definition
A (negligible)	Absolute value of the MH D-DIF is not significantly different from zero, or is less than one.
B (slight to moderate)	<ol style="list-style-type: none"> 1. Absolute value of the MH D-DIF is significantly different from zero but not from one, and is at least one; or 2. Absolute value of the MH D-DIF is significantly different from one, but less than 1.5. 3. Positive values are classified as “B+” and negative values as “B-”
C (moderate to large)	<ol style="list-style-type: none"> 1. Absolute value of the MH D-DIF is significantly different from 1, and is at least 1.5; and 2. Absolute value of the MH D-DIF is larger than 1.96 times the standard error of MH D-DIF. 3. Positive values are classified as “C+” and negative values as “C-“

Table 24. DIF Flagging Logic for Constructed-Response Items

DIF Category	Definition
A (negligible)	Mantel p-value >0.05 or chi-square $ SMD/SD \leq 0.17$
B (slight to moderate)	Mantel chi-square p-value <0.05 and $ SMD/SD >0.17$, but ≤ 0.25
C (moderate to large)	Mantel chi-square p-value <0.05 and $ SMD/SD > 0.25$

Items are classified into three DIF categories of “A,” “B,” or “C.” DIF Category A items contain negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large values of DIF. Positive values favor the focus group, and negative values are in favor of the reference group. The positive and negative values are reported for C-DIF item flagging. DIF analyses were not conducted if the sample size for either the reference group or the focal group was less than 100 or if the sample size for the two combined groups was less than 400. In subsequent tables, A levels of DIF are not flagged as they are too small to have perceptible interpretation.

Different DIF analysis procedures are used for dichotomous items (items with 0/1 score categories; selected-response items) and polytomous items (items with more than two score categories; constructed-response items). Statistics from two DIF detection methods are computed consisting of the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) and the standardized mean difference (SMD) procedure (Dorans & Kulick, 1983, 1986). Selected-response items are classified into DIF categories of A, B, and C, as described in Table 30.

For dichotomous items, the statistic described by Holland and Thayer (1988), known as Mantel-Haenszel D-DIF, is reported. This statistic is reported on the delta scale, which is a normalized transformation of item difficulty (p-value) with a mean of 13 and a standard deviation of 4. Items that are not significantly different based on the Mantel-Haenszel D-DIF ($p > 0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. For items where the statistical test indicates significant differences ($p < 0.05$), the effect size is used to determine the direction and severity of the DIF. The formula for the estimate of constant odds ratio is

$$\alpha_{MH} = \frac{\left(\frac{\sum_m R_{rm} W_{fm}}{N_m} \right)}{\left(\frac{\sum_m R_{fm} W_{rm}}{N_m} \right)},$$

where

R_{rm} = number in reference group at ability level m answering the item right;

W_{fm} = number in focal group at ability level m answering the item wrong;

R_{fm} = number in focal group at ability level m answering the item right;

W_{rm} = number in reference group at ability level m answering the item wrong; and

N_m = total group at ability level m .

This value can then be used as follows (Holland & Thayer, 1988):

$$MH\ D-DIF = -2.35 \ln[\alpha_{MH}].$$

The Mantel-Haenszel chi-square statistic used to classify items into the three DIF categories is

$$MH\ CHISQ = \frac{(\sum_m R_{rm} - \sum_m E(R_{rm}))^2}{\sum_m Var(R_{rm})},$$

where $E(R_{rm}) = N_{rm} R_{Nm} / N_m$, $Var(R_{rm}) = \frac{N_{rm} N_{jm} R_{Nm} W_{Nm}}{N_m^2 (N_m - 1)}$, N_{rm} and N_{jm} are the numbers of examinees in the

reference and focal groups, respectively, R_{Nm} and W_{Nm} are the number of examinees who answered the item correctly and incorrectly, respectively. The classification logic used for flagging items is based on a combination of absolute differences and significance testing. Items that are not statistically different based on the MH D-DIF ($p > 0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. For items where the statistical test indicates significant differences ($p < 0.05$), the effect size is used to determine the direction and severity of the DIF. The classification logic for selected-response items is based on a combination of absolute differences and significance testing, is shown in Table 23.

The standardized mean difference compares item performance of two subpopulations adjusting for differences in the distributions of the two subpopulations. The standardized mean difference statistic can be divided by the total standard deviation to obtain a measure of the effect size. A negative value of the standardized mean difference shows that the item is more difficult for the focal group, whereas a positive value indicates that it is more difficult for the reference group. The standardized mean difference used for polytomous items is defined as:

$$SMD = \sum p_{Fk} m_{Fk} - \sum p_{Rk} m_{Rk},$$

where p_{Fk} is the proportion of the focal group members who are at the k^{th} level of the matching variable, m_{Fk} is the mean score for the focal group at the k^{th} level, and m_{Rk} is the mean item score for the reference group at the k^{th} level. The standardized mean difference is divided by the total item group standard deviation to get a measure of the effect size. The classification logic for polytomous items is based on a combination of absolute differences and significance testing, as shown in Table 24. Items that are not statistically different based on the MH D-DIF ($p > 0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately.

A relatively small number of items showed some performance differences between student groups as indicated by C-DIF flagging criteria. Tables 25 and 26 show the number of items flagged for all categories of DIF for ELA/literacy and mathematics in grades 3 to 11. Note that the item flagging incorporates items that were administered across grades for vertical linking. A relatively small percentage of items were flagged for significant levels of DIF (C-DIF) in the Field Test for the collective item pool. All items had previously undergone bias reviews. Additional inspection of these C-DIF items was conducted by content editors before inclusion in operational tests administrations. Items with A level of DIF are not flagged because the level is too low to interpret meaningfully.

Table 25. Number of DIF Items Flagged by Category (ELA, Grades 3 to High School).

Focal Group Category									
Grade	DIF Flag	Female	Asian	African American	Hispanic	Native American	IEP	LEP	Title1
3	C+	4	5						
	C-		1		2			1	
	B	28	45	30	23	6	21	19	3
4	C+	8	7	2			2	1	
	C-	2	6	1	3			3	
	B	36	40	24	23	7	22	21	9
5	C+	18	5	3	1			2	
	C-	2		1	2		3	2	
	B	60	40	24	32	7	17	21	9
6	C+	6	11					1	
	C-	3	7	1	4		2	3	
	B	47	44	21	23	8	14	23	6
7	C+	7	8	2					
	C-	2	2	4	1				
	B	70	48	25	22	12	14	21	4
8	C+	16	12	1	3				
	C-	4	5	2	3		2	5	
	B	70	48	29	39	8	17	32	7
HS	C+	10	15	2	4		3	5	3
	C-	20	19	13	30	1	3	8	11
	B	180	161	77	138	12	60	73	74

Table 26. Number of DIF Items Flagged by Category (Mathematics, Grades 3 to High School).

Focal Group Category									
Grade	DIF Flag	Female	Asian	African American	Hispanic	Native American	IEP	LEP	Title1
3	C+	1	21	5	2		2	4	
	C-		5	1	1		1	3	
	B	14	74	80	58	1	22	39	1
4	C+	1	16	7	3	1		2	
	C-		3	2	2		1	5	
	B	25	73	40	41	14	18	41	4
5	C+		17			3	1		
	C-		5		1	1	2	5	
	B	22	61	43	9	15	21	27	3
6	C+	2	31	4	4				
	C-	2	5	3	1				
	B	29	49	18	19		7	21	7
7	C+	2	24	2			2	2	
	C-		4	1			1		
	B	27	66	19	18		26	19	7
8	C+	1	13	3			5	2	
	C-	1	6	1	2		1		
	B	11	46	22	22		24	22	2
HS	C+	10	46	4	7		1	4	4
	C-	5	2	4				2	
	B	76	60	57	59		26	18	22

Prospective Evidence in Support of Rater Agreement

Since CR items and performance tasks are an integral part of the score, it is important to establish that the results are consistent across raters and task types. This can be accomplished by evaluating the results of the Field Test administration and through careful management of the scoring processes, minimizing all possible sources of variance associated with these procedures. In order to minimize any sources of irrelevant variance, a comprehensive set of plans for evaluating and monitoring the scoring systems was implemented. The procedures described below provide a basis for monitoring whether the score categories and the underlying construct are maintained consistently in the Field Test and subsequent administrations.

Monitoring Scoring Processes. Pre-Field Test scoring procedures consist of range-finding, selection of calibration/benchmark papers, and the establishment of materials for rater training and qualification. Well-developed processes and procedures in the pre-operational phase determine the success of the operational phase. These processes include the following.

- Certification and training. Each qualified rater receives rigorous training in correctly applying the rubric at each specific score point and are required to successfully complete a certification test.
- Automated scoring. Automated scoring was implemented such that the scoring engines have to be established and “trained” to score the targeted items using a requisite number of student responses with known psychometric properties.
- Range-finding and calibration papers. Calibration papers with known psychometric properties are selected by experts and establish the standard for scoring various types of responses—these are also known as “benchmark” papers. These papers are distributed periodically during the course of scoring and are critical to determining the accuracy of scoring. In the pre-operational phase, it is critical that large and robust calibration papers are able to be selected.
- Operational. Operational scoring procedures include monitoring the method of distributing student responses, as well as real-time monitoring of rater accuracy and consistency and supervisory review and auditing.

Monitoring Raters and Associated Statistics. The statistics and methods used for monitoring rater agreement for evaluating the functioning of performance tasks may include, but are not limited to:

- number and proportion of students earning each rubric score;
- number and percentage of exact agreement between two human ratings or between automated and human scores after correcting for chance agreement rates;
- number and percentage of adjacent agreement between two human ratings or between automated and human scores after correcting for chance agreement rates;
- number and percentage of non-adjacent scores between two human ratings or between automated and human scores after correcting for chance agreement rates;
- unweighted Kappa statistics (Cohen, 1960), which characterize the degree of agreement or association between two human ratings or between automated and human scores after correcting for chance agreement rates;
- quadratic-weighted Kappa statistics (Fleiss, 1981), which have similar properties to unweighted Kappa, but are degraded disproportionately by the presence of large

disagreements between ratings of two human raters or between human and automated scores; and

- Pearson correlations, which provide another measure of the degree of agreement or association between two human ratings or between automated and human scores.

Monitoring Automated Scoring and Associated Statistics. Some validation of scoring, even for automated algorithms, is also necessary. Consistent with the procedures used with rater protocols for monitoring reliability, the following statistics can be produced.

- Similarity of human and automated score frequency distributions and means with standard deviations.
- Standardized differences (effect sizes) between human and automated score means. This is computed as the difference between means divided by the standard deviation of the human scores.
- Unweighted Kappa statistics (Cohen, 1960), which characterize the degree of agreement or association between automated and human scores after correcting for chance agreement rates.
- Quadratic-weighted Kappa statistics (Fleiss, 1981), which have similar properties to unweighted Kappa statistics but are degraded disproportionately by the presence of large disagreements between some human and automated scores.
- Pearson correlations provide another measure of agreement or association between automated and human scores.

External Assessments: NAEP and PISA.

Smarter Balanced established achievement levels with respect to the Consortium while also wanting to reference important national or international assessments such as NAEP and PISA. Inferences concerning relative performance on these items relied on assumptions concerning factors like test-delivery-mode effects, item-context effects, the impact of different testing windows and years, and the impact of different test purposes. The NAEP mathematics, reading or writing, and PISA literacy content and skills frameworks are also quite different from Smarter Balanced. Finally, NAEP and PISA data both derive from paper-based administrations, while Smarter Balanced assessments are computer-delivered (NAEP and PISA both plan computer administrations for 2015). Table 27 summarizes some high-level features of Smarter Balanced, NAEP, and PISA programs for purposes of comparison. This is followed by a brief narrative description of each program.

Table 27. Comparison of Features across the Smarter Balanced, NAEP, and PISA Assessment Programs

Design Feature	Smarter Balanced	NAEP	PISA
Construct Definition	ELA/literacy Claims—Reading, Writing, Listening, & Research Text Types: Literary & Information	Reading Frameworks: (Writing is separate.) Text Types: Literary & Information	Reading Aspects: Text Types: Exposition, Argumentation, Instruction, Transaction, & Description
	Math Claims—Concepts and Procedures, Problem solving, Model and Data Analysis, Communicating Reasoning	Math Frameworks: Number Properties and Operations, Measurement, Geometry, Data Analysis, Statistics and Probability, and Algebra	Math Aspects: Quantity, Uncertainty, Space & Shape, Change & Relationships
Item Context Effects and Test Administration Rules	<ul style="list-style-type: none"> The basic context will be maintained for NAEP and PISA items since they are administered as a set(s). The look and feel of NAEP and PISA item will likely be different from Smarter Balanced. The provision of glossaries, test manipulatives, and accommodation rules differ across programs. Smarter Balanced uses technology-enhanced items, while PISA and NAEP do not. 		
Testing Delivery Modes	LOFT delivery on computer and performance tasks online	Paper 2015: Paper Scale and Computer-based Testing Scale Study	Paper 2015: Computer-based Testing Scale
Testing Window	March–June	February	April/May
Untimed/Timed	Untimed	Timed	Timed
Delivery Design	<ul style="list-style-type: none"> Smarter Balanced Field Test LOFT blueprint(s) took into consideration the embedded set(s) properties such as testing length, reading load, and associated number of items. 		
Constructed-Response Scoring	<ul style="list-style-type: none"> Human scoring for external NAEP/PISA items was required. Approximately 30 percent of the PISA items and 30 to 40 percent of NAEP items were associated with sets requiring rater scoring. Scoring protocols such as training and qualification will need to be followed. Handwritten responses would need to be transcribed for anchors, training and qualification, and calibration papers. 		

Design Feature	Smarter Balanced	NAEP	PISA
Cohort/Population	Sample of 2014 Smarter Balanced Governing States	Based on 2013 U.S. national sample with state-level comparisons	Based on 2012 U.S. sample: 5,000 15-year-old students from 150 schools
Criterion-referenced Inferences	Designated achievement-level scores in 2014	Proficiency cut scores exist	Proficiency cut scores do not exist.
Anticipated Program Changes	No change after 2014 in content; schools still transitioning to the CCSS	Transitioning to computer in 2015 and math content domains will change.	Computer-based in 2015, and assessment framework will change.
IRT Model and Scaling Procedures	Scaling is at the overall content area level using the 2-PL/generalized partial credit model (GPCM).	3-PL and GPCM in reading and math; the main scales are weighted composites of subscales, and calibration is done at the subscale level.	Rasch (calibrated separately with relation to major domain and minor domain).
Anchor Item Requirements	<ul style="list-style-type: none"> • Construct-representative anchor sets were used. • More than one item block (test form) was implemented. 		

PISA Overview. The Program for International Student Assessment (PISA) is an international assessment that measures 15-year-old students' reading, mathematics, and science literacy. PISA also includes measures of general or cross-curricular competencies, such as problem solving. PISA is coordinated by the Organisation for Economic Cooperation and Development (OECD), an intergovernmental organization of industrialized countries, and is conducted in the United States by the National Center for Education Statistics (NCES). PISA emphasizes functional skills that students have acquired as they near the end of compulsory schooling. PISA was first administered in 2000 and is conducted every three years. The most recent assessment was in 2012. PISA 2012 assessed students' mathematics, reading, and science literacy. PISA 2012 also included computer-based assessments in mathematics literacy, reading literacy, and general problem solving, as well as an assessment of students' financial literacy. Results for the 2012 mathematics, reading, science, and problem-solving assessments are currently available.

NAEP Overview. The National Assessment of Educational Progress (NAEP) is a continuing and nationally representative assessment of what our nation's students know and can do. There are two types of NAEP assessments: the main NAEP and the long-term strand NAEP. The main NAEP was utilized for the Smarter Balanced Field Test. It is administered to fourth-, eighth-, and twelfth-grade students across the country in a variety of subjects. National results are available for all assessments and subjects. Each NAEP assessment is built from a content framework that specifies the types of knowledge and skill that students are expected to know in a given grade. When assessing performance for the nation only, approximately 6,000 to 20,000 students per grade from across the country are assessed for each subject. A sampling procedure is used to ensure that those selected to participate in NAEP will be representative of the geographical, racial, ethnic, and socioeconomic diversity of schools and students across the nation. NAEP is not designed to report

individual test scores, but rather to produce estimates of scale score distributions for groups of students.

Results

The subset of students who took NAEP and PISA items also took Smarter Balanced CAT items and performance tasks. A summary of the resulting item performance for NAEP and PISA and all Smarter Balanced students are presented in Table 28 for ELA/literacy and Mathematics showing the number of items administered, mean item difficulty (i.e., p-values) and discrimination. Figure 5 shows the p-values for NAEP items plotted against the ones obtained from the Smarter Balanced vertical scaling sample. The graphs suggest a reasonably linear relationship. The NAEP p-values are relatively higher (i.e., easier) than the ones obtained from the vertical scaling sample. There are other factors such as mode effects (i.e., on-line vs. paper) that might also account for these performance differences. It was not possible to obtain similar sorts of item difficulty information from the PISA program.

Table 28. Number of Items, Average Item Difficulty, and Discrimination for NAEP and PISA Items.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
ELA/literacy								
NAEP	No. of Items		28				30	27
	Difficulty		0.55				0.55	0.46
	Discrimination		0.56				0.53	0.54
PISA	No. of Items							33
	Difficulty							0.61
	Discrimination							0.62
Mathematics								
NAEP	No. of Items		30				33	28
	Difficulty		0.49				0.47	0.41
	Discrimination		0.56				0.57	0.56
PISA	No. of Items							74
	Difficulty							0.41
	Discrimination							0.59

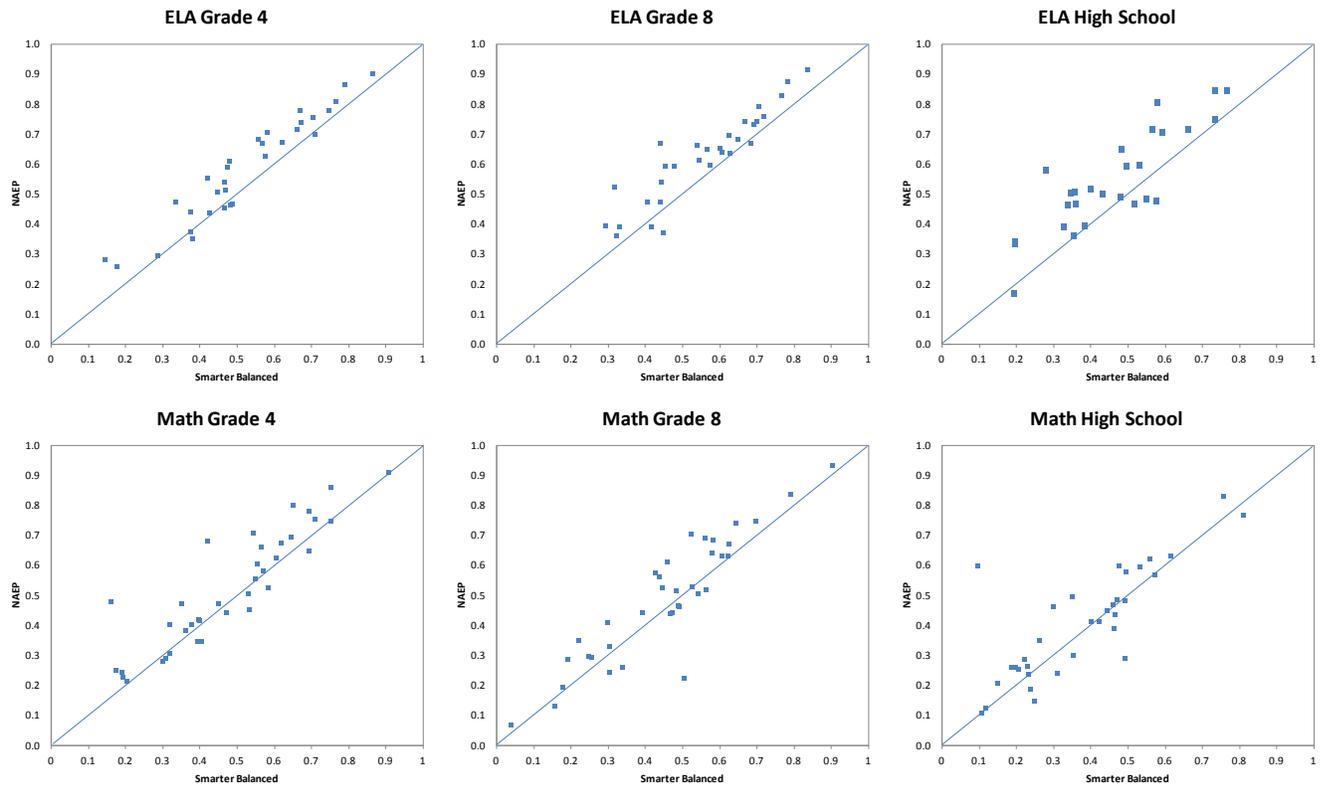


Figure 5. Comparison of NAEP Item Difficulty and Values Obtained from Smarter Balanced Samples

References

- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*, 37-46.
- Dorans, N. J., & Kulick, E. (1983). *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach (RR-83-9)*. Princeton, NJ: Educational Testing Service.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.
- Dragow, F. (1988). Polychoric and Polyserial Correlations. In L. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences, 7*, 69-74. New York: Wiley.
- Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. 2nd ed. (New York: John Wiley) pp. 38-46.
- Holland, P. W., & Thayer, D. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Mantel, N., & Haenszel, W. M. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute, 22*, 719-748.
- Olsson, U. (1979). Maximum Likelihood Estimation of the Polychoric Correlation Coefficient. *Psychometrika, 4*, 443-460.

Chapter 9 Field Test IRT Scaling and Linking Analyses

Introduction

The primary purposes of the Smarter Balanced assessments are to provide valid, reliable, and fair information concerning students' English Language Arts/literacy (ELA/literacy) and mathematics achievement with respect to the Common Core State Standards in grades 3 to 8 and high school. An important allied goal is to measure students' annual growth toward college and career readiness in grade 11 ELA/literacy and mathematics. For federal accountability purposes and potentially for state and local accountability systems, students' ELA/literacy and mathematics proficiency must also be reported. To meet these goals requires many technical characteristics to be demonstrated as evidence in support of validity. For instance, students must be measured on a common scale within a grade and content area. The methodology used to accomplish these varied goals is Item Response Theory (IRT). This chapter explains the methods used to construct the Smarter Balanced measurement scales using IRT. A description of the major Field Test scaling and linking activities in support of these goals are summarized in Figure 1.

As demonstrated by years of successful application in K-12 testing programs, IRT methods have the flexibility and strength to support the Smarter Balanced Consortium goals. IRT methods are ideally suited to the assessments and measurement goals of Smarter Balanced in both establishing a common scale and ongoing maintenance of the program, such as new item development and test equating and enabling computer adaptive testing (CAT) to be conducted (Wainer, 2000). Mixed-item-format tests, such as the Smarter Balanced assessments, that consist of dichotomous (selected-response) items, short answer responses, and performance tasks can be combined together and scaled concurrently (Ercikan, Schwarz, Julian, Burket, & Link, 1998; Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 2006). The purpose of the IRT horizontal calibration and scaling was to place items and ability estimates onto a common scale in a grade and content area. Since the Common Core State Standards (CCSS) were intended to be coherent and articulate across grade levels, they provided a foundation for developing Smarter Balanced assessments that support inferences concerning student change in achievement (i.e., growth). One approach to modeling student growth across grades is to report scores on a common vertical scale. For instance, comparing the means and standard deviations for scale scores across grades on the same scale is an intuitive approach for evaluating growth for a variety of test users. Vertical scales assume that increasing student proficiency is demonstrated across different levels of the assessment. For the CAT administration, vertical scaling permits items to be used across different grade levels if required. Another advantage of vertical scaling is that growth expectations concerning the establishment of achievement levels across grades can be inspected and ordered by standard setting panelists.

The IRT scaling for Smarter Balanced was performed in two steps. The first step was used to establish the horizontal and vertical scales that were used to set achievement levels. In the first step, items were initially scaled horizontally, where items in a single grade and content area were concurrently (i.e., simultaneously) calibrated. The vertical linking was accomplished using common items administered across grades (e.g., the same items given in 3rd and 4th grades) and then placing consecutive grades onto the vertical scale. In the second horizontal calibration step, the remaining, and much larger, item pool (containing noncommon items, each administered only to one grade) was scaled using the items from the first phase as linking/common items. Procedures associated with the IRT horizontal scaling are presented first. The horizontal scaling is followed by a discussion of assumptions, the methods used for vertical scaling, and the Field Test results. A cross validation of the vertical scaling is also briefly described. Next, the scale properties of selected NAEP and PISA items are presented, which were included to give further context to the establishment of the Smarter Balanced achievement levels concerning national and international comparisons.

Figure 1. Major Goals and Activities for Field Test Statistical Analysis

	Primary Goals	Major Analysis Activities
Phase 1 (Vertical Scaling)	<ul style="list-style-type: none"> Establish horizontal and vertical scales Analyze items and student “tests” that are scored on an expedited schedule to support achievement level setting Produce classical and IRT item statistics Provide provisional student proficiency estimates 	<ol style="list-style-type: none"> Performed classical item analysis and DIF analysis for Smarter Balanced vertical scaling items (on- and off-grade) in each grade/content area Calibrated all Smarter Balanced vertical scaling items (on- and lower-grade) at each grade/content Performed vertical scaling with Grade 6 as the pivot/base grade using embedded vertical scaling items from the lower-grade Estimated student proficiency Finalized recommendations for lowest and highest values of theta
NAEP/PISA Item Analysis	<ul style="list-style-type: none"> Provide IRT item parameters for embedded NAEP/PISA items on the Smarter Balanced vertical scales 	<ol style="list-style-type: none"> Calibrated all Smarter Balanced vertical scaling items and NAEP items Calibrated all Smarter Balanced vertical scaling items and PISA items (in high-school) Performed horizontal linking in the respective grades with on-grade Smarter Balanced items as linking items using their vertically-scaled item parameters Provided the resulting item parameters for NAEP/PISA items for use in standard setting
Phase 2 (Item Pool Calibration)	<ul style="list-style-type: none"> Provide classical and IRT item statistics for the remainder of the Field Test items on the Smarter Balanced scale 	<ol style="list-style-type: none"> Calibrated all Smarter Balanced on-grade items ($n \geq 500$) at each grade/content Performed horizontal linking in the respective grades with on-grade items as linking items with their vertically-scaled item parameters Provided IRT parameter estimates for the item pool

Horizontal Scaling: IRT Calibration for a Single Grade

Many K-12 programs scale, perform, and equate horizontally in the context of annual year-to-year assessments. For horizontal scaling in Smarter Balanced, methods using simultaneous, concurrent calibration of items were conducted at each content area/grade level. The calibration approach relied on a hybrid of the common items approach and the randomly equivalent groups linking approach. The “common items” approach requires that items and tasks partially overlap and are administered to different student samples. For the “equivalent groups” approach, the test material presented to different student samples is considered as comparably “on scale” by virtue of the

random equivalence of the groups. The random equivalence was implemented using the linear-on-the-fly test (LOFT) administration. Since neither type of linking method is guaranteed to work perfectly in practice, the linking design incorporated both types of approaches. This is done by assembling partially overlapping test content and randomly assigning them to students. The result is a design that is both reasonably efficient and well structured for IRT calibration. For further details concerning implementation of the data collection design and sampling, see Chapter 7, “Test Design and Field Test Design, Sampling, and Administration”, of the Technical Report.

The student response data consisted of the combined CAT and performance task (PT) components that were intended to measure the designated English language arts/literacy (ELA/literacy) or mathematics constructs as defined by the respective Field Test blueprints. The first step of the analysis was to create an item by student matrix reflecting item scores as well as missing information by design. For a given grade and content area, the dimension of this sparse data matrix was the total number of students times the total number of unique items (i.e., scorable units). Since each student only took a small subset of the available items, the remaining cells of the matrix represented items that were not administered. A provision in many IRT software programs is made for this “not-presented” or “not-reached” information necessary when multiple test forms are present. Students received a score that ranged from zero to the maximum permissible score level for the item administered. Using this sparse data matrix, a single grade-level concurrent calibration of all item data was performed. The procedures described here assumed that a unidimensional structure within each grade level is supported by the dimensionality analyses from the Pilot Test (see Chapter 6). Also based on the Pilot, the two-parameter logistic (2-PL) and generalized partial credit model (GPCM) models were chosen and implemented using the IRT program **PARSCALE** (Muraki & Bock, 2003). Chapter 6, on the Pilot Test, can be referenced for the results and decisions concerning of the IRT model comparison.

Vertical Scaling: Linking Across Multiple Grades

Determining whether students are making sufficient academic growth has received increased attention stemming from the No Child Left Behind (NCLB) Federal legislation. More recently, in the Race-to-the-Top legislation, there is a renewed emphasis on inferences concerning growth. These changes are intended to refocus instructional emphasis and facilitate inferences regarding change in academic achievement and readiness. Race-to-the-Top uses the Common Core State Standards (Common Core State Standards Initiative, 2010), which are articulated across grades and targeted at college- and career readiness. One method for evaluating change from one grade level to another is to develop a single common scale for use across multiple grade levels. Students are then ordered along the vertical scale implying that there is a progression of learning, primarily along a major or dominant dimension. The Common Core State Standards specifies across-grade-level articulation of content that is consistent with the general specifications for the construction of vertical scales. Another definition of growth embodied in the Common Core State Standards is learning progressions, which demonstrate how learning unfolds and characterize academic change at a finer level. As a result, there is increased interest in characterizing the amount of change that occurs for individual students or groups of students as they progress across grades. For these reasons, a continuous vertical scale reflecting growth was desired for the Smarter Balanced ELA/literacy and mathematics assessments ranging from grade 3 to 8 and high school. By contrast, in the NCLB legislation, there was no requirement for defining the relationships between content and performance standards across grades. In many instances, assessments and content standards were developed in a somewhat piecemeal fashion because the legislation was phased in over several years. For example, NCLB legislation began with a requirement (in reading and mathematics) in each of three grade spans (grades 3 to 5, 6 to 9, and 10 to 12). Subsequently, states were required to “fill in the gaps” and have assessments at grades 3 to 8. In many states, scale scores were established independently in each grade, which made inferences across grade levels more difficult. By contrast,

vertical scales permit scores on assessments administered at different grades to be reported using a common scale. The difference in scale scores for students from one grade to the next higher grade is used as an indicator of “growth” or change in achievement.

To conduct vertical scaling, common items are often administered to students at other grade levels than the targeted grade level for which they were developed, and primarily administered. Vertical scales assume that there is substantial overlap in the construct across grade levels. An assumption for test equating and interchangeable test scores is that test content and technical characteristics are parallel. Comparing students or groups of students who take parallel forms will generally be strongly supported, and most decisions of consequence for students and schools have depended on these types of cohort comparisons. Vertical scaling is not strictly equating. Holland and Dorans (2006) proposed a taxonomy consisting of three levels of linking that correspond to prediction, aligning, or equating. In this taxonomy, vertical scaling is a form of aligning in which tests have similar constructs and reliability but have different levels of difficulty and test taker populations. Overlap in content standards at adjacent grades may support the proposition that test forms for adjacent grades measure a common construct, but differences in the standards and psychometric properties of the test forms (e.g., test difficulty) imply that these forms are not parallel, so they may be linked but not equated. In the case in which scores are not parallel but a common proficiency is measured, linking can still occur (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999).

Assumptions and Interpretive Cautions Concerning Vertical Scales. The establishment of a vertical scale implies a) an increase in the difficulty of the assessments as the grade level increases, and b) a generally greater student proficiency in higher grades relative to lower grades. Accordingly, at the item level, it is assumed that students at a higher-grade level will generally have a higher probability of correctly answering an item than students at a lower grade level. This basic proposition must be substantiated to ensure the validity and plausibility of the vertical scale. With a sufficiently large and diverse sample of students, scale score means and other quantiles of the score distribution are expected to increase with grade level with a somewhat smooth pattern that is not erratic.

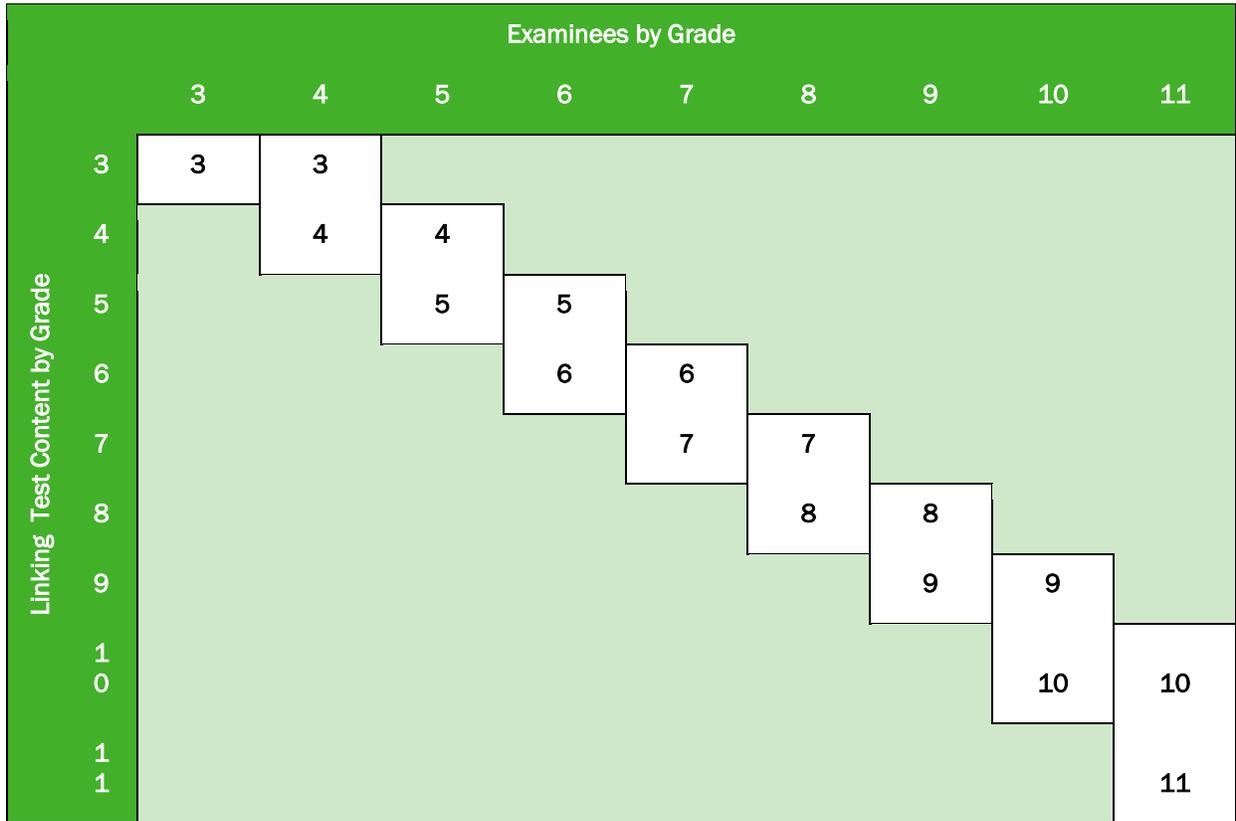
Validity evidence that vertical scales are appropriate for measuring students’ annual progress would include items that are easier in upper grades and have sensible patterns of variability within and across grades (i.e., meaningful separation of means and proficiency distributions across grades). The notion of interval level measurement dates back to taxonomy of measurement suggested by Stevens (1946). The familiar measures of height and weight, for example, exhibit this property. Interval level units are at best approximated in scales built for measuring latent variables such as academic ability. A ten-point difference in scale score units may mean something different at the low end of the score scale than it does at the middle or high score range. Achievement tests are constructed to have a strong first factor (i.e., essentially unidimensional), but multidimensionality to some extent will be reflected by changes in the content sampling across grades. The multidimensionality of the scale will be impacted by the relative importance of content subdomains at a given grade level and will determine the strength of these connections. Yen (2007) suggested that vertical scales are similar to a folding ruler that curves through space when held out. Connections among some levels of the scale are firmer, while others are somewhat looser. As a result, interpreting changes in scale scores is made more challenging when vertical scales are utilized. Additional validation efforts are appropriate when the change in scale scores over time is a focus of interest or for accountability purposes. In evaluating the vertical scale, an important question is whether the growth demonstrated by the vertical scale is consistent with expectations and the scope and sequence of grade-level instruction reflected in the Common Core State Standards. Growth needs to be interpreted in the context of the underlying scale variability. Finally, different vertical scaling methods can interact to stretch or compress the scale. Briggs and Weeks (2009), on the assessment they examined, suggested that the choice of the IRT scaling model had the largest impact on how growth

is depicted, followed by the choice of the calibration design and IRT scale score type (e.g., MLE, MAP, or EAP) chosen.

The IRT scaling and linking procedures described here were conducted in three successive stages. First, the scaling began with the evaluation of separate, horizontal (i.e., grade-specific), and concurrent scaling of items in the targeted item pool during Phase 1. After the grade-specific horizontal scaling was conducted in a content area, a separate, cross-grade vertical linking occurred using common items, also in Phase 1. The vertical scaling/linking was undertaken using the test characteristic curve transformation methods (Stocking & Lord, 1983). Using grade 6 as the baseline, each grade was successively linked onto the vertical scale separately for ELA/literacy and mathematics. Once the Smarter Balanced horizontal and vertical scales were established, the remaining items (i.e., the larger calibration item pool including the noncommon items) were linked horizontally onto this final scale in each grade and subject area in Phase 2. Another method for conducting vertical scaling is the multiple-group concurrent approach, which calibrates all grades simultaneously in a single step. The concurrent approach in vertical scaling context is a multigroup, nonequivalent group method that estimates underlying population distributions (latent means and standard deviations) for each group (Mislevy, 1987; Bock & Zimowski, 1997). Multigroup IRT permits the examination of group characteristics as a unit of analysis rather than as just individuals. For vertical scaling, the latent means should increase monotonically across grade levels. This method calibrates all students and grade levels in a single step.

Concurrent calibration uses all available item response information in the calibration and is therefore more efficient. Several studies that have investigated separate versus concurrent calibration have been inconclusive or limited in some respects, or found no substantive differences (Kim & Cohen, 1998; Hanson & Béguin, 1999; Ito, Sykes, & Yao, 2008). However, concurrent calibration can have limitations, such as convergence problems and restrictions on the number of items and observations the IRT software can handle. Kolen and Brennan (2004) suggested that separate linking steps might be preferable since it is more difficult to detect how items behave across grade levels or to diagnose any convergence problems in estimation, and violations of unidimensionality can be more problematic with the concurrent approach. The separate calibration approach produces two sets of item-parameter estimates, which can help identify and remediate potential problems. For example, if an item functioned poorly or was highly unstable across levels, it could be dropped as a vertical linking (common) item. This type of problem would be essentially undetectable with concurrent calibration, where all item parameters are estimated simultaneously, assuming common parameters across grade levels. Despite the utility of the multiple-group concurrent approach in other applications, and to reduce risk, concurrent calibration was not used as the vertical scaling method. The factors mentioned here and unknown effects of other factors such the size of the data matrix (item by student) across all grades, and the possibility of poor item functioning in the context of a Field Test led to this decision.

Figure 2. Summary of Field-Test Vertical Linking Item Configuration



Vertical Scale Linking Design. To implement the vertical scaling, representative sets of off-grade items were administered to an adjacent-upper grade. For example, grade 4 items were also administered to grade 5 students. To the extent possible, vertical linking item sets were intended to sample the construct that included both the CAT and performance task components and associated item types, and claims that conformed to the test blueprint. Linking items from the lower grade were administered to the upper-adjacent-grade level students, as shown in Figure 2. Content experts designated a target grade for each item and a minimum and maximum grade designation. Table 1 shows the vertical scaling linking design in terms of the number of CAT items and performance tasks. A set of six performance tasks was given on-grade, and the same set was administered off-grade for vertical linking. Each performance task had five or six items associated with it according to the test specifications. The same set of six performance tasks was administered in grades 9, 10, and 11 (high school). Table 2 presents the number of CAT items and performance tasks available for the vertical linking after test delivery and some item exclusions. In mathematics, particularly in grades 6 to 8 and high school (HS), a reduced number of items were available for vertical linking after test delivery, relative to the original test design. The total shown in Table 2 on the right is the number of items surviving after the IRT flagging criteria were applied, resulting in exclusion of some items (discussed in the next section). Other items might have been excluded in prior steps based on poor classical statistics. A full description of the item and test exclusion rules is given in Chapter 8 on “Field Test Dastep and Classical Test Analysis”. In some cases, a single item was eliminated from a given performance task. The resulting Smarter Balanced claim distributions for on-grade items and those targeted for vertical linking are presented in Table 3. In ELA/literacy, the claims are, respectively, Reading, Writing, Speaking/Listening, and Research, respectively. In mathematics, the

claims are Concepts and Procedures, Problem Solving, Communicating/Reasoning, and Modeling/Data Analysis. Tables 4 and 5 give a summary of item types by their purpose for ELA/literacy and mathematics.

IRT Preprocessing and Item Inclusion/Exclusion Criteria. Item functioning was evaluated prior to calibration and during the course of calibration in which items for which parameter estimates did not converge, or poorly functioning items, were excluded. In the data step, items were required to have 10 observations in a score category level for constructed-response (CR) items or 500 observations overall in order to obtain sufficiently stable IRT estimates. Many items, particularly in high school, were eliminated due to low numbers of observations. Chapter 8 has a complete description of the item- and student-exclusion rules applied and the resulting number of items available for vertical scaling. Some additional IRT-based rules are:

1. Local item dependence. ELA/literacy performance tasks contained a single “long-write” writing prompt that was subsequently scored for the dimensions of organization, elaboration, and conventions. These resulting scores were very highly correlated in the Field Test. Very high correlations between ratings of a single writing response can lead to local item dependence, which is a violation of IRT assumptions (Yen, 1993). As a result, the two dimensions for organization (0 to 4 score points) and elaboration (0 to 4 score points) were averaged and rounded for IRT scaling. This resulted in a score that ranged from zero to four points for the long-write performance tasks. In some cases, it was also necessary to collapse the top score because it had few or no observations.
2. Non-convergence results when item parameters could not be estimated in **PARSCALE**. Poor item parameter estimation was defined by either not achieving the criterion of largest estimate change lower than 0.005 or an erratic pattern of loglikelihood values. Standard errors were also evaluated along with item-parameter estimates as to their reasonableness.
3. For IRT analysis, all items with *a*-parameter estimates (i.e., discrimination) below 0.10 or the combination of *a*-parameter estimates below 0.20 and *b*-parameter estimates (i.e., difficulty) above 4.0 were excluded.
4. IRT parameter estimates, item characteristic curve plots, associated standard errors, along with item goodness-of-fit statistics were evaluated holistically to determine the quality of the resulting item parameter estimates. After examining these item characteristics, additional items were excluded due to poor functioning (e.g., a combination of very low discrimination and poor fit).
5. These criteria resulted in a subset of items in each grade being excluded due to poor IRT functioning. If a vertical linking item was excluded in the on-grade designation, it was also eliminated as a vertical linking item.

In general, after these exclusions were implemented overall convergence was met, and the resulting IRT item/ability parameter estimates under each model combination were reasonable.

Table 1. Number of Items by Type in the Vertical Linking Design.

Grade	CAT		PT		NAEP/PISA
	On-Grade	Off Grade	On-Grade	Off Grade	
ELA/literacy					
3	300	--	6		
4	300	150	6	6	75
5	300	150	6	6	
6	300	150	6	6	
7	300	150	6	6	
8	300	150	6	6	75
9		150		6	
10		150		6	75
HS	300	150	6	6	75
Mathematics					
3	300	--	6		
4	300	150	6	6	75
5	300	150	6	6	
6	300	150	6	6	
7	300	150	6	6	
8	300	150	6	6	75
9		150		6	
10		150		6	75
HS	300	150	6	6	75

Table 2. Unique Number of CAT Items and Performance Tasks (PTs) Administered and the Survivorship for Vertical Scaling.

Administered						Survivorship				
Grade	CAT		PT		NAEP/ PISA	CAT		PT		NAEP/ PISA
	On-grade	Off-grade	On-grade	Off-grade		On-grade	Off-grade	On-grade	Off-grade	
ELA/literacy										
3	306	-	6			261	-	6		
4	280	159	6	6	31	242	120	6	6	28
5	313	156	6	6		256	133	6	6	
6	292	160	6	6		232	131	6	6	
7	289	158	6	6		238	107	6	6	
8	300	161	6	6	30	243	123	6	6	30
HS	602	153	6	6	31/34	410	107	6	6	27/33
Mathematics										
3	320	-	6			304	-	6		
4	332	128	6	6	37	306	104	6	6	30
5	325	126	6	6		306	95	6	6	
6	327	127	6	6		222	102	6	6	
7	319	126	6	6		239	71	6	6	
8	333	128	6	6	36	230	73	6	6	33
HS	573	129	6	6	35/82	319	81	6	6	28/74

Table 3. Summary of ELA/literacy and Mathematics Items by Purpose and Claim.

Grade	Purpose	Number of Items	Claims (Percent)				
			1	2	3	4	Not Assigned*
ELA/literacy							
3	On-Grade	261	36	27	19	18	
	Off-grade	-	-	-	-	-	
4	On-Grade	242	30	28	21	21	
	Off-grade	120	33	28	22	17	
5	On-Grade	256	36	26	18	20	
	Off-grade	133	41	24	19	17	
6	On-Grade	232	31	29	19	21	
	Off-grade	131	40	24	19	17	
7	On-Grade	238	32	29	19	20	
	Off-grade	107	38	27	17	18	
8	On-Grade	243	34	27	20	19	
	Off-grade	123	40	28	20	13	
HS	On-Grade	410	44	31	10	16	
	Off-grade	107	45	23	20	12	
Mathematics							
3	On-Grade	304	61	6	15	6	12
	Off-grade	-	-	-	-	-	-
4	On-Grade	306	59	6	17	8	11
	Off-grade	104	56	4	15	7	18
5	On-Grade	306	59	6	16	7	12
	Off-grade	95	58	3	19	6	14
6	On-Grade	222	48	9	18	9	16
	Off-grade	102	59	4	13	7	18
7	On-Grade	239	56	4	16	9	15
	Off-grade	71	39	6	21	8	25
8	On-Grade	230	57	7	15	7	14
	Off-grade	73	49	4	12	10	25
HS	On-Grade	319	60	7	14	10	9
	Off-grade	81	57	6	15	7	15

Note: *Not Assigned refers to items that were not assigned to a claim at the time of the Field Test.

Table 4. Summary of ELA/literacy by Type and Purpose.

Item Purpose	Item Response Type	Score Type	Number of Items per Grade						
			3	4	5	6	7	8	HS
On-grade	SR*		115	92	95	81	77	91	142
	Other	Dichotomous	110	119	122	114	122	113	216
		Polytomous	36	31	39	37	39	39	52
Off-grade Vertical Linking Items	SR			57	53	46	27	38	39
	Other	Dichotomous		45	62	63	58	62	58
		Polytomous		18	18	22	22	23	10
NAEP	SR			22				20	12
	Other	Dichotomous		2				2	4
		Polytomous		4				8	11
PISA	SR								17
	Other	Dichotomous							12
		Polytomous							4

Note: *SR refers to selected-response.

Table 5. Summary of Mathematics by Type and Purpose.

Item Purpose	Item Response Type	Score Type	Number of Items per Grade						
			3	4	5	6	7	8	HS
On-grade	SR		48	65	78	21	39	41	66
	Other	Dichotomous	221	212	175	174	185	159	203
		Polytomous	35	29	53	27	15	30	50
Off-grade Vertical Linking Items	SR			11	12	31	9	7	18
	Other	Dichotomous		76	71	56	55	60	56
		Polytomous		17	12	15	7	6	7
NAEP	SR			20				19	18
	Other	Dichotomous		2				6	4
		Polytomous		8				8	6
PISA	SR								19
	Other	Dichotomous							44
		Polytomous							11

IRT Models and Software

Unidimensional IRT models were used to calibrate the selected-response and constructed-response (i.e., polytomous) items. Using the criteria and results from the Pilot Test and consultation with Smarter Balanced, the two-parameter logistic and the generalized partial credit models were chosen for use in the Field Test to establish the scale. For selected-response items, the two-parameter logistic (2PL) model was used (Birnbaum, 1968). The 2PL model is given by

$$P_i(\theta_j) = \exp[Da_i(\theta_j - b_i)] / \{1 + \exp[Da_i(\theta_j - b_i)]\},$$

where $P_i(\theta_j)$ is the probability of a correct response to item i by a test taker with ability θ_j ; a_i is the discrimination parameter; b_i is the difficulty parameter, for item i , and D is a constant that puts the θ ability scale into the same metric as the normal ogive model ($D=1.7$).

For constructed-response items, the generalized partial credit model (GPCM; Muraki, 1992) or partial credit model (PCM; Masters, 1982) is employed. The generalized partial credit model is given by

$$P_{ih}(\theta_j) = \frac{\exp \sum_{v=1}^h [Da_i(\theta_j - b_i + d_{iv})]}{\sum_{c=1}^{n_i} \exp \left[\sum_{v=1}^c Da_i(\theta_j - b_i + d_{iv}) \right]},$$

where $P_{ih}(\theta_j)$ is the probability of examinee j obtaining a score of h on item i , n_i is the number of item score categories, b_i is the item location parameter, d_{iv} is the category parameter for item i for category v , and D is a scaling constant given previously.

PARSCALE (Muraki & Bock, 2003) was used for the IRT calibrations. **PARSCALE** is a multipurpose program that implements a variety of IRT models associated with mixed-item formats and associated statistics. The psychometric properties of **PARSCALE** are well known, and it can efficiently and accurately calibrate large data sets such as those of Smarter Balanced assessments. The program implements marginal maximum likelihood (MML) estimation techniques for items and MLE estimation of theta.

The software program **STUIRT** (Kim & Lee, 2004) was used to conduct the vertical linking and horizontal linking in the item-pool calibration step. **STUIRT** implements four IRT scale transformation methods using the mean/sigma, mean/mean Haebara (1980) and Stocking-Lord (1983) methods. Consistent with previous research, the Stocking-Lord and Haebara methods are expected to have highly similar results (Hanson & Beguin, 2002). The Stocking-Lord transformation constants consisting of the slope (A) and intercept (B) terms were estimated and then applied to targeted item parameter estimates to place them onto the common vertical scale.

Since **PARSCALE** is limited in the types of graphical output, the program **PARPLOT** (ETS, 2009) was used to obtain item characteristic curves used for evaluating item functioning. A useful way to understand item functioning is to examine plots showing the observed and expected performance based on the item-parameter estimates. Figures 3, 4, 5, and 6 show example plots for a dichotomous and a polytomous item that demonstrate items with both good and poor fit. The solid line represents the expected item performance based on IRT, and the triangles represent the observed item performance, with the size of the triangles proportional to student sample size at a given level of theta. For an item to show good model data fit, it is expected that the triangles, especially the large-size ones, adhere closely around the item characteristic curves. Evaluation of item functioning was conducted visually using **PARPLOT** in conjunction with the goodness-of-fit statistic. Based on evaluation of the plots, any items demonstrating poor functioning were flagged and excluded from the calibrated item pool as previously described.

Figure 3. Sample ICC Plot for a Dichotomous Item Demonstrating Good Fit

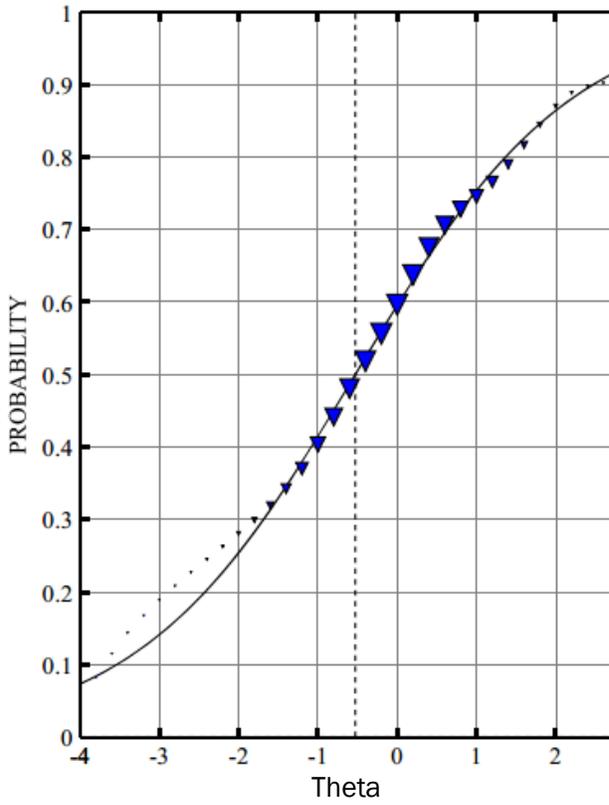


Figure 4. Sample ICC Plot for a Dichotomous Item Demonstrating Poor Fit

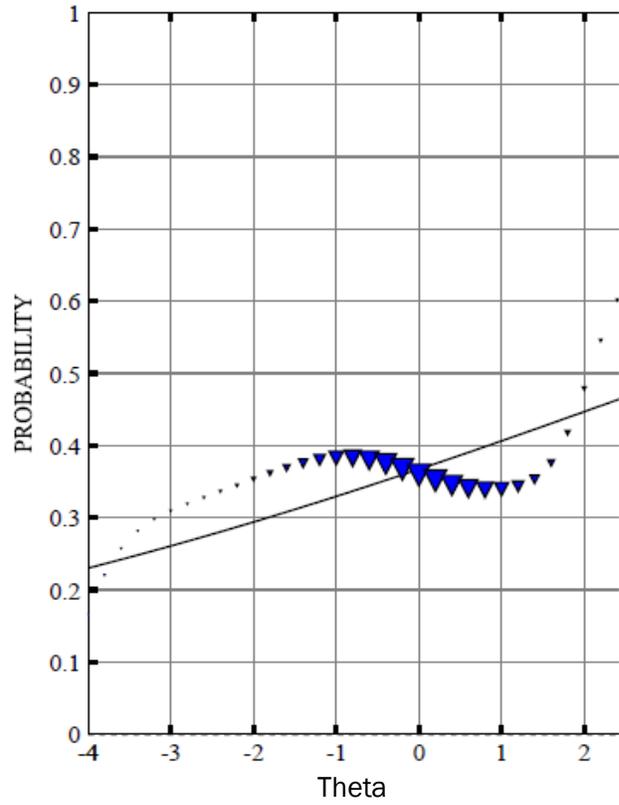


Figure 5. Sample ICC Plot for a Polytomous Item Demonstrating Good Fit

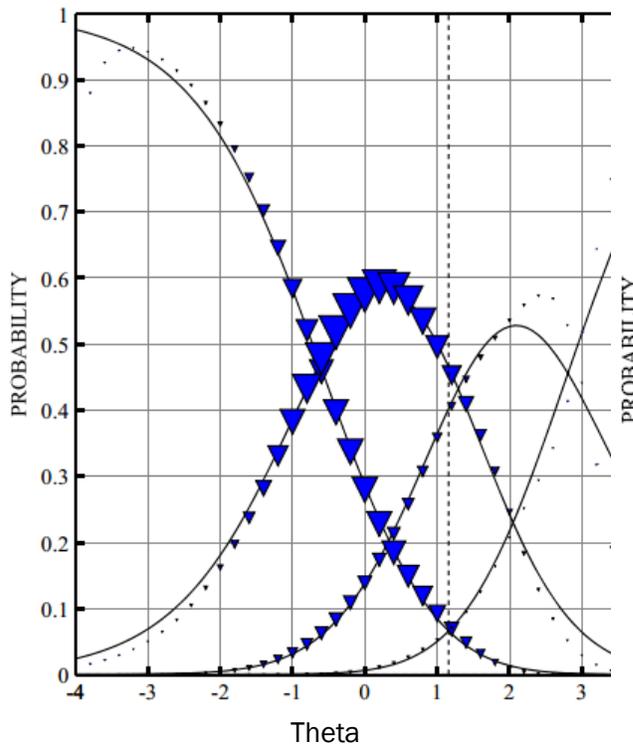
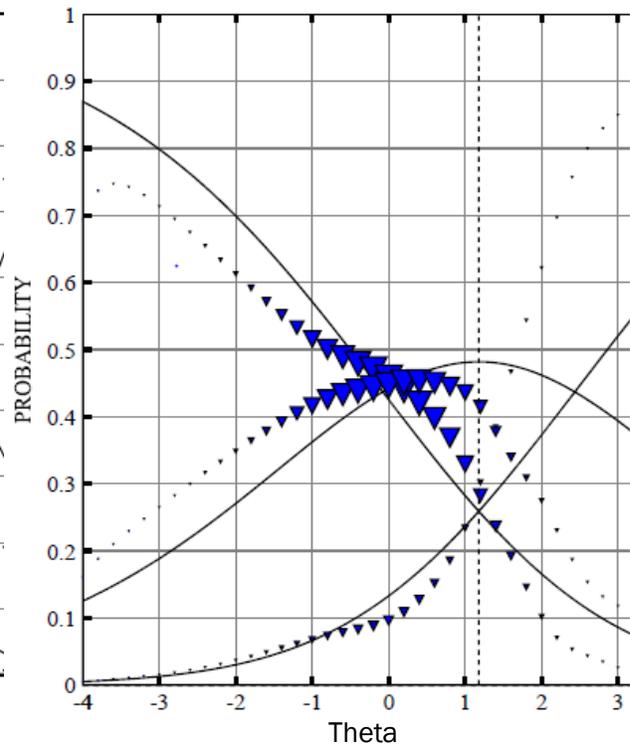


Figure 6. Sample ICC Plot for a Polytomous Item Demonstrating Poor Fit



Item Fit. The usefulness of IRT models is dependent on the extent to which they effectively reflect the data. Assessing fit in item response models usually involves validating assumptions underlying the models and evaluating goodness-of-fit, which specifically refers to how effectively the model describes the outcome data. IRT fit evaluation was conducted for calibrations using the two-parameter-logistic/generalized partial credit model (2PL/GPC) combination. The goodness-of-fit information contained in **PARSCALE** uses the likelihood ratio χ^2 test.

The likelihood ratio χ^2 test statistic can be used to compare the frequencies of correct and incorrect responses in the intervals on the θ continuum with those expected based on the fitted model (du Toit, 2003)

$$\chi_j^2 = 2 \sum_{h=1}^{n_g} \left[r_{hj} \log_e \left\{ \frac{r_{hj}}{N_h P_j(\bar{\theta}_h)} \right\} + (N_h - r_{hj}) \log_e \left\{ \frac{N_h - r_{hj}}{N_h [1 - P_j(\bar{\theta}_h)]} \right\} \right],$$

where n_g is the total number of intervals, r_{hj} is the observed frequency of correct responses to item j in interval h , N_h is the number of examinees in interval h , $\bar{\theta}_h$ is the average ability of examinees in interval h , and $P_j(\bar{\theta}_h)$ is the value of the fitted response function for item j at $\bar{\theta}_h$. The residuals are not under linear constraints, and there is no loss of degrees of freedom due to fitting the item parameters. The number of degrees of freedom is equal to the number of intervals remaining after neighboring intervals are merged, if necessary, to avoid expected values less than 5. Chi-square-type statistics tend to be sensitive to sample size (i.e., flagging more items with large sample size). Item

fit was evaluated in conjunction with other psychometric criteria and the plots described previously. No items were excluded based solely on fit.

Vertical Linking Via Stocking-Lord. The Stocking-Lord method was used as the primary method of linking adjacent grade levels to construct the vertical scale. In general, test-characteristic-curve methods such as the Stocking-Lord method have some advantages when compared to moment methods such as mean/mean or mean sigma (Baker & Al-Karni, 1991; Hanson & Béguin, 2002; Kolen & Brennan, 2004). When used with separate calibration, the test-characteristic-curve methods are more robust to violation of the IRT assumptions and produce less error when compared with moment methods. The Stocking-Lord procedure minimizes the sum of the squared differences over students between the target and reference test characteristic curves based on common items. Specifically, the procedure seeks to determine the slope (A) and intercept (B) that minimize the function

$$E = \frac{1}{N} \sum_{a=1}^N [T(\theta_a) - T^*(\theta_a)]^2,$$

where $T(\theta_a)$ is the test characteristic curve of linking items on the reference vertical scale and $T^*(\theta_a)$ is the test characteristic curve of linking items from the grade to be transformed onto the vertical scale. The linking takes place by applying the resulting slope and intercept to the targeted item parameters.

For 2-PL and GPC models, the transformations for discrimination and difficulty parameter estimates are

$$a^T = \frac{a}{A};$$

$$b^T = A \cdot b + B,$$

and for GPC model item-category parameters, the transformation is

$$d^T = A \cdot d.$$

The following transformations are applied to theta (ability)

$$\hat{\theta}^T = A \cdot \hat{\theta} + B.$$

The associated standard errors for the parameter estimates were transformed as follows

$$s.e.(\hat{\theta}^T) = A \cdot s.e.(\hat{\theta})$$

$$s.e.(\hat{b}^T) = A \cdot s.e.(\hat{b})$$

$$s.e.(\hat{a}^T) = s.e.(\hat{a}) / A$$

$$s.e.(\hat{d}^T) = A \cdot s.e.(\hat{d}).$$

The **STUIRT** program was used to implement the vertical linking using test-characteristic methods. An example of the **STUIRT** linking output comparing the different methods is shown in Table 6 for linking grade 3 ELA/literacy to grade 4. In this example, all methods produced similar slope and intercept

values. To implement the Stocking-Lord linking, the weights and quadrature points of the latent ability distribution output from **PARSCALE** were used. These quadrature points were transformed the same way as student abilities. The slope and intercept transformation parameters (A & B) were applied to the latent distributions produced by **PARSCALE** in each grade. **STUIRT** was also used to conduct the linking in Phase 2, where horizontal scaling of the remaining on-grade item pool was conducted.

Evaluation of Vertical Anchor Item Stability. An inspection of the differences between the off-grade estimates and the reference, on-grade ones for each vertical linking item was conducted. The weighted root mean squared difference (WRMSD) is calculated as

$$WRMSD = \sqrt{\sum_{j=1}^{N_g} w_j [P_n(\hat{\theta}_j) - P_r(\hat{\theta}_j)]^2},$$

where abilities are grouped in the intervals of 0.5 between -4.0 and 4.0 , $\hat{\theta}_j$ is the mean of the abilities in the interval j , N_g is the number of intervals, w_j is a weight equal to the proportion of estimated abilities from the transformed new form in interval j , $P_n(\hat{\theta}_j)$ is the probability of correct response based on the transformed new-item-parameter estimates at ability level $\hat{\theta}_j$, and $P_r(\hat{\theta}_j)$ is the probability of correct response at ability level $\hat{\theta}_j$ based on the reference-item-parameter estimates. A criterion of WRMSD greater than 0.125 was used to evaluate the linking. This criterion has produced reasonable results in other programs in year-to-year horizontal-equating contexts (Gu, Lall, Monfils, & Jiang, 2010). The distributions of WRMSD were evaluated; no linking items were eliminated based on the WRMSD statistic.

Table 6. Example of STUIRT Linking Methods and Output.

Method	Slope A	Intercept B
Mean/Mean	0.9627	-1.1864
Mean/Sigma	0.9585	-1.1823
Haebara	0.9533	-1.1683
Stocking-Lord	0.9444	-1.1889

Vertical Scale Evaluation. In the process of constructing the vertical scale, it was evaluated using a number of methods that included:

- correlation and plots of (common) item difficulties across grade levels;
- progression in test difficulty across grades;
- comparison of mean scale scores across grades;
- comparison of scale scores associated with proficiency levels across grades;
- comparison of overlap/separation of proficiency distributions across grades; and

- comparison of variability in scale scores (ability) within and across grades comparing scale score standard deviations.

Grade-to-grade change can be displayed as the differences between means and percentiles (10, 25, 50, 75, and 90) across grades. Separation of ability distributions can also be displayed by plotting the scale score cumulative distributions across grades. An index of separation in grade distributions suggested by Yen (1986) is the effect size. It standardizes the grade-to-grade difference in the means by the square root of the average of the within-grade variances. The effect size is defined as

$$\frac{\bar{\theta}_{higher} - \bar{\theta}_{lower}}{\sqrt{\frac{\sigma_{higher}^2 + \sigma_{lower}^2}{2}}}$$

where $\bar{\theta}_{higher}$ is the average ability estimate for the higher grade level, $\bar{\theta}_{lower}$ is the average ability estimate for the lower grade level, σ_{higher}^2 is the variance of the ability estimates for the higher grade, and σ_{lower}^2 is the variance of the ability estimates for the lower one.

Horizontal and Vertical Scaling Results

During classical item analysis, the performance of on-grade and vertical linking items was evaluated by comparing the item difficulty across the two adjacent grades. For a vertical sale to demonstrate change in achievement and be plausible, items are expected to be easier in the higher grade. For example, when an item is administered in grades 5 and 6, the p -value should be relatively higher (easier) in grade 6. The on- and off-grade average item difficulty and item-test correlations are given in Tables 7 and 8 for ELA/literacy and mathematics, respectively. Item difficulty is defined as the percentage of the maximum possible raw score. The average item difficulty is consistent with the notion of better performance in the higher grade-level necessary to establish a vertical scale. Since the tests differed widely in the number of items delivered, theta was used as the criterion rather than the typical total raw score for the item-test correlation.

The distributions (i.e., the five-number summary) for the IRT item discrimination and location parameters that resulted from the initial horizontal calibration in the vertical scaling are given in Table 9. The difficulty parameters (i.e., b -parameters) indicate that the tests were difficult, particularly in high school. Figures 7 and 8 present plots of chi-square item fit for ELA/literacy and mathematics. In general, there were only a relatively small number of outliers. Table 10 provides a summary of the χ^2 statistics and sample size ranges per item. Tables 11 to 22 show the distributions of untransformed item a - and b -parameter estimates from common items for ELA/literacy and mathematics, respectively. Figures 9 to 14 for ELA/literacy and 15 to 20 for mathematics show plots of untransformed, vertical linking, item parameter estimates across grades. The item parameter estimates for the most part cluster along the diagonal. As a rule-of-thumb, the b -parameter estimate correlations should typically be above .90 and the a -parameter estimates above .85, which indicate the items behaved consistently across grades. The distributions of the untransformed a - and b -parameter estimates are given for the common, vertical linking items across grades, and these parameters are plotted along with the correlations.

After the conclusion of the horizontal scaling and IRT item exclusion steps, the vertical scaling was conducted. Using grade 6 as the base and the common items across grade levels, each grade level was successively linked onto the vertical scale using the associated Stocking-Lord transformation constants. Table 23 presents the Stocking-Lord transformation constants that were obtained from **STUIRT**. For evaluative purposes, the WRMSD was computed, and histograms of the resulting values

were plotted in Figures 21 and 22 for each vertical linking set. To construct a criterion for evaluation, a vertical line plotted at 0.125 was used as a criterion for identifying items with large values. Consistent with the high correlations between linking items, the values WRMSD were for the most part below 0.10. This information was used diagnostically to evaluate the linking, and no items were removed based on the WRMSD.

The ability or theta estimates used were maximum likelihood estimates (MLEs) produced by **PARSCALE**. In cases in which an MLE could not be produced by **PARSCALE**, table driven sufficient statistics (Lord, 1980) were used to derive a theta estimate. Table 24 summarizes the resulting theta distribution for the ELA/literacy and mathematics vertical scales. It presents the five-number summaries, the means and standard deviations, and sample sizes along with the effect sizes. The effect size demonstrates the degree of change over grades and is not uniform with a larger change observed in the lower grades. Figures 23 and 24 display the cumulative distributions of ability (theta) for the vertical scale. The cumulative distributions of ability are more widely separated at the lower-grade levels, with diminishing amounts of change in the upper-grade levels in both ELA/literacy and mathematics.

Table 7. Summary of Classical Statistics by Purpose for ELA/literacy.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
On-grade	Number of Items	261	242	256	232	238	243	410
	Mean Difficulty	0.34	0.35	0.38	0.35	0.34	0.36	0.34
	Item-Total Correlation	0.51	0.50	0.52	0.49	0.49	0.49	0.47
Off-grade (Vertical Linking)	Number of Items		120	133	131	107	123	107
	Mean Difficulty		0.45	0.45	0.42	0.36	0.38	0.36
	Item-Total Correlation		0.54	0.52	0.52	0.51	0.51	0.49

Table 8. Summary of Classical Statistics by Purpose for Mathematics.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
On-grade	Number of Items	304	306	306	222	239	230	319
	Mean Difficulty	0.39	0.36	0.32	0.30	0.27	0.24	0.24
	Item-Total Correlation	0.59	0.58	0.56	0.60	0.59	0.53	0.53
Off-grade (Vertical Linking)	Number of Items		104	95	102	71	73	81
	Mean Difficulty		0.51	0.40	0.37	0.32	0.31	0.32
	Item-Total Correlation		0.62	0.61	0.58	0.62	0.59	0.56

Table 9. Summary of Item Parameter Estimates for Horizontal Calibration Step.

Grade	ELA/literacy							Mathematics						
	3	4	5	6	7	8	HS	3	4	5	6	7	8	HS
No. of Items	261	362	389	363	345	366	517	304	410	401	324	310	303	400
a-parameter														
Mean	0.598	0.594	0.599	0.580	0.577	0.577	0.546	0.756	0.765	0.752	0.736	0.814	0.749	0.727
SD	0.215	0.207	0.204	0.208	0.207	0.222	0.208	0.253	0.255	0.296	0.274	0.360	0.329	0.318
Min	0.117	0.171	0.138	0.165	0.164	0.119	0.129	0.149	0.206	0.200	0.126	0.103	0.146	0.126
10%	0.331	0.326	0.342	0.320	0.307	0.299	0.283	0.431	0.437	0.405	0.380	0.295	0.338	0.351
25%	0.430	0.451	0.451	0.433	0.429	0.419	0.405	0.563	0.577	0.537	0.555	0.513	0.477	0.478
Median	0.578	0.586	0.597	0.553	0.556	0.564	0.540	0.769	0.751	0.708	0.723	0.835	0.735	0.693
75%	0.723	0.736	0.735	0.719	0.723	0.722	0.675	0.941	0.952	0.936	0.923	1.103	0.971	0.929
90%	0.897	0.854	0.867	0.856	0.843	0.870	0.795	1.075	1.102	1.167	1.100	1.241	1.179	1.161
Max	1.186	1.317	1.222	1.320	1.249	1.434	1.349	1.379	1.457	1.816	1.587	2.053	1.831	1.885
b-parameter														
Mean	1.060	0.779	0.637	0.817	0.996	0.923	1.110	0.588	0.504	0.783	0.852	1.128	1.335	1.397
SD	1.256	1.295	1.187	1.251	1.220	1.364	1.358	1.261	1.134	1.021	1.190	1.140	1.205	1.172
Min	1.707	2.587	3.101	2.025	2.196	3.183	2.019	3.261	2.987	1.909	4.064	2.522	1.893	2.425
10%	0.528	0.757	0.879	0.814	0.478	0.721	0.577	0.985	1.017	0.518	0.794	0.278	0.172	0.103
25%	0.073	0.246	0.244	0.069	0.093	0.084	0.122	0.361	0.265	0.113	0.075	0.486	0.458	0.639
Median	0.971	0.671	0.618	0.753	0.969	0.925	0.977	0.701	0.552	0.831	0.923	1.187	1.390	1.445
75%	2.028	1.613	1.459	1.699	1.754	1.713	1.904	1.362	1.290	1.518	1.696	1.808	2.052	2.164
90%	2.773	2.552	2.247	2.347	2.569	2.772	2.946	2.208	1.956	2.074	2.297	2.228	2.952	2.790
Max	4.745	4.897	4.296	4.479	4.720	4.875	5.781	4.865	3.562	4.700	4.139	5.008	4.810	4.372

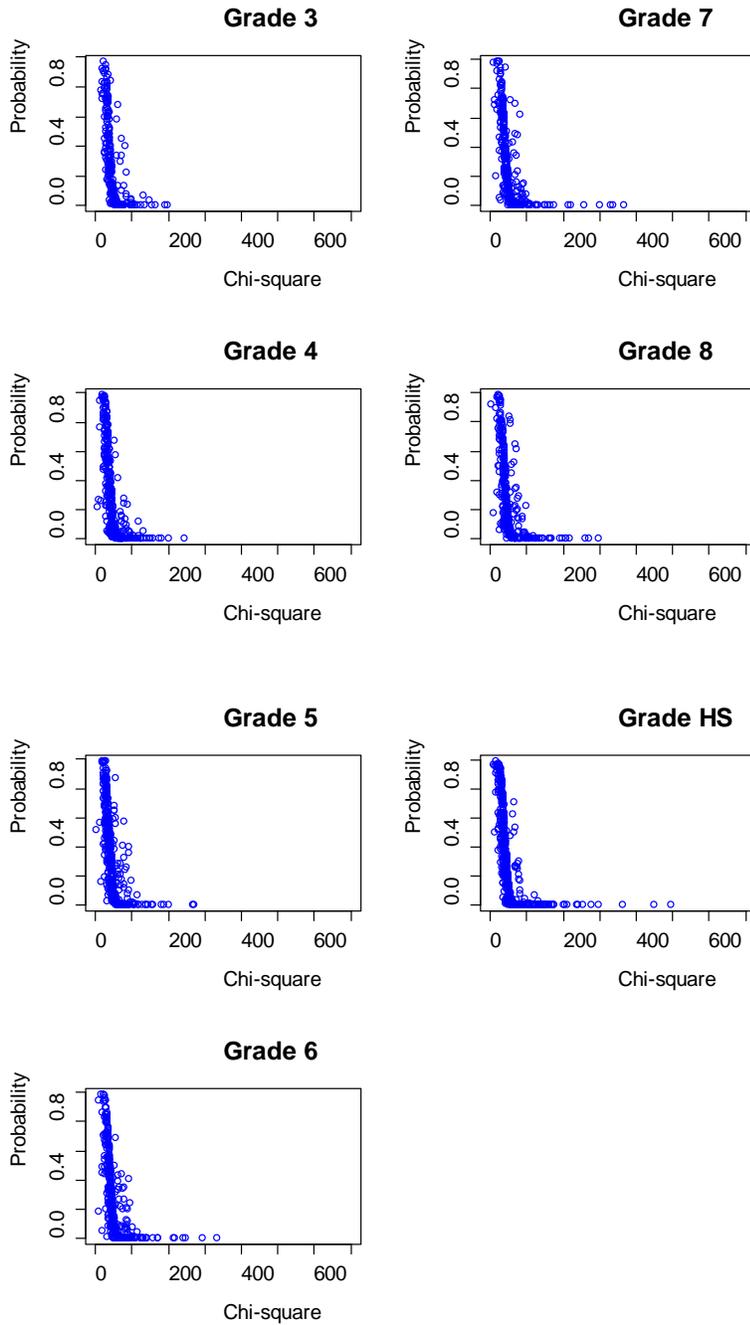


Figure 7. ELA/literacy Item Fit Chi-Square Plots (Vertical Scaling)

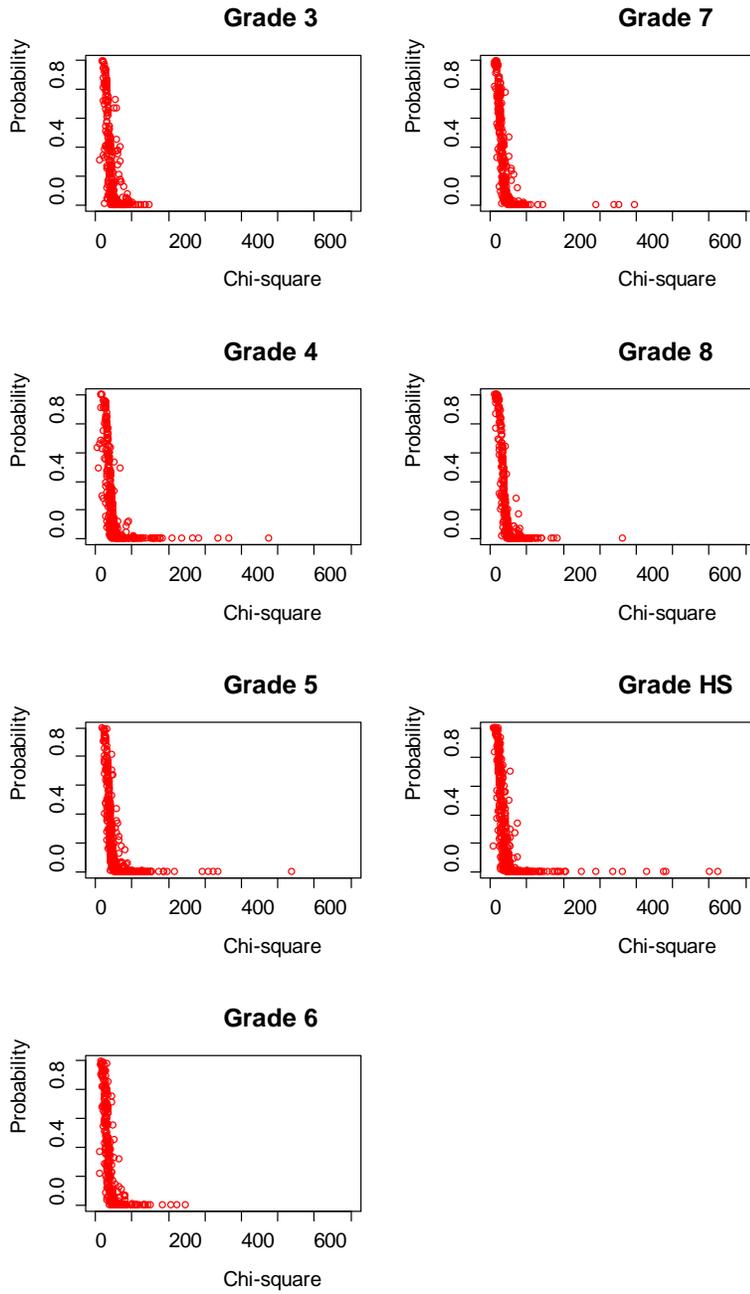


Figure 8. Mathematics Item Fit Chi-Square Plots (Vertical Scaling)

Table 10. Summary of Likelihood Ratio χ^2 Test Statistics by Grade and Content Area.

Grade	No. of Students per Item (Range)	Mean	SD	Min.	Max.	Prob.<.05	Prob.<.01
ELA/literacy							
3	530 - 9,804	50	27	15	199	68	41
4	550 - 22,291	51	29	7	244	102	57
5	759 - 10,853	50	30	3	270	78	44
6	527 - 13,746	56	37	9	335	87	53
7	509 - 20,748	57	42	9	367	83	54
8	508 - 12,981	57	37	1	297	104	57
HS	526 - 16,646	58	47	9	497	156	98
Mathematics							
3	693 - 5,952	49	21	14	150	106	61
4	519 - 13,845	60	46	7	475	155	107
5	502 - 19,614	62	47	19	538	166	119
6	509 - 7,722	47	31	11	247	98	71
7	502 - 10,188	44	40	12	395	97	49
8	536 - 13,005	52	33	12	364	119	79
HS	501 - 14,521	56	67	9	626	122	78

Table 11. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 3 to 4.

No. of Items = 120	<i>a</i> -parameter		<i>b</i> -parameter	
	3	4	3	4
Mean	0.62	0.63	0.87	0.35
SD	0.21	0.20	1.14	1.12
Min	0.12	0.17	-1.09	-1.54
10%	0.36	0.39	-0.55	-1.06
25%	0.47	0.51	-0.11	-0.55
Median	0.63	0.61	0.86	0.32
75%	0.73	0.74	1.58	1.07
90%	0.91	0.87	2.29	1.87
Max	1.19	1.32	3.76	3.22

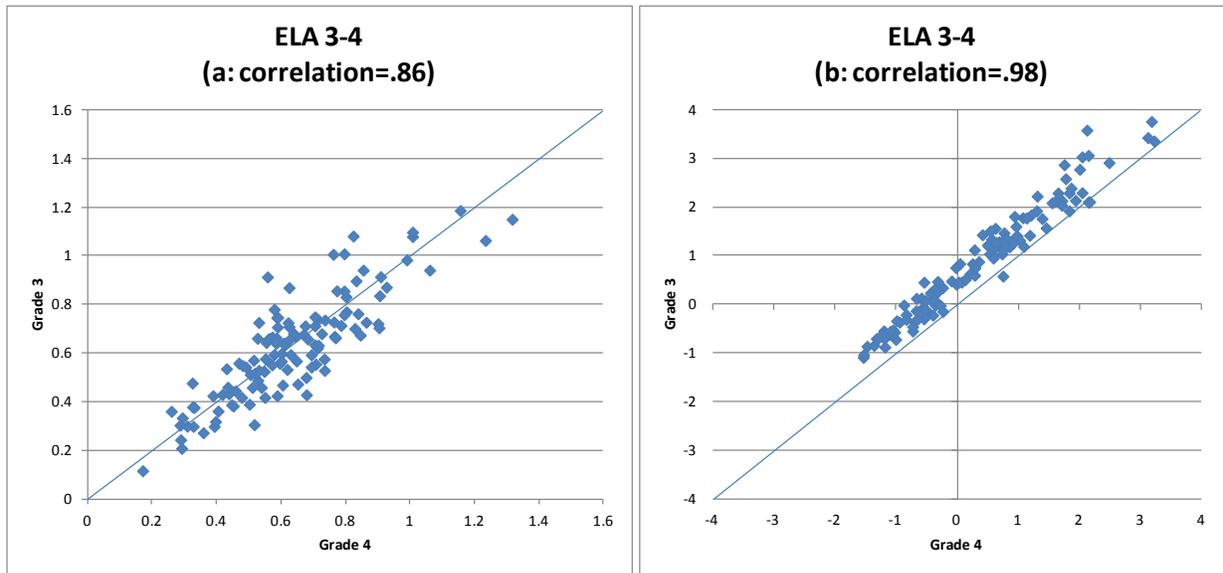


Figure 9. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 3 to 4

Table 12. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 4 to 5.

No. of Items = 133	<i>a</i> -parameter		<i>b</i> -parameter	
	4	5	4	5
Mean	0.60	0.60	0.79	0.37
SD	0.21	0.19	1.15	1.14
Min	0.22	0.18	-2.33	-3.10
10%	0.31	0.36	-0.63	-0.89
25%	0.43	0.45	0.08	-0.31
Median	0.59	0.62	0.63	0.21
75%	0.76	0.73	1.50	1.07
90%	0.84	0.82	2.52	2.18
Max	1.13	1.17	3.61	3.28

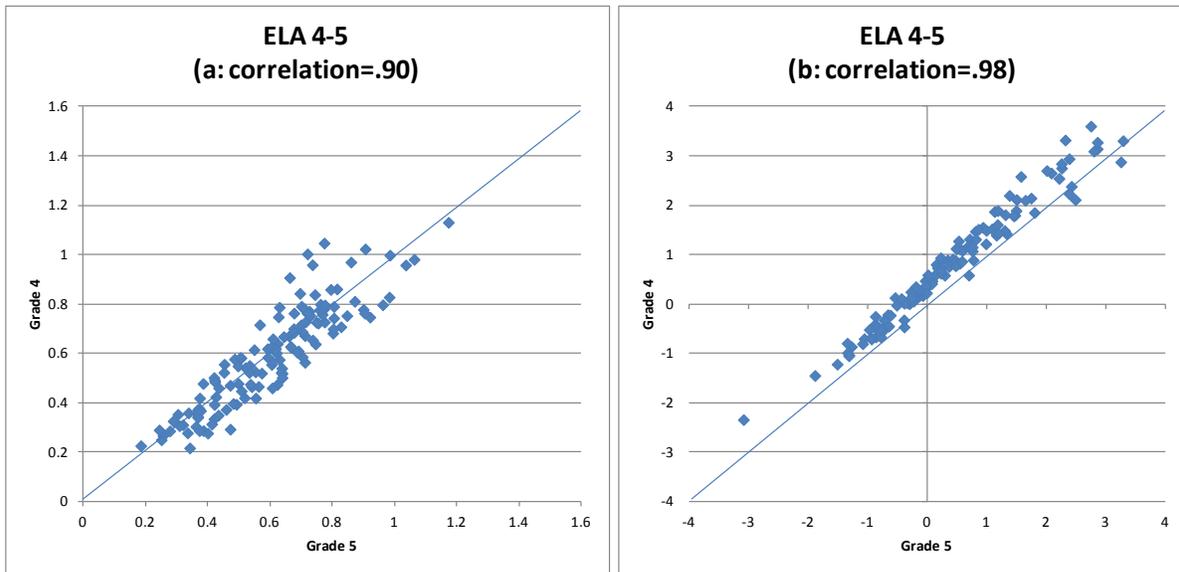


Figure 10. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 4 to 5

Table 13. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 5 to 6.

No. of Items = 131	<i>a</i> -parameter		<i>b</i> -parameter	
	5	6	5	6
Mean	0.61	0.60	0.75	0.53
SD	0.21	0.22	1.17	1.16
Min	0.14	0.17	-2.07	-2.00
10%	0.35	0.36	-0.82	-0.98
25%	0.46	0.46	-0.04	-0.24
Median	0.60	0.57	0.66	0.49
75%	0.74	0.72	1.69	1.52
90%	0.87	0.87	2.29	2.06
Max	1.22	1.32	3.25	3.00

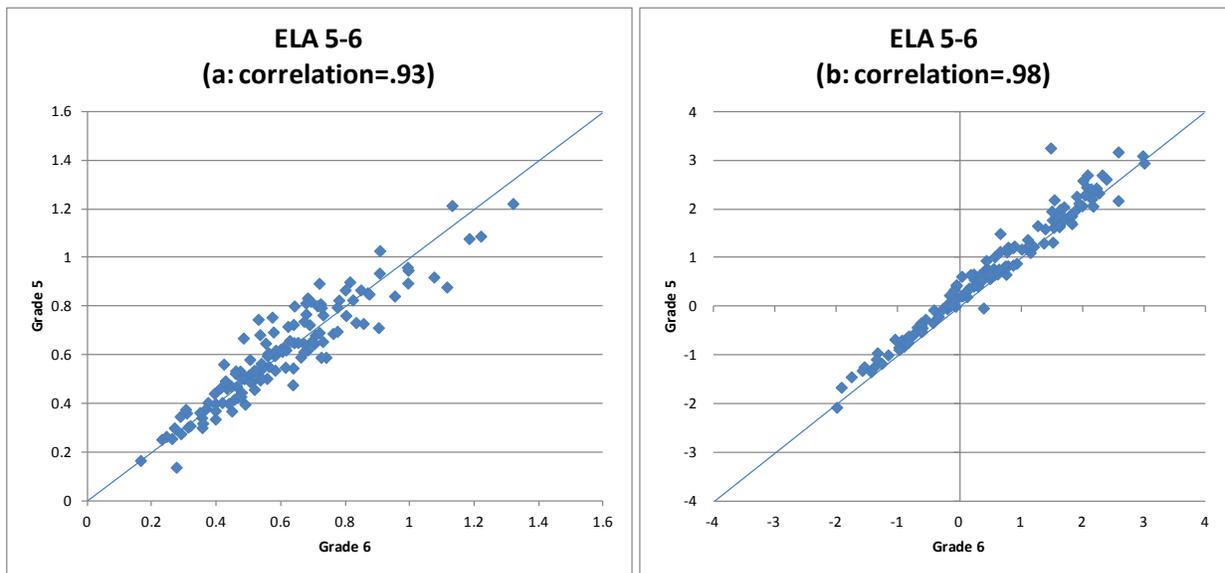


Figure 11. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 5 to 6

Table 14. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 7 to 6.

No. of Items = 107	<i>a</i> -parameter		<i>b</i> -parameter	
	6	7	6	7
Mean	0.59	0.60	1.06	0.84
SD	0.20	0.19	1.22	1.13
Min	0.18	0.21	-1.88	-2.02
10%	0.34	0.32	-0.41	-0.60
25%	0.44	0.46	0.19	0.07
Median	0.57	0.59	1.15	0.96
75%	0.73	0.73	1.86	1.58
90%	0.85	0.83	2.60	2.18
Max	1.01	1.09	4.48	3.50

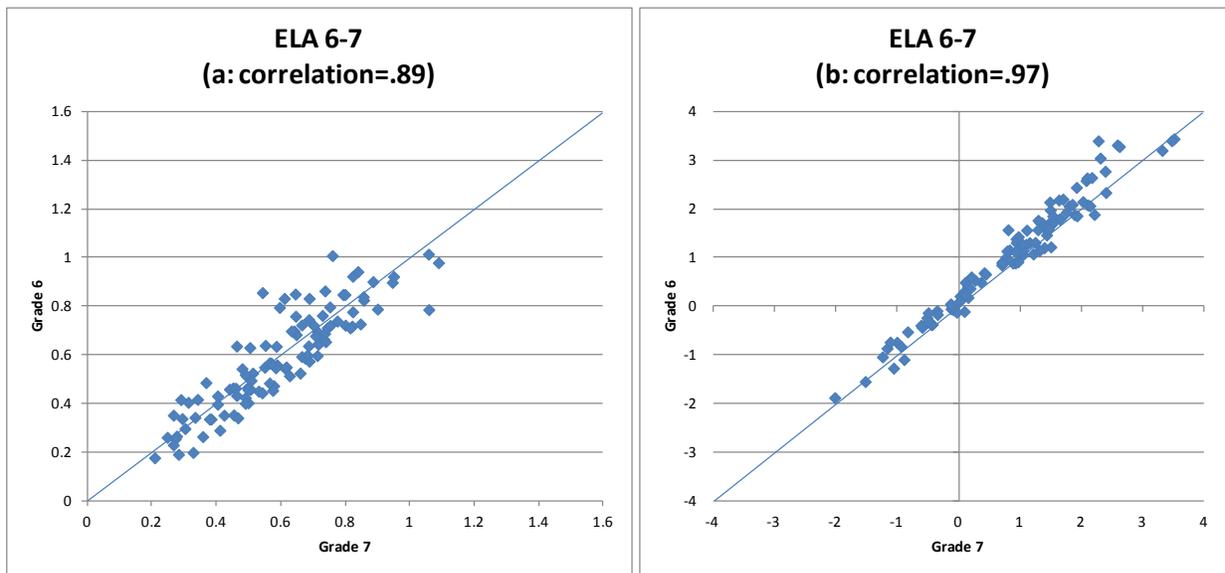


Figure 12. Comparison of ELA/literacy *a*- and *b*-parameter estimates for linking Grade 7 to 6

Table 15. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: Grade 8 to 7.

No. of Items = 123	<i>a</i> -parameter		<i>b</i> -parameter	
	7	8	7	8
Mean	0.59	0.59	1.01	0.77
SD	0.21	0.20	1.23	1.23
Min	0.19	0.19	-2.20	-2.42
10%	0.32	0.34	-0.34	-0.65
25%	0.44	0.44	0.09	-0.16
Median	0.56	0.56	0.98	0.79
75%	0.72	0.70	1.83	1.52
90%	0.85	0.88	2.53	2.12
Max	1.22	1.09	4.61	4.16

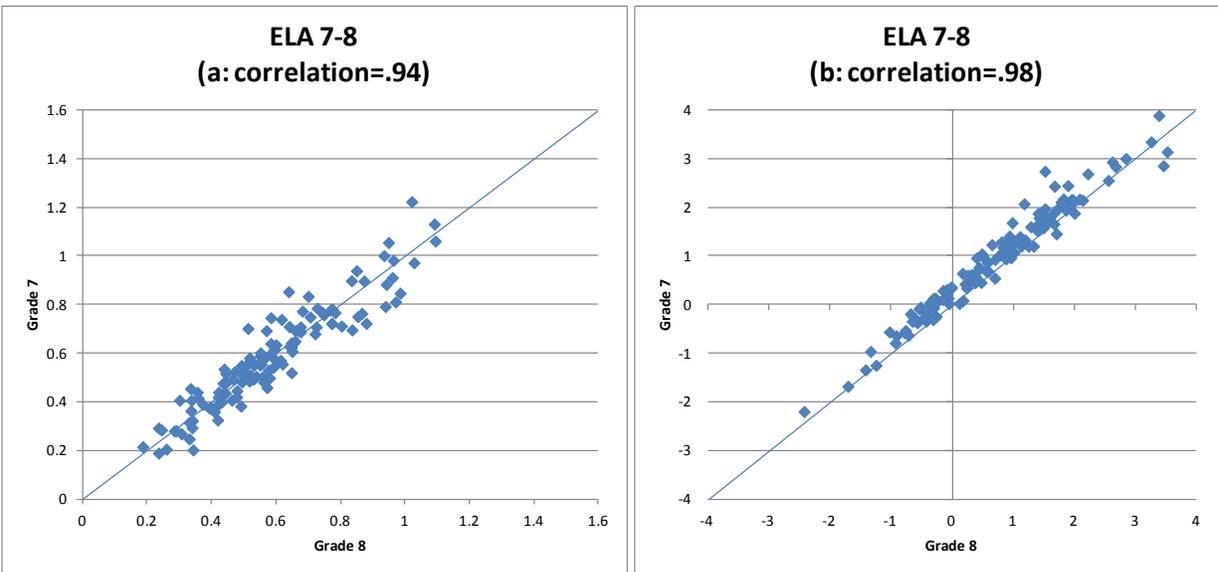


Figure 13. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking Grade 8 to 7

Table 16. Distribution of ELA/literacy Vertical Linking Untransformed Parameter Estimates: High School to Grade 8.

No. of Items = 107	<i>a</i> -parameter		<i>b</i> -parameter	
	8	HS	8	HS
Mean	0.57	0.60	1.16	0.93
SD	0.25	0.26	1.43	1.27
Min	0.12	0.13	-2.01	-2.02
10%	0.28	0.30	-0.58	-0.69
25%	0.38	0.40	0.05	0.03
Median	0.57	0.59	1.05	0.86
75%	0.70	0.77	2.09	1.63
90%	0.88	0.94	3.18	2.57
Max	1.43	1.35	4.23	4.20

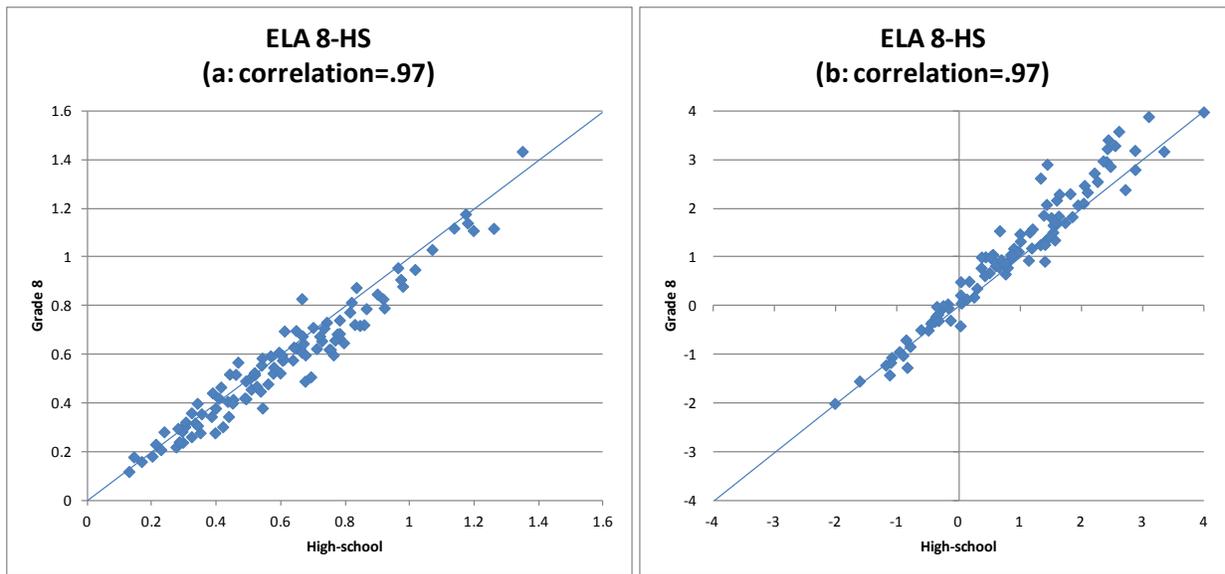


Figure 14. Comparison of ELA/literacy *a*- and *b*-parameter estimates for Linking High School to Grade 8

Table 17. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 3 to 4.

No. of Items = 104	<i>a</i> -parameter		<i>b</i> -parameter	
	3	4	3	4
Mean	0.77	0.81	0.61	-0.07
SD	0.25	0.23	1.30	1.17
Min	0.15	0.26	-2.17	-2.88
10%	0.44	0.52	-0.97	-1.49
25%	0.58	0.63	-0.42	-0.92
Median	0.77	0.81	0.65	-0.07
75%	0.95	1.00	1.39	0.67
90%	1.09	1.09	2.23	1.50
Max	1.38	1.29	4.83	3.47

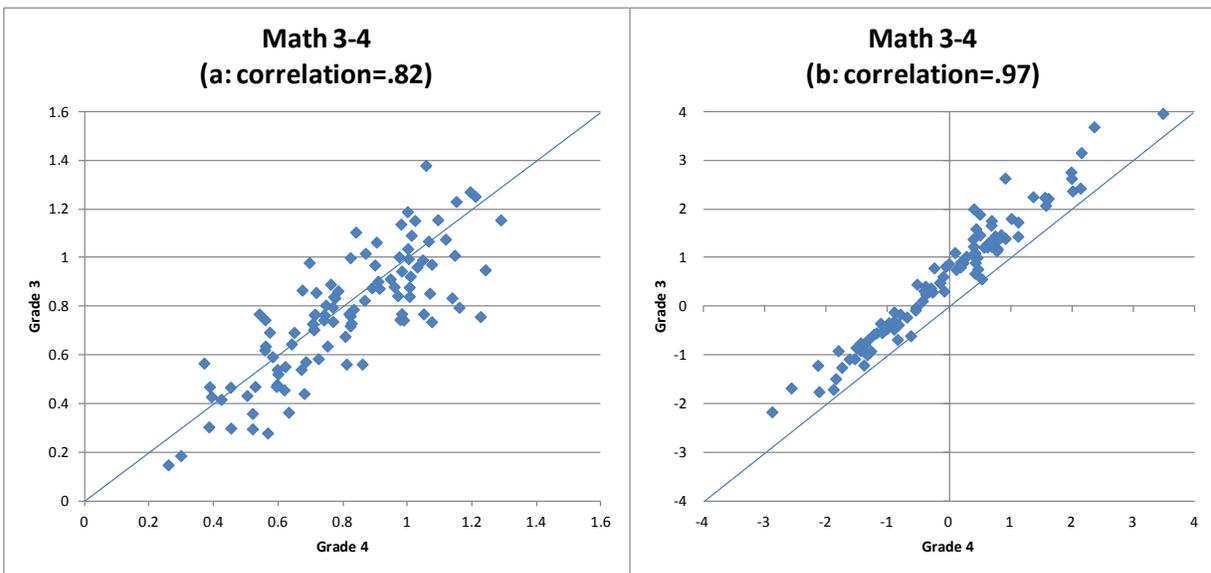


Figure 15. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 3 to 4

Table 18. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 4 to 5.

No. of Items = 95	<i>a</i> -parameter		<i>b</i> -parameter	
	4	5	4	5
Mean	0.77	0.81	0.87	0.45
SD	0.26	0.25	0.98	0.95
Min	0.33	0.35	-1.73	-1.81
10%	0.42	0.49	-0.52	-0.87
25%	0.58	0.65	0.17	-0.32
Median	0.75	0.79	0.89	0.43
75%	0.94	0.94	1.57	1.24
90%	1.11	1.16	2.17	1.65
Max	1.46	1.49	2.88	2.80

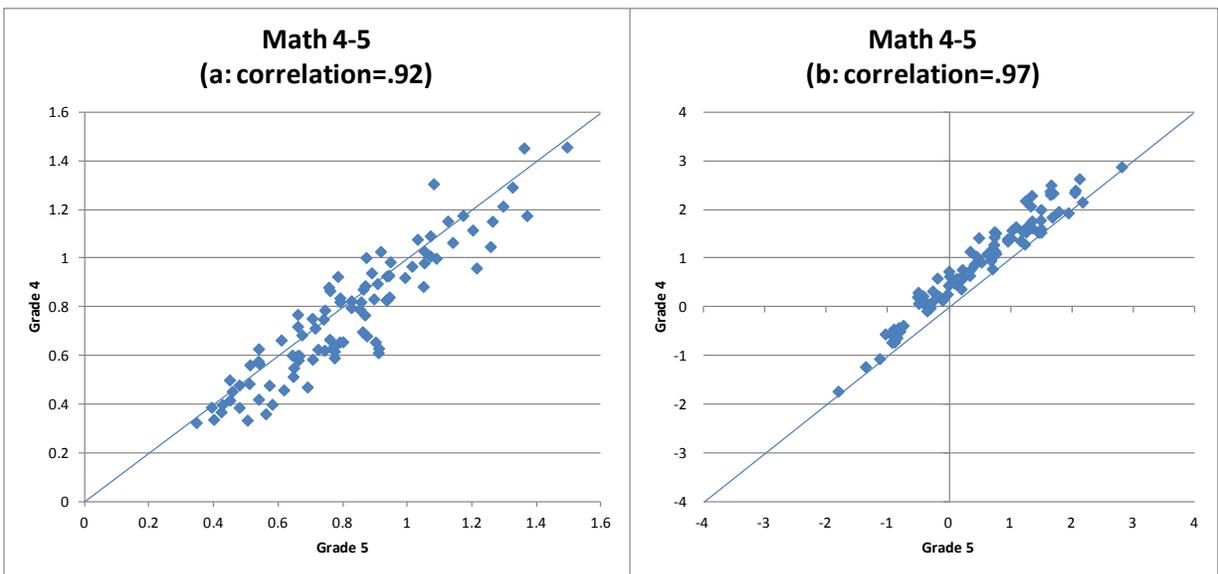


Figure 16. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 4 to 5

Table 19. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 5 to 6.

No. of Items = 102	<i>a</i> -parameter		<i>b</i> -parameter	
	5	6	5	6
Mean	0.69	0.69	0.84	0.58
SD	0.28	0.25	1.00	1.05
Min	0.26	0.23	-1.78	-1.95
10%	0.40	0.40	-0.45	-0.76
25%	0.49	0.53	0.19	-0.21
Median	0.62	0.64	0.89	0.58
75%	0.85	0.85	1.40	1.37
90%	1.06	1.01	2.17	1.93
Max	1.61	1.47	3.14	3.82

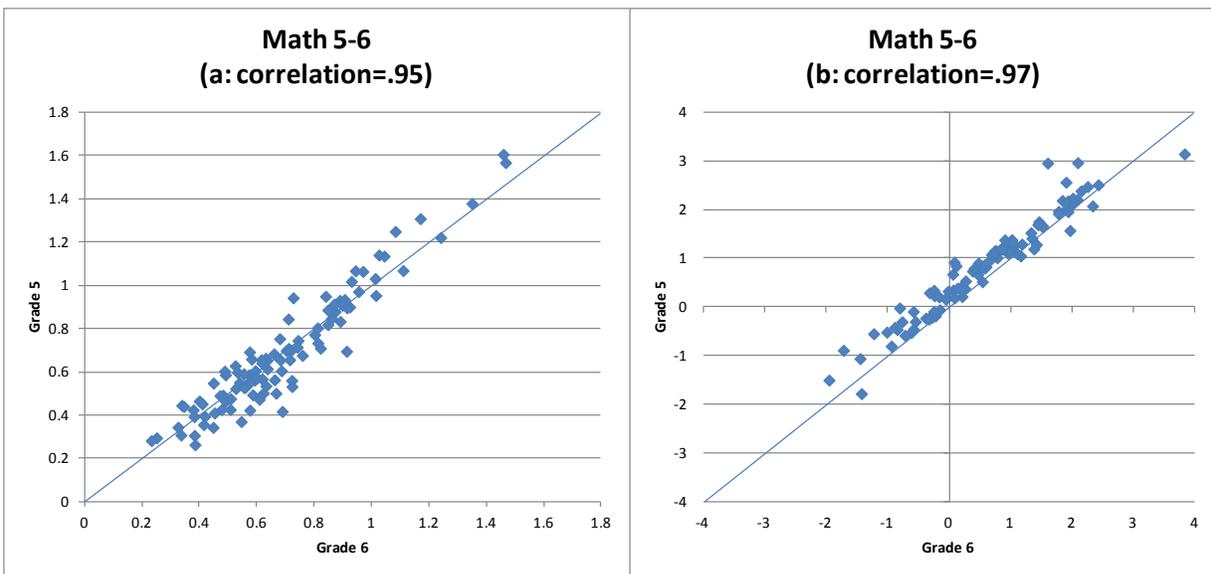


Figure 17. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 5 to 6

Table 20. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 7 to 6.

No. of Items = 71	<i>a</i> -parameter		<i>b</i> -parameter	
	6	7	6	7
Mean	0.76	0.85	1.04	0.77
SD	0.30	0.31	1.11	1.00
Min	0.13	0.20	-1.74	-1.70
10%	0.36	0.48	-0.22	-0.47
25%	0.56	0.65	0.40	0.22
Median	0.76	0.86	1.01	0.74
75%	0.97	1.12	1.87	1.52
90%	1.14	1.22	2.37	1.99
Max	1.59	1.37	3.28	2.62

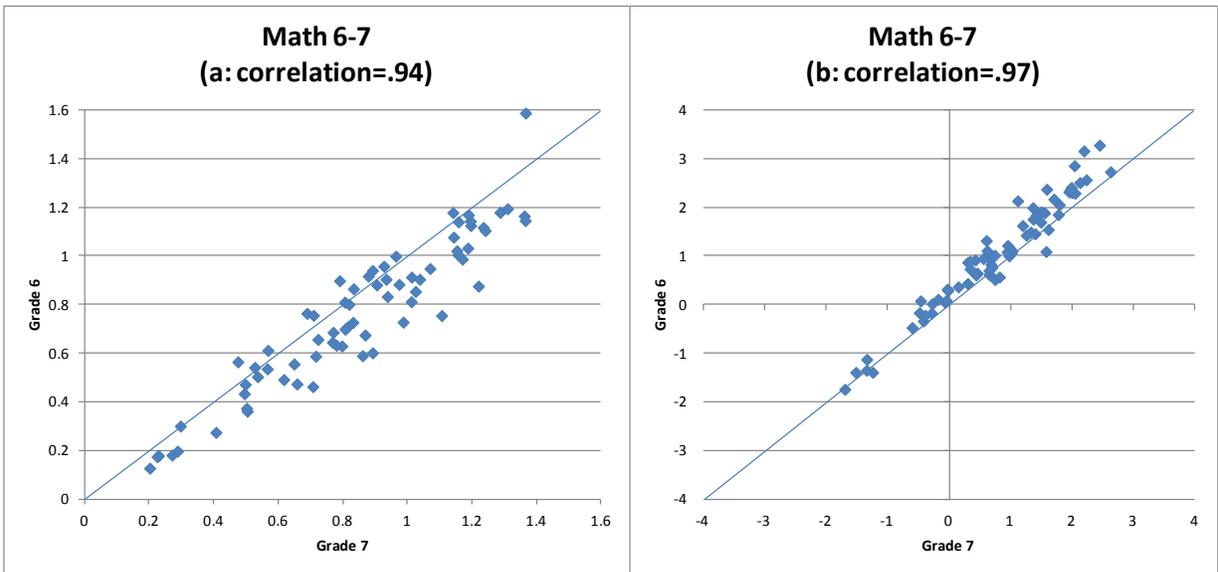


Figure 18. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 7 to 6

Table 21. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: Grade 8 to 7.

No. of Items = 73	<i>a</i> -parameter		<i>b</i> -parameter	
	7	8	7	8
Mean	0.84	0.86	1.18	0.97
SD	0.34	0.34	1.20	1.20
Min	0.22	0.22	-1.66	-1.80
10%	0.34	0.38	-0.17	-0.35
25%	0.59	0.63	0.36	0.10
Median	0.88	0.84	1.06	1.10
75%	1.10	1.14	1.87	1.65
90%	1.23	1.33	2.80	2.41
Max	1.80	1.60	4.33	4.13

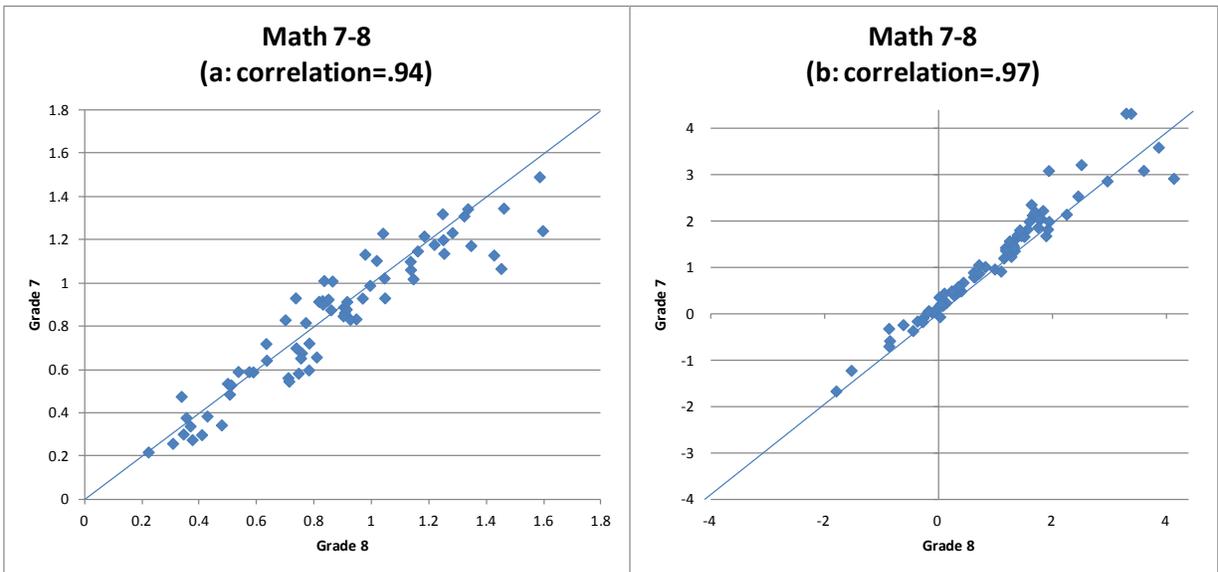


Figure 19. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking Grade 8 to 7

Table 22. Distribution of Mathematics Vertical Linking Untransformed Parameter Estimates: High School to Grade 8.

No. of Items = 81	<i>a</i> -parameter		<i>b</i> -parameter	
	8	HS	8	HS
Mean	0.67	0.76	1.33	0.87
SD	0.31	0.32	1.34	1.12
Min	0.15	0.26	-1.89	-1.47
10%	0.30	0.37	-0.10	-0.43
25%	0.40	0.50	0.46	0.21
Median	0.62	0.76	1.45	0.92
75%	0.93	1.00	2.09	1.50
90%	1.09	1.16	3.33	2.52
Max	1.32	1.48	4.16	3.39

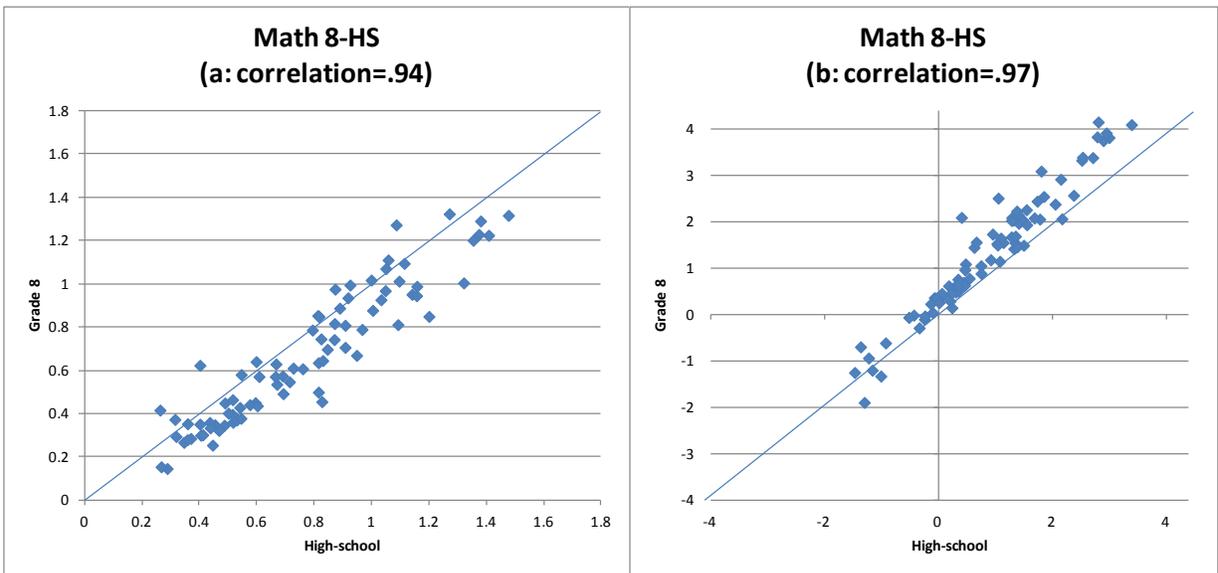


Figure 20. Comparison of Mathematics *a*- and *b*-parameter estimates for Linking High School to Grade 8

Table 23. Vertical Linking Transformation Constants from the Stocking-Lord Procedure.

Grade Pairs	Slope A	Intercept B
ELA/literacy		
3 to 4	0.944421	-1.188941
4 to 5	0.973260	-0.683668
5 to 6	1.002164	-0.256198
6 (Base Grade)		
7 to 6	1.027782	0.172946
8 to 7	1.033673	0.437905
HS to 8	1.105322	0.583021
Mathematics		
3 to 4	0.872487	-1.240565
4 to 5	0.938657	-0.666856
5 to 6	1.004384	-0.279283
6 (Base Grade)		
7 to 6	1.103163	0.147206
8 to 7	1.137342	0.340534
HS to 8	1.311837	0.630426

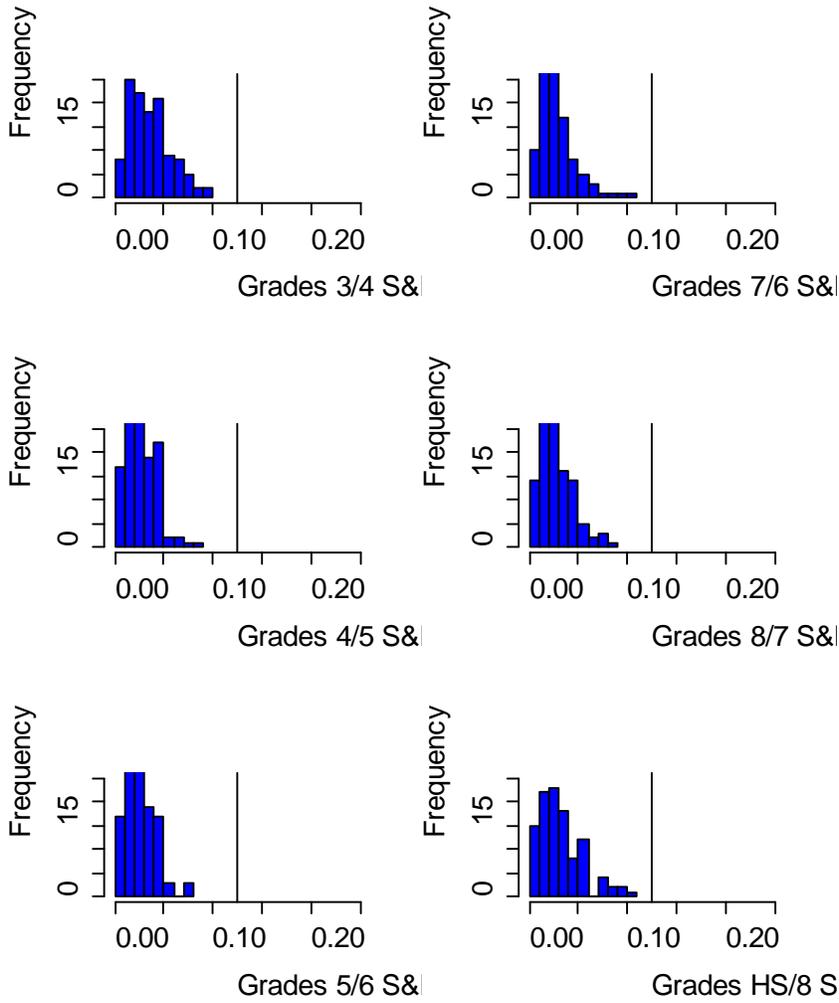


Figure 21. Distribution of WRMSD for ELA/literacy (Vertical Linking Items)

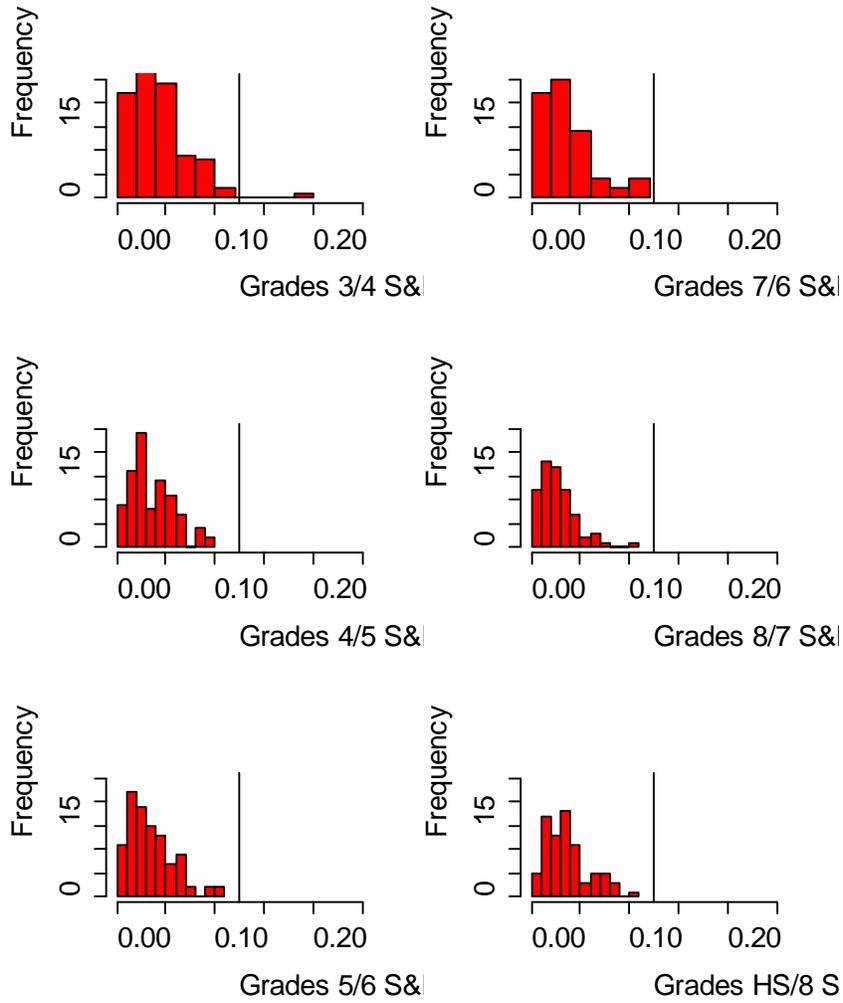


Figure 22. Distribution of WRMSD for Mathematics (Vertical Linking Items)

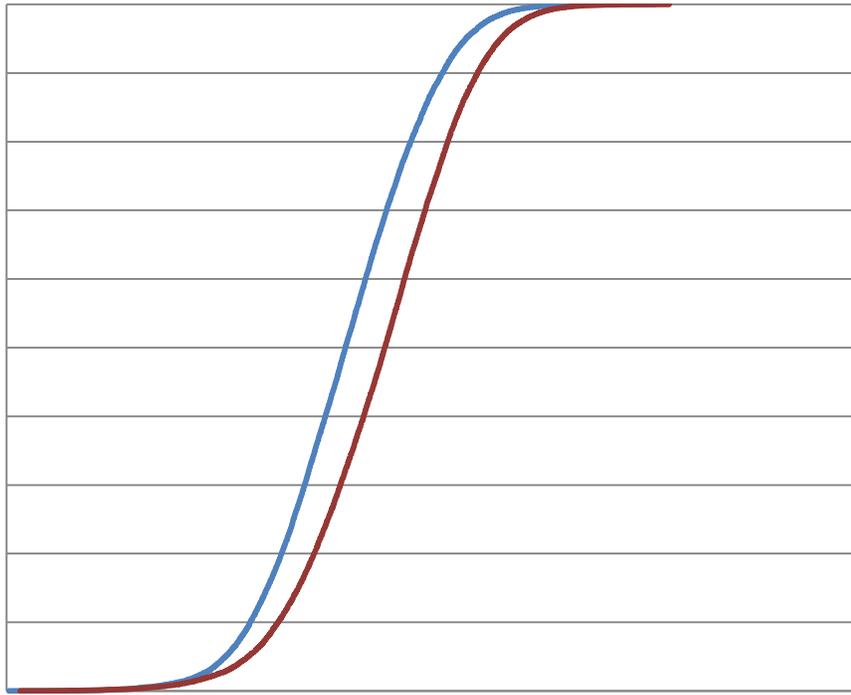


Figure 23. ELA/literacy Cumulative Distributions of Student Ability across Grades

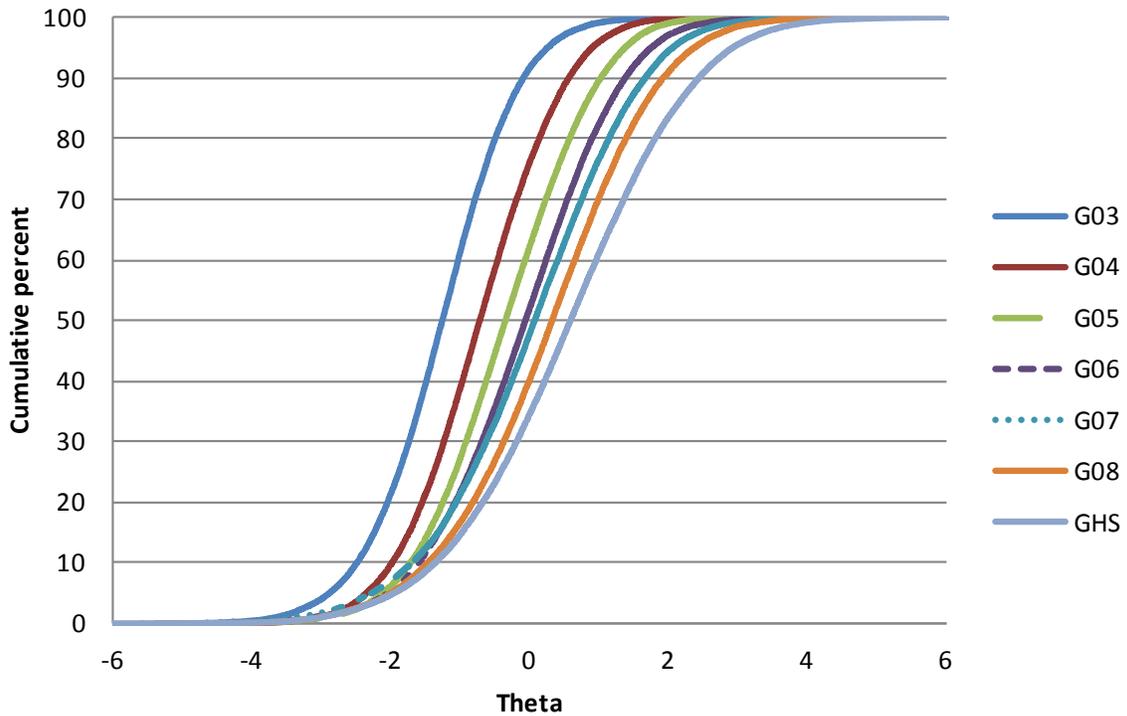


Figure 24. Mathematics Cumulative Distributions of Student Ability across Grades

Figures 25 and 26 present plots of the univariate theta (i.e., student ability) distributions for ELA/literacy and mathematics. Figures 27 and 28 present the ability distributions using boxplots for ELA/literacy and mathematics. The boxplots show that both the means and standard deviations tend to increase with grade level. No constraints were placed on the minimum and maximum thetas in the box plots (refer to the section on establishing the minimum and maximum thetas). The properties of the vertical scale are consistent with the comments of Kolen (2011) that an acceptable vertical scale should display increasing mean scores from grade-to-grade, the amount of growth should be decelerating, and the within-grade variability (*SD*) should be increasing from grade to grade.

Table 24. Summary of Vertically Scaled Student Ability Estimates and Effect Size.

Grade	<i>N</i>	Mean	<i>SD</i>	10%	25%	Median	75%	90%	Effect Size
ELA/literacy									
3	23,223	-1.227	1.051	-2.568	-1.961	-1.229	-0.487	0.138	
4	35,689	-0.737	1.106	-2.176	-1.476	-0.674	0.054	0.631	0.45
5	31,594	-0.305	1.102	-1.752	-1.052	-0.254	0.479	1.075	0.39
6	31,535	-0.048	1.107	-1.491	-0.785	-0.002	0.731	1.342	0.23
7	30,913	0.119	1.139	-1.357	-0.633	0.166	0.933	1.555	0.15
8	35,913	0.385	1.142	-1.099	-0.377	0.432	1.197	1.816	0.23
HS	50,657	0.527	1.211	-1.050	-0.301	0.573	1.400	2.050	0.12
Mathematics									
3	24,799	-1.265	0.960	-2.480	-1.872	-1.249	-0.634	-0.079	
4	38,925	-0.700	0.997	-1.972	-1.367	-0.705	-0.020	0.581	0.58
5	42,380	-0.330	1.073	-1.699	-1.052	-0.327	0.421	1.048	0.36
6	29,946	-0.083	1.183	-1.586	-0.858	-0.049	0.726	1.391	0.22
7	28,271	0.030	1.336	-1.691	-0.821	0.083	0.945	1.681	0.09
8	34,880	0.276	1.333	-1.469	-0.582	0.324	1.183	1.926	0.18
HS	47,608	0.571	1.494	-1.351	-0.408	0.598	1.578	2.448	0.21

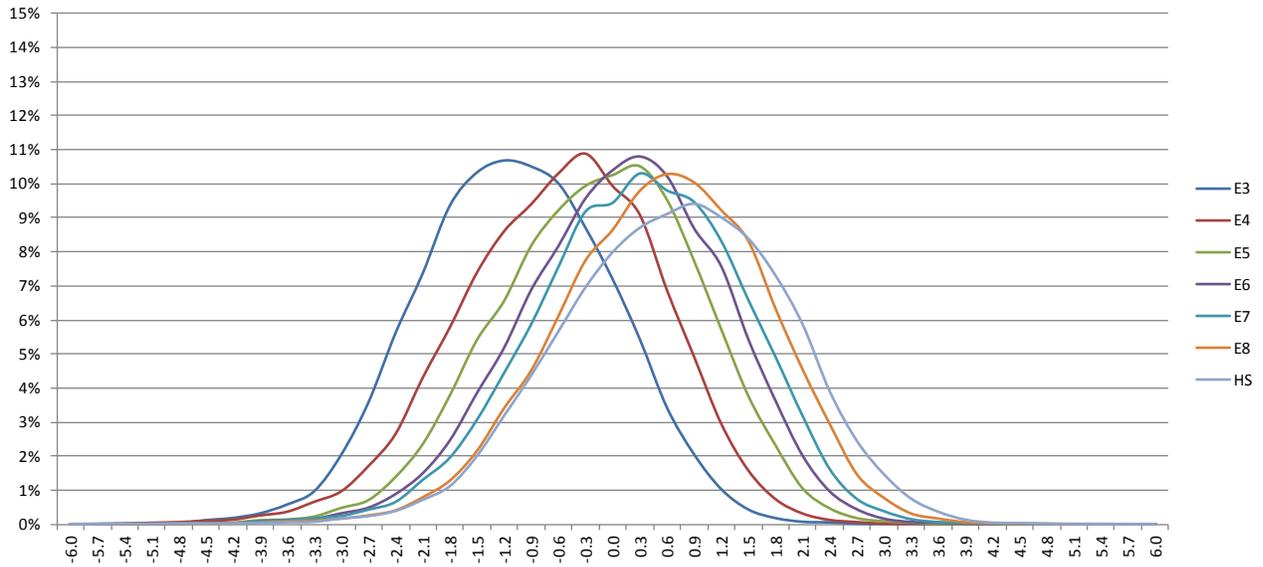


Figure 25. ELA/literacy Student Ability Distributions Across Grades 3 to High School

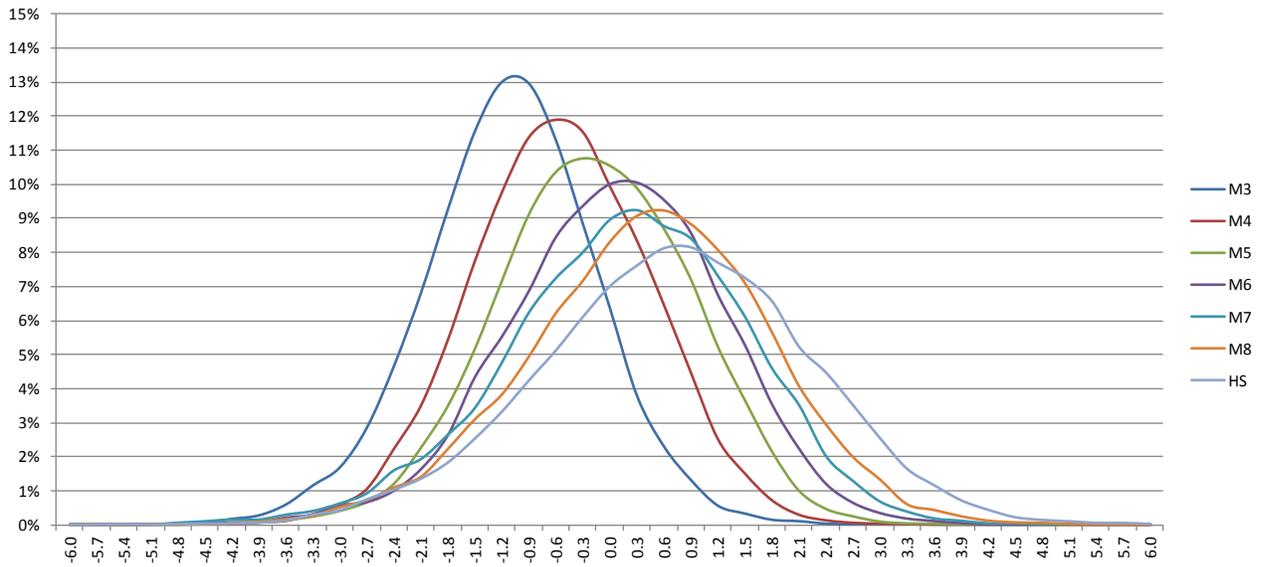


Figure 26. Mathematics Student Ability Distributions Across Grades 3 to High School

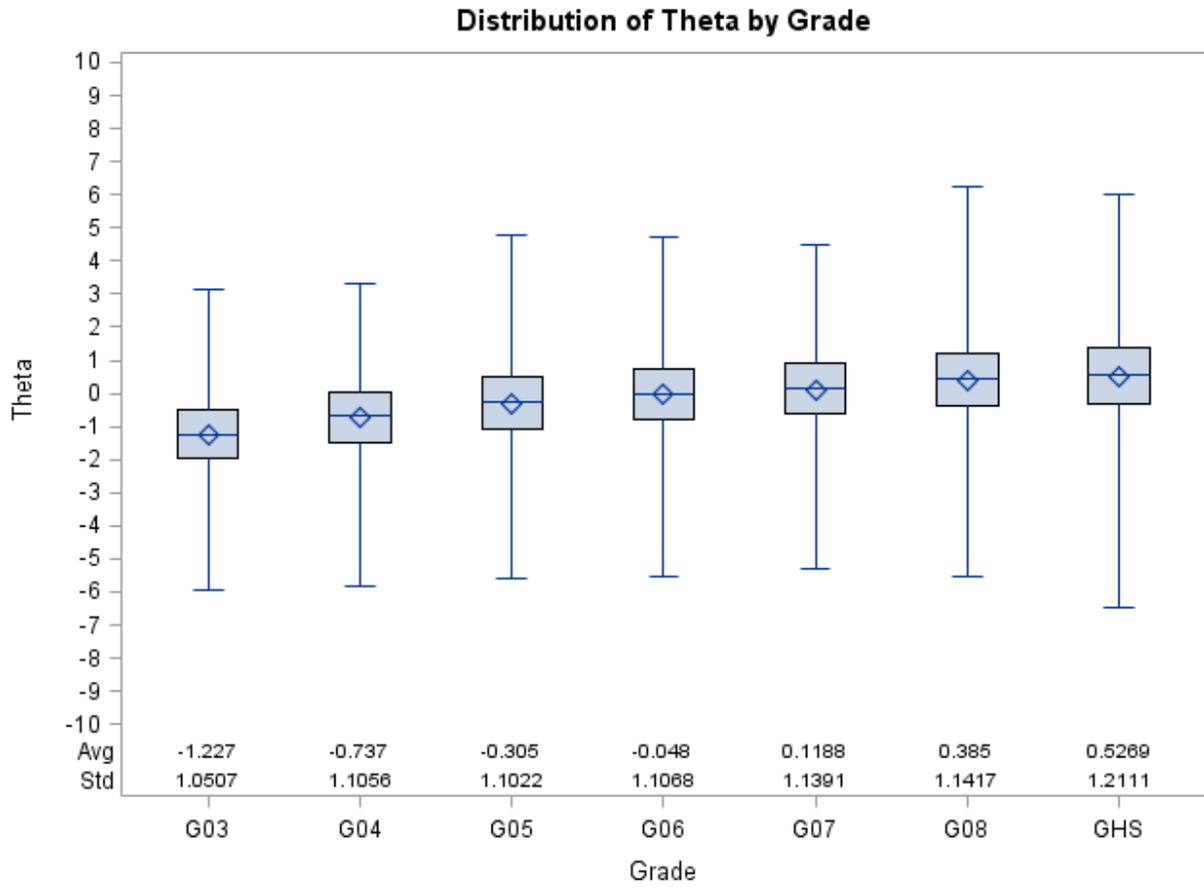


Figure 27. Boxplots of Theta Estimates across Grade Level for ELA/literacy

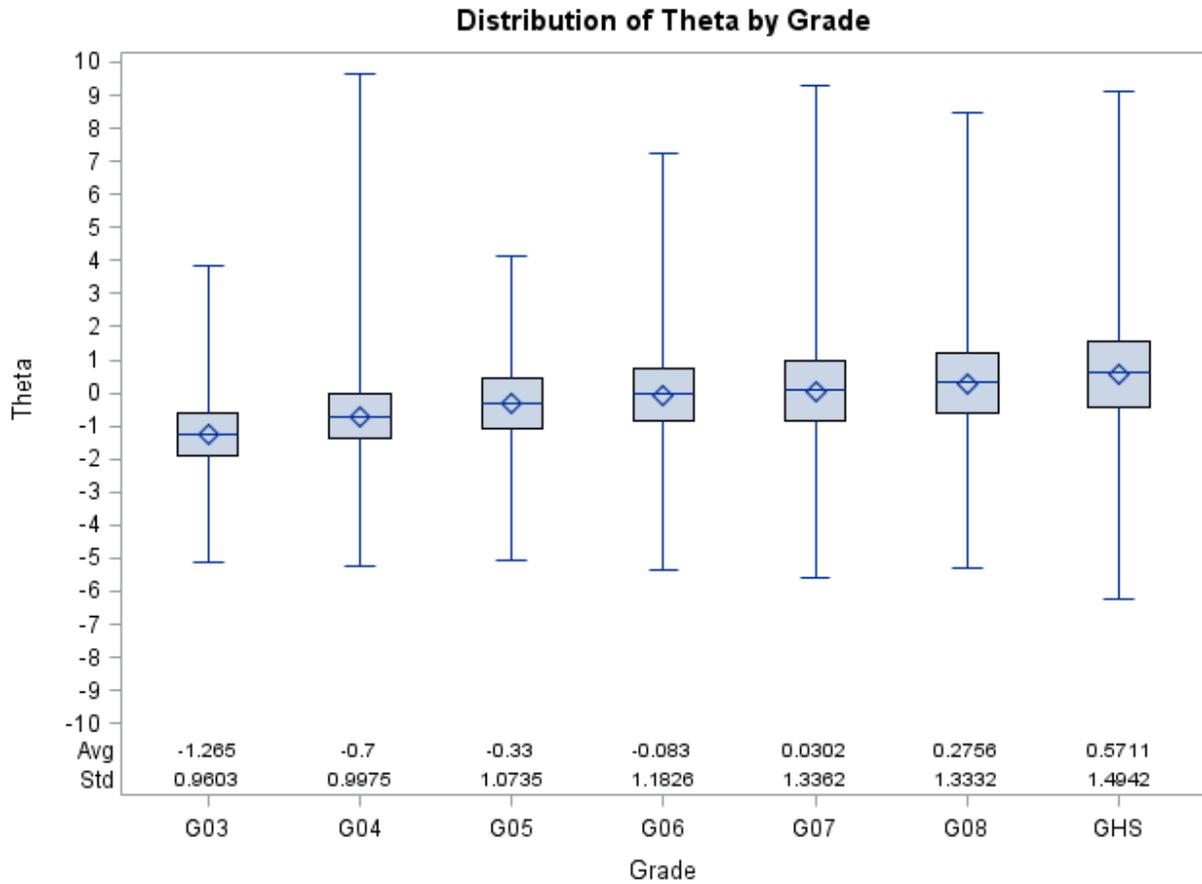


Figure 28. Boxplots of Theta Estimates Across Grade Level for Mathematics

Figures 29 to 34 display IRT information functions by score level and the total combined across all score levels (i.e., All) for ELA/literacy and mathematics. These displays reflect all items, both on-grade and off-grade vertical linking items, administered in a given grade level for the vertical scaling. Figures 31 and 34 show total information across grades for ELA/literacy and mathematics.

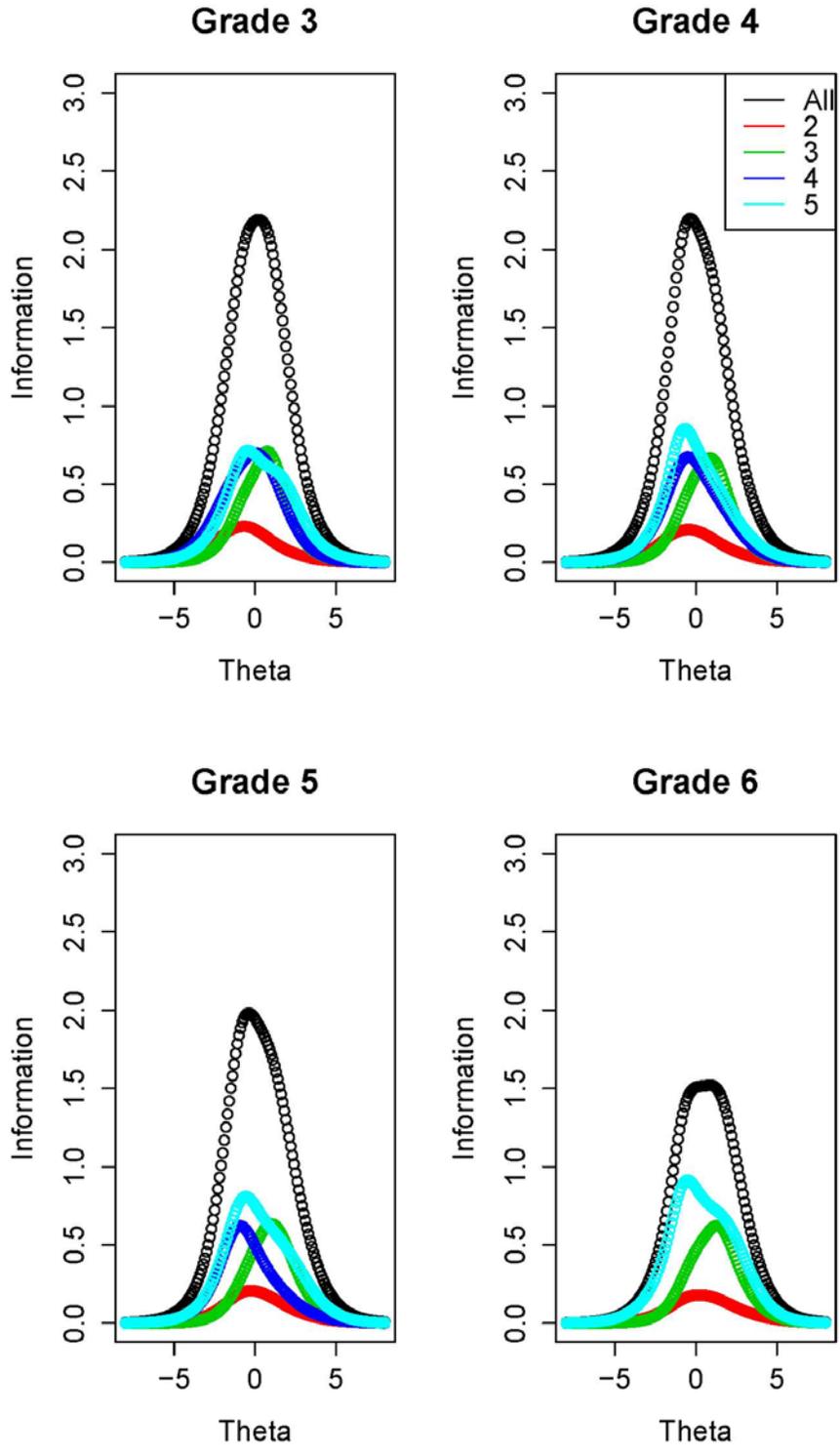


Figure 29. ELA/literacy Test Information by Score Level and Combined for Grades 3 to 6

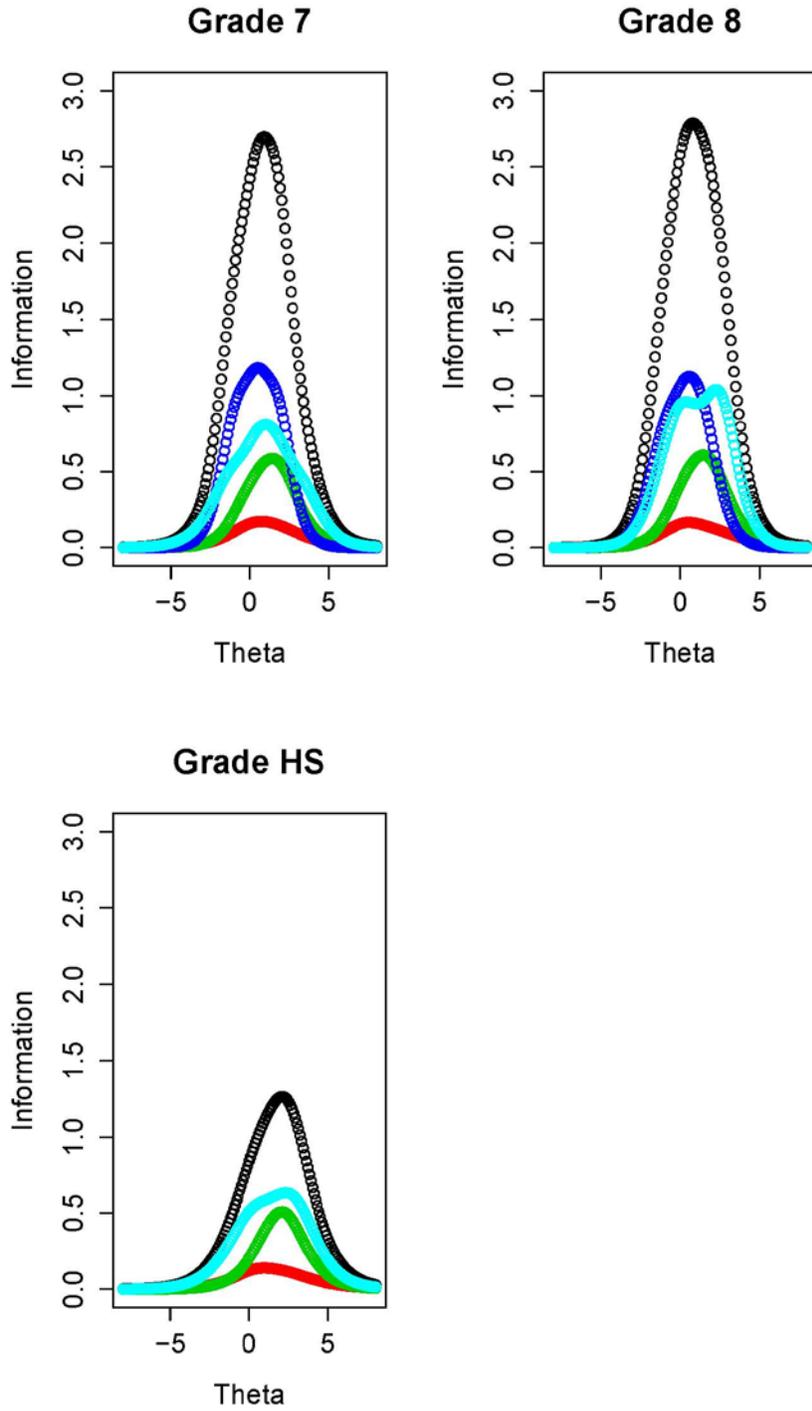


Figure 30. ELA/literacy Test Information by Score Level and Combined for Grades 7 to High School

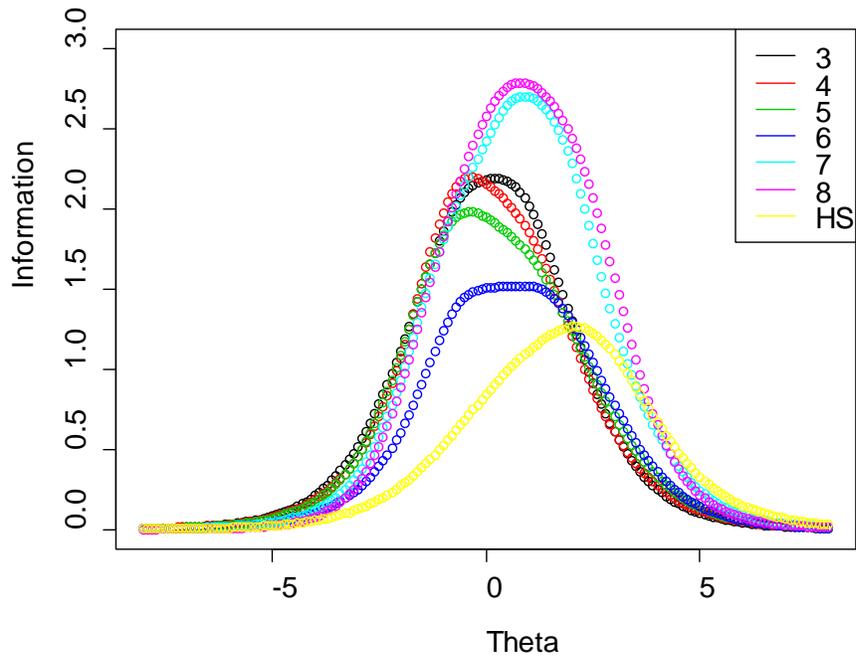


Figure 31. ELA/literacy Total Test Information for Grades 3 to High School

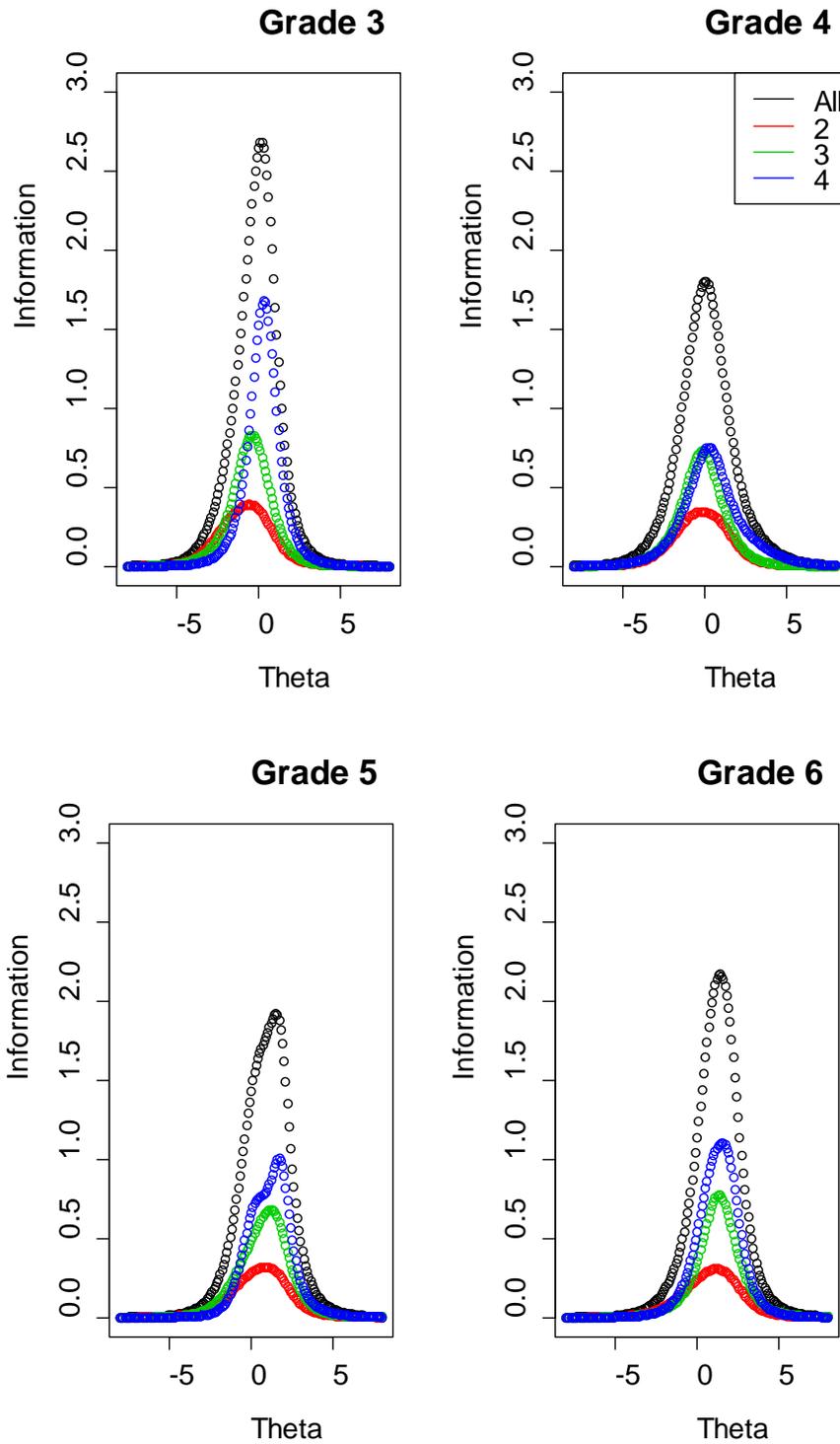


Figure 32. Mathematics Test Information and Score Level and Combined for Grades 3 to 6

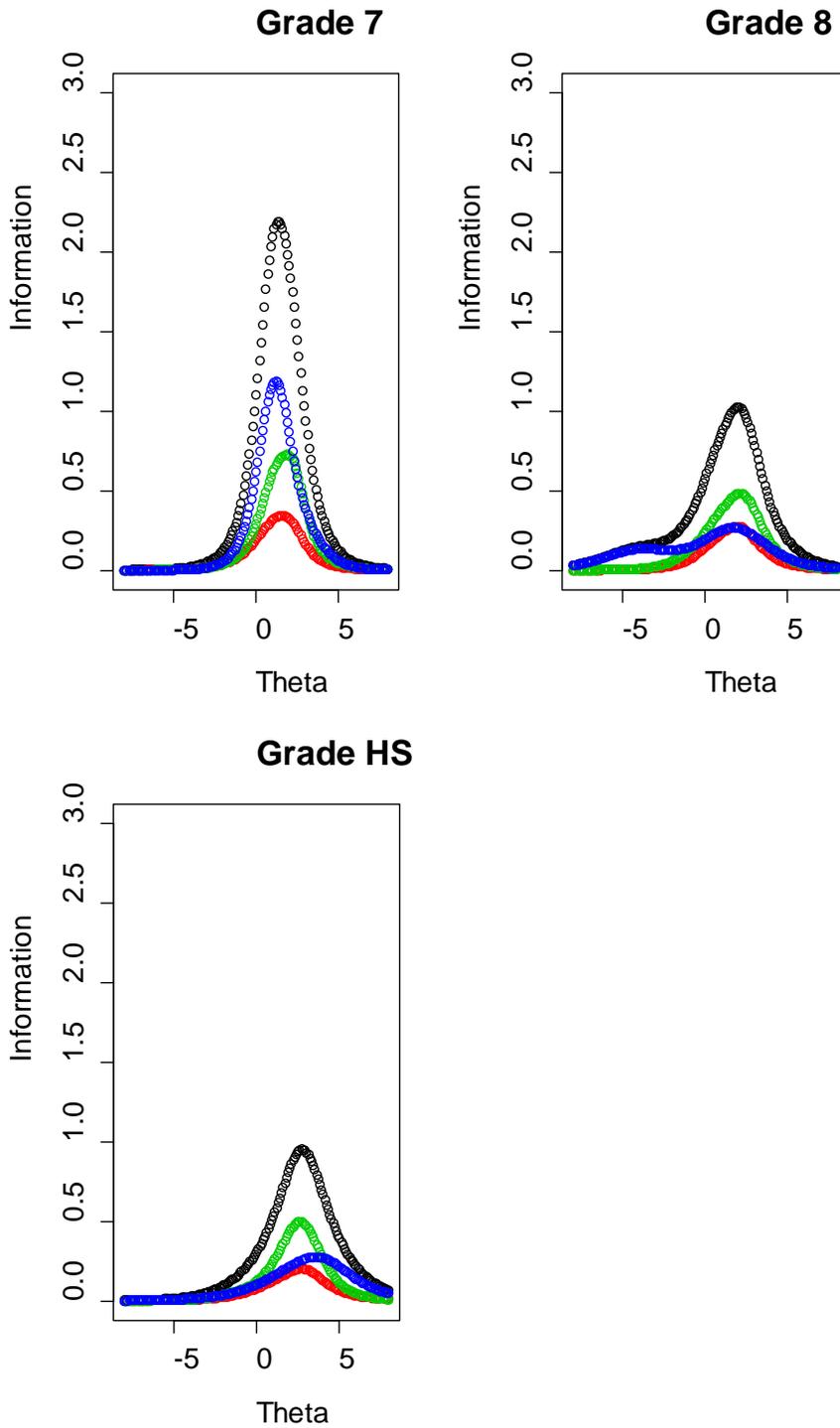


Figure 33. Mathematics Test Information and Score Level and Combined for Grades 7 to High School

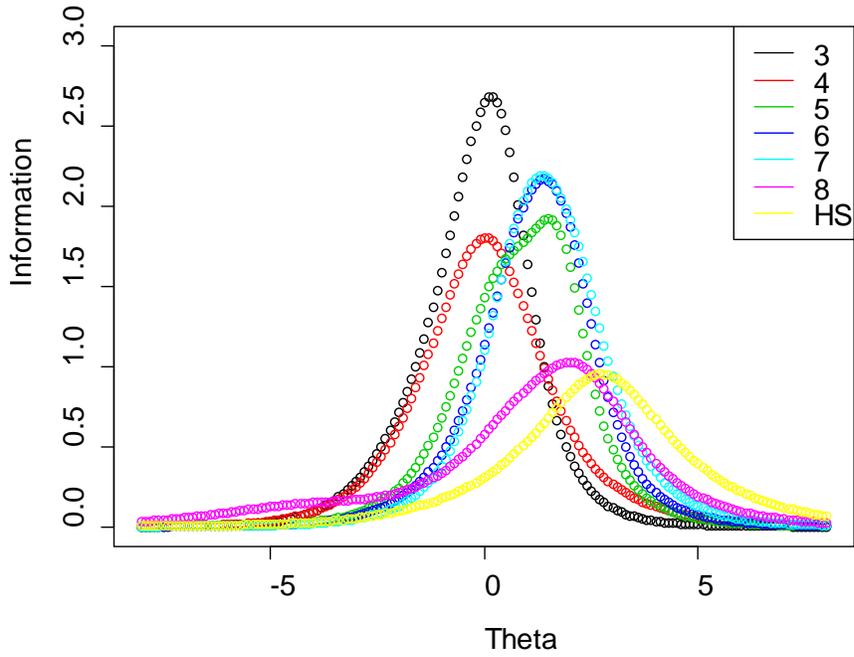


Figure 34. Mathematics Total Test Information for Grades 3 to High School

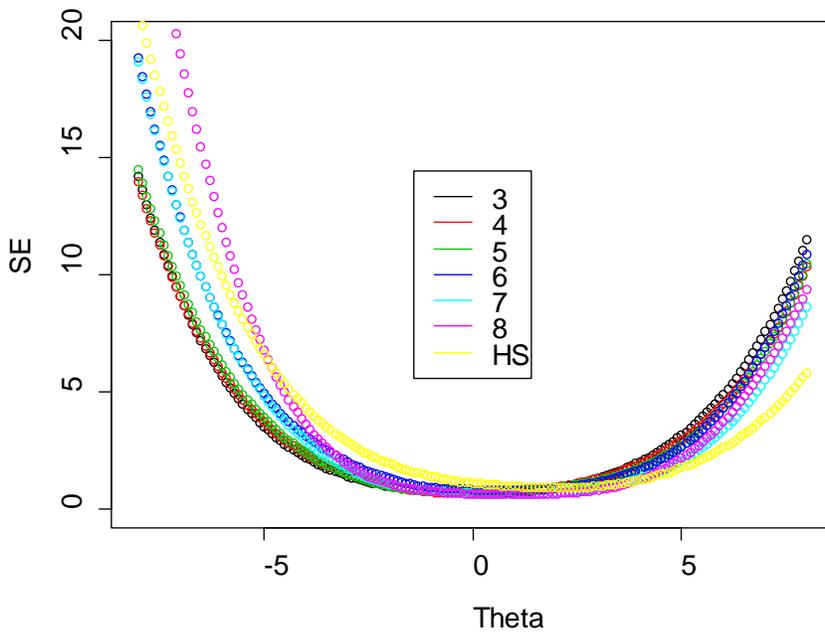


Figure 35. IRT Standard Error Plots for ELA/literacy Grades 3 to High School (HS)

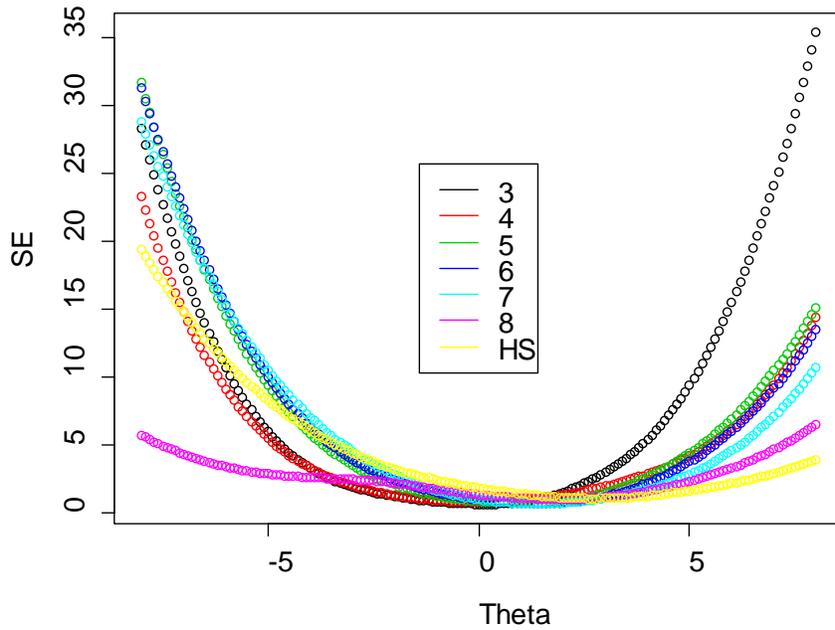


Figure 36. IRT Standard Error Plots for Mathematics Grades 3 to High School (HS)

IRT conditional standard errors were calculated as one over the inverse of information for a given level of theta. Plots of the conditional standard errors of measurement (*CSEM*) for ELA/literacy and mathematics are displayed in Figures 35 and 36 for grades 3 to high school. The item pools for the vertical Scale all tended to measure well over the middle part of the theta distribution. Separation between grades is primarily in the low and high theta ranges.

Establishing the Minimum and Maximum Scale Score

A maximum likelihood procedure will not result in theta estimates for students with perfect or zero scores. Scale scores can be established for these extreme values following a non-maximum likelihood but logical procedure. These minimum and maximum values are called the Lowest Obtainable Theta (LOT) and the Highest Obtainable Theta (HOT). The guidelines for establishing the LOT and HOT values were as follows.

1. The HOT should be high enough so that it does not cause an unnecessary pileup of scale scores at the top of the scale. Likewise, the LOT should be low enough so that it does not cause an unnecessary pileup of scale scores at the bottom part of the scale.
2. The HOT should be low enough so that $CSEM(HOT) < 10 * \text{Min}(CSEMs \text{ for all scale scores})$, where *CSEM* is the conditional standard error of measurement. The LOT should be high enough so that $CSEM(LOT) < 15 * \text{MIN}(CSEMs \text{ for all scale scores})$.
3. For multiple test levels placed on the same vertical scale, the HOT and LOT values should increase and transition smoothly over levels.

Table 25 provides recommendations for Smarter Balanced for the LOT and HOT values. The LOT and HOT values give the effective range of the ELA/literacy and mathematics scales. The ELA/literacy scale ranges from a value of -4.5941, which is the LOT for grade 3, to the HOT of 3.3392 for high school. In mathematics, the range was from -4.1132 to 4.3804. The means and SDs for theta given in Table 26 reflect the application of these LOT and HOT values.

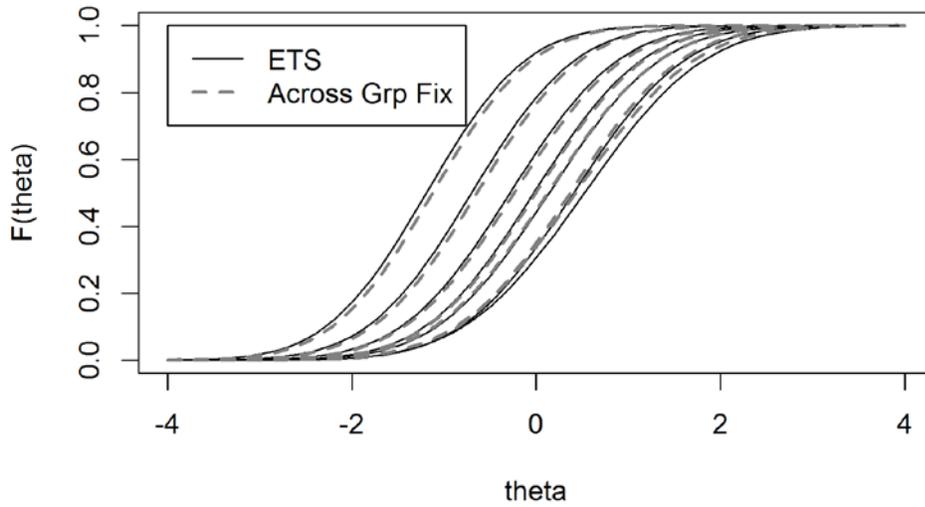
Table 25. Lowest and Highest Obtainable Theta Values and Resulting Theta Scale Summary.

LOT/HOT on Theta Scale						
Grade	LOT	<i>CSEM</i>	HOT	<i>CSEM</i>	Mean	<i>SD</i>
ELA/literacy						
3	-4.5941	1.22	1.3374	0.35	-1.240	1.06
4	-4.3962	1.20	1.8014	0.53	-0.748	1.11
5	-3.5763	1.03	2.2498	0.46	-0.310	1.10
6	-3.4785	1.10	2.5140	0.40	-0.055	1.11
7	-2.9114	0.84	2.7547	0.43	0.114	1.13
8	-2.5677	1.08	3.0430	0.35	0.382	1.13
HS	-2.4375	1.00	3.3392	0.47	0.529	1.19
Mathematics						
3	-4.1132	1.00	1.3335	0.44	-1.285	0.98
4	-3.9204	0.74	1.8191	0.47	-0.708	1.00
5	-3.7276	1.43	2.3290	0.34	-0.345	1.09
6	-3.5348	2.13	2.9455	0.66	-0.131	1.17
7	-3.3420	2.46	3.3238	0.56	-0.060	1.29
8	-3.1492	2.05	3.6254	0.68	0.080	1.36
HS	-2.9564	2.02	4.3804	0.64	0.417	1.47

Cross-validation of Vertical Linking Results

National Center for Research on Evaluation, Standards, and Student Testing (CRESST) was contracted by Smarter Balanced to conduct an independent replication of the Field Test item calibration and vertical linking. These analyses were conducted on the vertical scaling data in which classical item exclusion logic had already been applied. CRESST replicated the within-group concurrent calibration followed by the stepwise application of the Stocking-Lord to perform the vertical linking. They also reported the cross-validation by applying a multigroup approach in which all test levels are calibrated simultaneously. These analyses were implemented using the program **flexMIRT** (Cai, 2013), which implement Bayesian approaches for IRT parameter estimation. There was good agreement between both the CRESST Stocking-Lord and the multigroup approaches with the original ones reported here. Figure 37 shows the cumulative distributions for both the CRESST multigroup approach and the ones implemented for Smarter Balanced. Note the Expected A Posteriori (EAP) scores were computed by CRESST rather than the MLE estimates reported here. Using different methods (i.e., multigroup and Bayesian estimation) had essentially the same outcomes.

CDF of Normal Approx. using EAP Scores: ELA



CDF of Normal Approx. using EAP Scores: Math

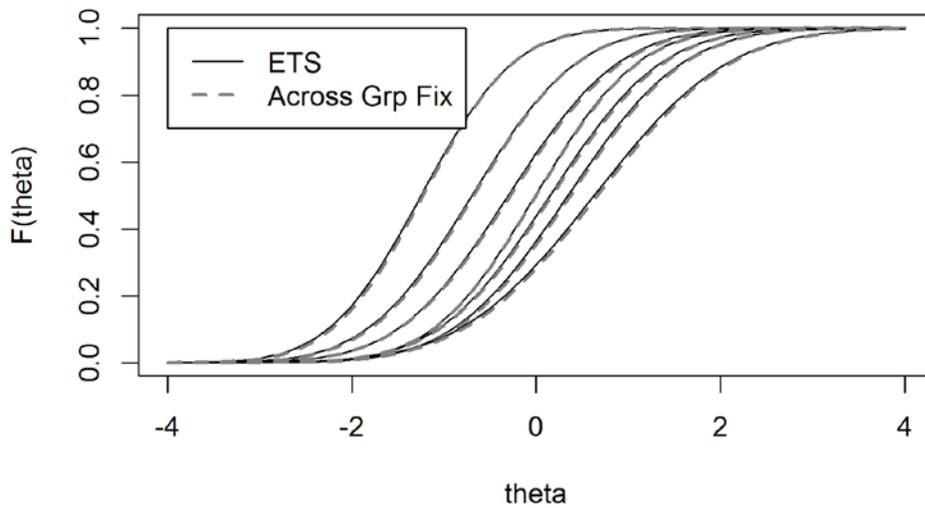


Figure 37. Cross-validation of Vertical Linking Results Comparing cumulative frequency distributions of theta (EAP) for ELA and mathematics obtained from the CRESST cross-validation

Calibration Step for Item Pool

In a second step after the completion of the vertical scaling, all items in the remaining pool were calibrated. This horizontal item-pool calibration involved a much larger number of items and students compared with the initial vertical scaling. It resulted in the final operational item pool at the conclusion of the Field Test. To perform this linking, on-grade, vertical scaling items were administered to students targeted in the horizontal item pool calibration. Using the common items from the vertical scaling step, these on-grade items were linked horizontally onto the scale in each grade using Stocking-Lord test-characteristic-curve methods. Table 26 shows the distribution of observations per student in the item-pool calibration sample after test delivery. Items with fewer than 500 observations were not calibrated. Table 27 presents the mean and standard deviation for the parameter estimates and the number of combined CAT and performance task items. To compare the results of the two respective scaling steps, Figures 38 and 39 show the plots of cumulative theta estimate distributions for the vertical scaling (achievement level setting sample) and the item-pool calibration step. These figures show the outcomes for ELA/literacy and mathematics at each grade. The figures show close agreement for the outcomes from the vertical scaling and item-pool calibration step.

Table 26. Distribution of Student Observations per Item in the Field Test Pool.

Item Response Frequency Percentiles											
Grade	Min	1	5	10	25	Median	75	90	95	99	Max
ELA/literacy											
3	72	270	636	846	1,514	2,438	3,899	6,763	8,123	13,474	24,446
4	80	279	655	957	1,350	2,192	4,171	6,852	8,724	15,945	43,327
5	57	146	658	941	1,296	2,166	4,090	6,740	8,905	12,688	25,518
6	63	242	763	1,128	1,619	2,175	4,202	7,175	11,286	17,193	21,795
7	50	188	556	932	1,284	2,292	4,169	6,735	8,785	17,430	37,351
8	61	253	558	938	1,244	2,057	4,487	7,524	12,702	16,500	20,843
HS	50	73	180	406	849	1,633	3,052	5,635	7,682	13,858	27,229
Mathematics											
3	1,142	1,173	1,251	1,580	1,786	1,914	3,026	4,295	4,838	10,305	14,182
4	913	1,000	1,053	1,220	1,852	2,063	3,782	5,496	6,870	13,121	20,497
5	926	997	1,108	1,258	1,905	2,163	3,769	5,853	8,278	19,590	21,085
6	959	970	987	996	1,331	1,829	2,362	3,993	4,870	11,905	15,702
7	506	910	943	965	1,420	1,940	2,887	3,968	4,656	13,769	14,204
8	618	905	939	957	1,289	1,989	2,765	4,559	5,445	12,964	15,932
HS	164	302	324	339	557	1,066	1,376	2,694	3,241	8,227	30,833

Table 27. Summary of IRT Item Parameter Estimates for the Field Test Item Pool.

		<i>a</i> -parameter		<i>b</i> -parameter	
Grade	No. of Items	Mean	SD	Mean	SD
3	896	0.654	0.23	-0.208	1.21
4	856	0.593	0.21	0.259	1.29
5	823	0.613	0.20	0.607	1.23
6	849	0.568	0.22	1.101	1.35
7	875	0.567	0.23	1.333	1.39
8	836	0.555	0.21	1.464	1.43
HS	2,371	0.491	0.18	1.819	1.47
3	1,114	0.851	0.29	-0.759	1.06
4	1,130	0.814	0.29	-0.052	1.03
5	1,043	0.766	0.30	0.669	1.02
6	1,018	0.715	0.26	1.029	1.18
7	942	0.727	0.29	1.670	1.24
8	894	0.626	0.27	2.174	1.41
HS	2,026	0.536	0.26	2.668	1.55

For the item-pool calibration sample, Tables 28 and 29 present the distributions for theta estimates and the conditional standard errors of estimate in a grade and content area using the five-number summary. Table driven methods using sufficient statistics were applied to produce the estimated theta values. The CSEM was reciprocal of the inverse of test information for a given student. In this case, the LOT and HOT values were applied for these theta values in each grade and content area.

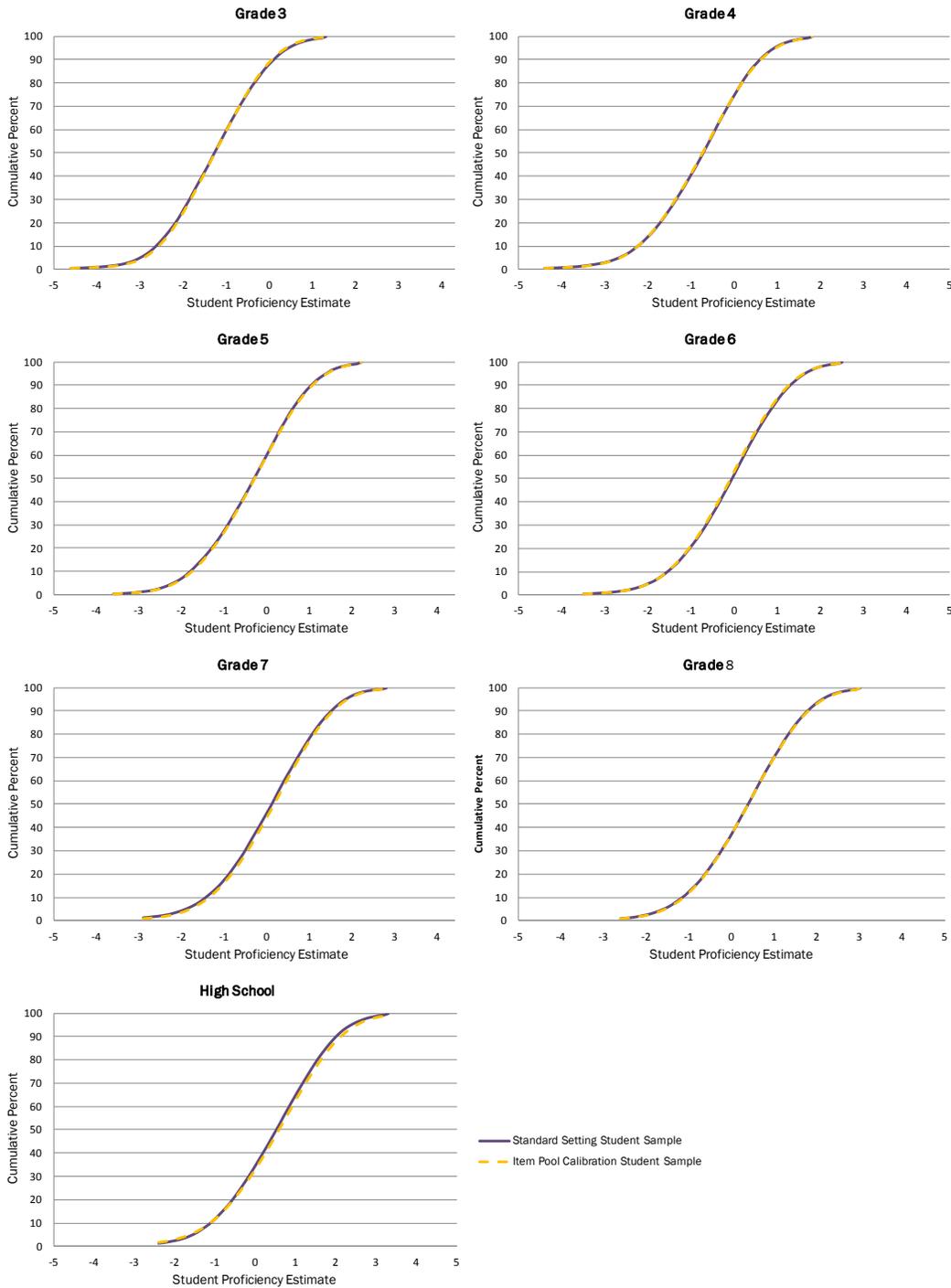


Figure 38. Comparison of Student Proficiency Estimates (theta) for the Vertical Scaling (Achievement Level Setting Sample) and the Item Pool Calibrations Step for ELA/literacy

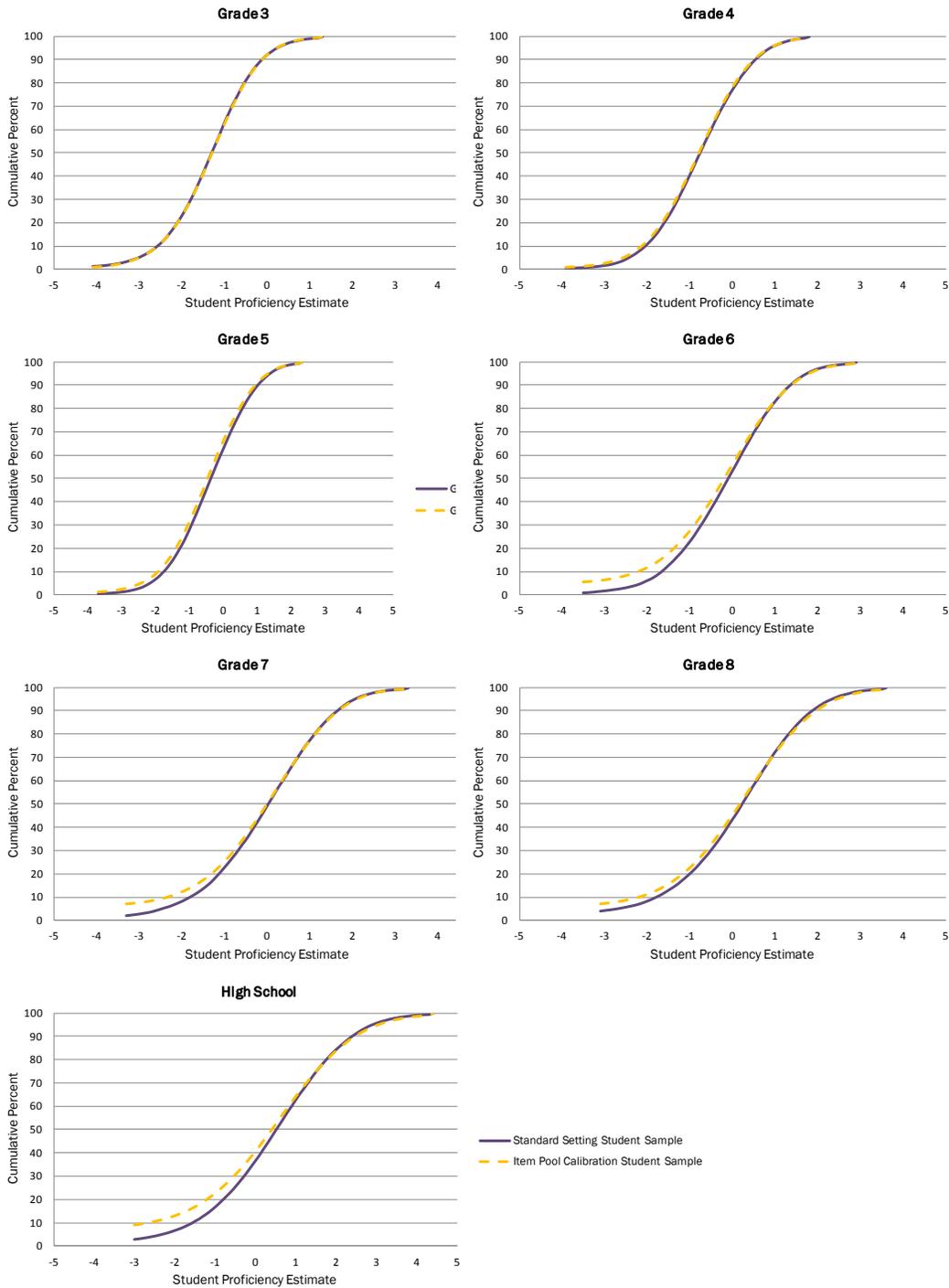


Figure 39. Comparison of Student Proficiency Estimates (theta) for the Vertical Scaling (Achievement Level Setting Sample) and the Item Pool Calibrations Step for Mathematics

Table 28. Distributions of ELA/literacy Theta Estimates and Conditional Standard Error of Measurement.

Grade	3		4		5		6		7		8		HS	
<i>N</i>	83,531		92,595		85,885		90,814		89,332		93,877		228,136	
	Theta	CSEM	Theta	CSEM										
Mean	-1.22	0.38	-0.74	0.43	-0.29	0.40	-0.08	0.40	0.18	0.40	0.41	0.40	0.60	0.49
SD	0.99	0.12	1.07	0.15	1.06	0.12	1.06	0.13	1.08	0.12	1.08	0.11	1.17	0.14
Min	-4.59	0.24	-4.39	0.25	-3.56	0.25	-3.48	0.25	-2.91	0.25	-2.57	0.25	-2.44	0.28
Max	1.34	1.65	1.80	1.90	2.25	1.64	2.51	1.86	2.75	1.62	3.04	1.69	3.34	1.87
10	-2.50	0.28	-2.16	0.30	-1.69	0.30	-1.48	0.30	-1.26	0.30	-1.03	0.31	-0.98	0.36
25	-1.92	0.30	-1.48	0.34	-1.02	0.32	-0.80	0.32	-0.56	0.32	-0.35	0.33	-0.23	0.39
50	-1.22	0.35	-0.69	0.39	-0.25	0.37	-0.04	0.36	0.22	0.36	0.44	0.36	0.64	0.45
75	-0.50	0.41	0.04	0.48	0.49	0.43	0.68	0.43	0.96	0.43	1.20	0.43	1.47	0.54
90	0.08	0.50	0.63	0.59	1.08	0.54	1.29	0.55	1.57	0.54	1.81	0.53	2.13	0.66

Table 29. Distributions of Mathematics Theta Estimates and Conditional Standard Error of Measurement.

Grade	3		4		5		6		7		8		HS	
<i>N</i>	91,325		106,124		105,335		108,667		103,296		102,176		202,115	
	Theta	CSEM	Theta	CSEM	Theta	CSEM	Theta	CSEM	Theta	CSEM	Theta	CSEM	Theta	CSEM
Mean	-1.27	0.33	-0.75	0.37	-0.43	0.44	-0.10	0.56	0.12	0.57	0.31	0.73	0.61	0.92
SD	0.93	0.10	0.99	0.14	1.07	0.19	1.18	0.26	1.24	0.26	1.30	0.37	1.46	0.46
Min	-4.11	0.21	-3.92	0.21	-3.73	0.21	-3.53	0.25	-3.34	0.22	-3.15	0.27	-2.96	0.29
Max	1.33	2.41	1.82	2.00	2.33	2.20	2.94	3.99	3.32	2.49	3.63	4.39	4.38	4.53
10	-2.48	0.25	-2.03	0.26	-1.81	0.28	-1.66	0.33	-1.54	0.31	-1.41	0.39	-1.33	0.48
25	-1.87	0.27	-1.41	0.29	-1.15	0.31	-0.89	0.39	-0.73	0.39	-0.57	0.47	-0.40	0.59
50	-1.24	0.30	-0.73	0.34	-0.42	0.37	-0.07	0.49	0.15	0.51	0.33	0.62	0.62	0.80
75	-0.63	0.35	-0.07	0.41	0.32	0.49	0.73	0.64	1.00	0.68	1.23	0.86	1.63	1.12
90	-0.09	0.44	0.52	0.52	0.95	0.67	1.41	0.86	1.72	0.91	2.00	1.21	2.52	1.52

References

- Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. pp. 395-479. In Lord, F. M. and Novick, M. R. *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley.
- Bock, R. D. & Zimowski, M. F. (1997). Multiple Group IRT. In W.J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 433-448). New York: Springer-Verlag.
- Briggs, D. C. & Weeks, J. P. (2009). The Impact of Vertical Scaling Decisions on Growth Interpretations. *Educational Measurement: Issues & Practice*, 28, 3-14.
- Cai, L. (2013). flexMIRT® 2.0: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.
- du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.
- Ercikan, K., Schwarz, R., Julian, M., Burket, G., Weber, M., & Link, V. (1998). Calibration and Scoring of Tests with Multiple-choice and Constructed-Response Item Types. *Journal of Educational Measurement*, 35, 137-155.
- Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling Performance Assessments: A Comparison of One-Parameter and Two-Parameter Partial Credit Models. *Journal of Educational Measurement*, 33, 291-314.
- Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C., (1999). *Uncommon Measures: Equivalence and Linkage Among Educational Tests*. Washington, DC: National Academy Press.
- Gu, L., Lall, V. F., Monfils, L., & Jiang, Y. (2010). *Evaluating Anchor Items for Outliers in IRT Common Item Equating: A Review of the Commonly Used Methods and Flagging Criteria*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.
- Haebera, T. (1980). Equating Logistic Ability Scales by Weighted Least Squares Method. *Japanese Psychological Research*, 22, 144-149.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.
- Hanson, B. A., & Beguin, A. A. (1999). *Separate Versus Concurrent Estimation of IRT Parameters in the Common Item Equating Design*. ACT Research Report 99-8. Iowa City, IA: ACT.
- Hanson, B. A., & Beguin, A. A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. *Applied Psychological Measurement*, 26, 3-24.
- Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and Separate Grade-Group Linking Procedures for Vertical Scaling. *Applied Measurement in Education*, 21, 187-206.
- Karkee, T., Lewis, D. M., Hoskins, 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.
- Kim, S. H., & Cohen, A. S. (1998). A Comparison of Linking and Concurrent Calibration Under Item Response Theory. *Applied Psychological Measurement*, 22, 131-143.

- Kim, S. H., & Kolen, M. J. (2004). *STUIRT: A Computer Program for Scale Transformation Under Unidimensional Item Response Theory Models (Version 1.0)*. Iowa Testing Programs, University of Iowa.
- Kolen, M. J., & Brennan, R. L. (2004). *Test Equating: Methods and Practices*. (2nd ed.). New York, NY: Springer-Verlag.
- Kolen, M. J. (2011). *Issues Associated with Vertical Scales for PARCC Assessments*. PARCC.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.
- Mislevy, R. J. (1987). Recent Developments in Item Response Theory. *Review of Research in Education*, 15, 239-275.
- Mislevy, R. J., & Bock, R. J. (1990). *BLOG3: Item Analysis and Test Scoring with Binary Logistic Model* (2nd ed.) [Computer program]. Mooresville, IN: Scientific Software.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E., & Bock, R. D. (2003). *PARSCALE 4.1: IRT based item analysis and test scoring for rating-scale data*. Chicago, IL: Scientific Software International, Inc.
- Orlando, M., & Thissen, D. (2003) Further Examination of the Performance of S-X2, an Item Fit index for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, 27(4), 289-98.
- PARPLOT (2009). ETS: Author.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, Norming, and Equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-262). New York, NY: Macmillan.
- Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103, 677-680.
- Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, 7, 201-210.
- Stone, C. A., & Zhang, B. (2003). Assessing Goodness of Fit of Item Response Theory Models: A Comparison of Traditional and Alternative Procedures. *Journal of Educational Measurement*, 40, 331-352.
- Stout, W. (1987). A Nonparametric Approach for Assessing Latent Trait Unidimensionality. *Psychometrika*, 52, 589-617.
- Sykes, R. C., & Yen, W. M. (2000). The Scaling of Mixed-Item-Format Tests with the One-Parameter and Two-Parameter Partial Credit. *Journal of Educational Measurement*, 37, 221-244.
- Wainer, H. (Ed.). (2000). *Computerized Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Williams, V. S., Pommerich, M., & Thissen, D. (1998). A Comparison of Developmental Scales Based on Thurstone Methods and Item Response Theory. *Journal of Educational Measurement*, 35, 93-107.
- Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.
- Yen, W., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111-153). Westport, CT: Praeger Publishers.

- Yen, W. M. (1993). Scaling performance assessments – strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.
- Yen, W. (1981). Using Simulation Results to Choose a Latent Trait Model. *Applied Psychological Measurement*, 5, 245-262.
- Yen, W. M. (1986). The Choice of Scale for Educational Measurement: An IRT Perspective. *Journal of Educational Measurement*, 23, 299-325.
- Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. D. (1997). *BILOG-MG: Multiple Group Item Analysis and Test Scoring with Binary Logistic Models*. Mooresville, IN: Scientific Software.

Chapter 10 Achievement Level Setting

Background

In a request for proposals (RFP) issued in October 2013, Smarter Balanced called for a contractor to provide services for a multi-phase standard setting process and to plan and execute a comprehensive Communication Plan. The standard setting plan was to be executed in five phases from conducting distributed standard setting to finalizing achievement level descriptors. The communication plan was to “proactively explain the rationale for setting common achievement standards tied to the Common Core State Standards, describe the standard-setting process in layman’s terms, and make the case for approval of the performance standards derived from the standard setting process” (RFP, p. 10).

The word “standard” is used in many in educational and assessment contexts with varied meanings. To clarify the process of choosing proficiency thresholds, the consortium uses the term “achievement level setting” (ALS) which is used throughout this chapter in reference to the specific consortium activity. The assessment research literature calls this process “standard setting”, which is used in reference to the general process.

This chapter describes in some detail key outcomes of the ALS project. Specifically, this chapter documents the application and subsequent revision of achievement level descriptors, the in-person standard-setting activity, and the approved cut scores.

Achievement Level Setting Process

Achievement level setting is the culminating set of activities in a four-year enterprise to create, field test, and implement a set of rigorous computer-adaptive assessments closely aligned to the Common Core and to provide guidance to educators regarding the achievements of their students, with particular reference to college and career readiness. The goal of the process is to identify assessment scores that delineate levels of achievement described by achievement level descriptors. Smarter Balanced has adopted four levels of achievement. For each grade and subject, there are three threshold cuts: Level 1 and level 2; Level 2 and Level 3; Level 3 and level 4. The division between Levels 2 and 3 is used as the proficiency criterion in accountability.

The ALS process used two components, an online panel that allowed broad stakeholder participation and provided a wide data set, and a more traditional in-person workshop that provided focused judgment from a representative stakeholder panel. The in-person workshop included a final cross-grade review stage. The online panel and in-person workshop used a Bookmark procedure (Lewis, Mitzel, Mercado, & Schultz, 2012), while the vertical articulation (cross-grade review) employed a procedure described by Cizek & Bunch (2007, Chapter 14). Details of both procedures are described in the sections below.

The Bookmark Procedure

The Bookmark standard setting procedure (Lewis et al., 2012) is an item response theory-based item mapping procedure developed in 1996 in response to the need for a robust standard setting procedure for high-stakes assessments of mixed format. Since 1996, it has become the most widely used procedure for setting cut scores on statewide assessments and other high stakes educational assessments. Its psychometric foundation is well documented (e.g., Cizek & Bunch, 2007), and its usefulness has been well established through adoption of cut scores produced by Bookmark-based standard-setting activities.

Creating ordered item booklets.

The bookmark method relies on presenting panelists with sets of test items sorted by difficulty and representing test content. This item collection is called the ordered item booklet (OIB). An important consideration when creating an ordered item booklet is to ensure appropriate content coverage. Psychometricians and content specialists from MI worked together closely to construct content specifications that matched Smarter Balanced guidelines with respect to targets and claims used to inform item and test development. The OIBs contained at least 70 items pages and with content weighted according to the specifications. Each OIB contained an entire performance task, that is, a set of 5-6 items/tasks all related to a set of stimuli. In order to minimize the reading load of the panelists, the ELA booklets included reading passages with a minimum of three associated items.

Since item order is the basis for panelist judgment, statistical considerations are of primary importance when building the OIBs. Thus, the booklets contained items that had a wide range of difficulty across the score scale with items at generally equal intervals of difficulty. All OIB items exhibited acceptable classical statistics, and showed no differential functioning. Combining the content and statistical constraints decreased the number of items for selection, but the final OIBs were very representative of the specified test content.

All OIBs were reviewed by MI, CTB, and Smarter Balanced's content and measurement experts. The reviews resulted in the removal and insertion of several items within each grade-content area until Smarter Balanced staff gave their final approval.

In a typical Bookmark procedure, each item in an OIB is mapped to an underlying construct in terms of the amount of that construct the examinee must possess in order to have a reasonable chance of answering the item correctly (in the case of a selected-response item) or obtaining a given score point or higher (in the case of an open-ended item or performance task).

In the three-parameter logistic (3PL) model, the Bookmark procedure relies on the basic relationship between person ability (θ) and item difficulty (b), discrimination (a), and pseudo-guessing (c), where the probability of answering a dichotomously scored item correctly (P) can be expressed as shown in equation (11.1).

$$P_j(X=1|\theta) = c_j + (1 - c_j)/\{1+\exp[-1.7a_j(\theta - b_j)]\} \quad (11.1)$$

where P_j is the probability of answering correctly, θ is the ability required, a_j is the item discrimination index, \exp is the exponential function, and b_j is the item difficulty index. The way that guessing is accounted for is critical to the mapping. For most bookmark procedures, the c (pseudo-guessing) parameter is set to zero, so that the response probability specified is associated with the likelihood of a student knowing the correct response without guessing, as shown in equation (11.2). For this project, the two-PL model (with c set to 0) was used.

$$P_j(X=1|\theta) = 1/\{1 + \exp[-1.7a_j(\theta - b_j)]\} \quad (11.2)$$

For items with two or more score points, the probability of achieving any score k point or better given student ability $P_{jk}(\theta)$ in a 2-parameter logistic model can be expressed as shown in equation 11.3 from Mitzel, Lewis, Patz & Green (2001).

$$P_{jk}(\theta) = \frac{m_j \exp(z_{jk})}{\sum_{i=1}^{m_j} \exp(z_{ji})}, \quad (11.3)$$

where m_j is the number of score points or steps for item j , and $z_{jk} = (k - 1)\alpha_j - \sum_{i=0}^{k-1} \gamma_{ji}$; α_j is the discrimination index of item j , k is the number of this score point or step, and γ_{ji} is the step value for item j at step i . Thus, the probability of scoring at step k is a joint function of examinee ability, item discrimination, and the likelihood of obtaining any of the $k - 1$ other scores. In this formulation, the value for a score of 0 (step 0) is set equal to zero; i.e., $\gamma_{j0} = 0$ for all items.

In practice, item maps show each item ordered in terms of the underlying value of θ required to answer dichotomously scored items correctly and the value of θ required to obtain at least each score point for multi-point items. Such maps may also contain other data, such as content domain, or other item metadata. It is also possible to show validation data.

In the Bookmark procedure, panelists are typically asked to find an item that a certain percentage of examinees at a critical threshold will be able to answer correctly. The cut score is identified at the point in an ordered item booklet beyond which panelists can no longer say that the target group would have the specified likelihood of answering correctly. The choice of that percentage is critical not only to defining the group of examinees but to defining the threshold between adjacent ability groups. This percentage is commonly called the RP value. In practice, users of the Bookmark procedure have employed 50 percent, 60 percent, 67 percent, and other values. For this project, upon the advice of the Technical Advisory Committee (TAC), RP50 was used.

Solving equation (11.2) for θ produces equation (11.4):

$$\theta = b_j + \ln(1/P_j - 1)/(-1.7a_j) \quad (11.4)$$

where \ln is the natural logarithm and other values are as defined above. For any value other than 50%, the value for $\ln(1/P_j - 1)$ is nonzero. However, when $P_j = .50$, the value of $\ln(1/P_j - 1)$ reduces to $\ln(1)$, which is 0, and the value of θ reduces to the item difficulty b_j , and item discrimination plays no part in the determination of the threshold ability level. Solving equation 11.3 for θ involves the simultaneous calculation of the probabilities of obtaining each score point or better and is described in detail in Cizek & Bunch (2007). Thus the OIBs used in the consortium's achievement level setting process were ordered on the b parameter.

Item mapping.

Item mapping allows individual items to be located along the scale score continuum so that interpretations about what students know and can do at individual scale score points may be facilitated. Item mapping is a component in the process of setting performance standards in the Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996). Item mapping is based in item response theory, exploiting the use of a common scale to express item difficulty and examinee proficiency. It requires the human judgmental process because panelists must make a decision about the response probability (RP; the likelihood that a person answers the item correctly) in order to align an item with a specific score point.

In addition to purely psychometric information, item maps may also contain item metadata (content standard, depth of knowledge, etc.) and other information. For this project, the contractor developed item maps that contained the content standard to which each item was aligned, the depth of knowledge associated with that item, ability level (expressed in scale score units), and, for the grade 11 tests, a region corresponding to a projection of college and career scale score levels on the ACT Assessment.

External data.

Some of the items in the OIBs for grades 4, 8, and 11 are not Smarter Balanced items but actually come from other tests such as the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA). These items were embedded in the spring 2014 field test to provide panelists with an external reference range to the performance of students on other tests. They were to be used as part of the internal ALS process, not as a broad indicator.

In addition, for both Math and ELA in grade 11, panelists could see an area of the item map where ACT benchmark scores were projected. These benchmarks are estimates of scores students need to attain on the ACT in orderable to be considered ready to enter credit-bearing coursework at the postsecondary level.

Facilitators presented and discussed the external data rather briefly. Because many factors differentiate the Smarter Balanced tests from these other assessments, the facilitators maintained

the focus of the panelists on the Smarter Balanced ALDs, relevant claims and targets, and the items in the OIBs.

Typical application of the Bookmark procedure.

In a typical application of the Bookmark procedure, panelists receive extensive training in the content standards, the achievement level descriptors, the test to be reviewed, and the Bookmark procedure itself. This training typically takes a day or more. Panelists are then organized into small groups of 5-6 and instructed to review the OIB and place one or more bookmarks in accordance with the training procedures. Each such small group is led by a panelist serving as a table leader. Several such small groups make up a panel of 15 or more panelists, led by a facilitator in addition to the several table leaders. The facilitator provides ongoing instruction and leads discussions between rounds of item review. There are typically two or three rounds of item review.

After training in the bookmark procedure, panelists typically complete a practice round, setting a single bookmark in a very short OIB (usually 6-9 pages) and discuss the experience among themselves with leadership by the facilitator. Once all panelists confirm that they understand the process and the task, they begin Round 1.

In Round 1, panelists review the items in the OIB with a series of questions in mind:

1. *What do you know about a student who responds successfully to this item; that is, what skills must a student have in order to know the correct answer?*
2. *What makes this item more difficult than preceding items?*
3. *Would a student at the threshold have at least a 50% chance of earning this point?*
 - *Yes: Move on to the next item.*
 - *No: Place your bookmark here.*

Panelists then place a bookmark on the first page in the OIB where they believe the student at the threshold for that level would NOT have at least a 50% chance of answering correctly. They complete this task once for each cut score.

After Round 1, bookmarks are tallied and shared among panelists for a given table. Those five or six panelists compare their Round 1 bookmark placements, discuss their rationales and understandings of the threshold student at each level, and review the procedures for placing bookmarks. After this discussion, they answer a brief questionnaire indicating readiness to begin Round 2.

In Round 2, panelists once again review the OIB, this time bypassing pages that clearly did not contribute to bookmark placement. They continue to discuss the contents of the items with others at their table but place their own bookmarks. Using the same set of guiding questions they used in Round 1, panelists place a single bookmark for each cut score.

After Round 2, bookmarks are tallied, and a median bookmark for each cut score is calculated. These results are shared with the entire panel, along with impact data – percentages of students who would be classified at each level as well as percentages classified at or above all but the lowest level. Panelists, led by their facilitator, discuss the bookmark distributions as well as the impact data. After the discussion, panelists complete a brief questionnaire indicating their readiness to begin Round 3.

In Round 3, panelists once again review the OIB as in Round 2, but with the knowledge of the impact of their bookmark placements. Each panelist enters a bookmark for each cut score and submits his or her final bookmarks. After receiving the final median bookmark placements and associated impact data, panelists complete a final questionnaire and evaluation form.

Software development.

MI staff consulted with Smarter Balanced staff to create a detailed development schedule defining essential tasks and timelines for the online standard-setting web site. Using the approved requirements documentation, MI developers designed the online application and finalized in-person application software, continuing to work closely with Smarter Balanced staff in accordance with the timeline shown in Table 1.

Table 1 Software Development Timeline.

Software Development Task/Deliverable	Begin	End
Gather requirements/modify application design	2/3/14	3/7/14
Develop online tool	3/10/14	4/25/14
QA application	4/28/14	5/16/14
Receive additional SBAC feedback	5/19/14	5/30/14
Implement changes/make updates	6/2/14	8/1/14
Deploy and field test application	8/4/14	8/15/14
Address issues	8/18/14	9/19/14
Demonstrate for Smarter Balanced	9/22/14	10/3/14
Go live	10/6/14	10/20/14

The basic elements of the system were the home page, item map, and ordered item booklet. The home page contained all instructions, links to external resources (e.g., the Smarter Balanced website to allow panelists to take practice tests), and links to internal resources (instructions on applying the Bookmark procedure, Common Core State Standards, and Achievement Level Descriptors). The item map had many features that could be turned on or off, depending on the round and nature of the task to be performed. The OIB contained the items as well as metadata, sample responses, and links to the ALDs.

The home page.

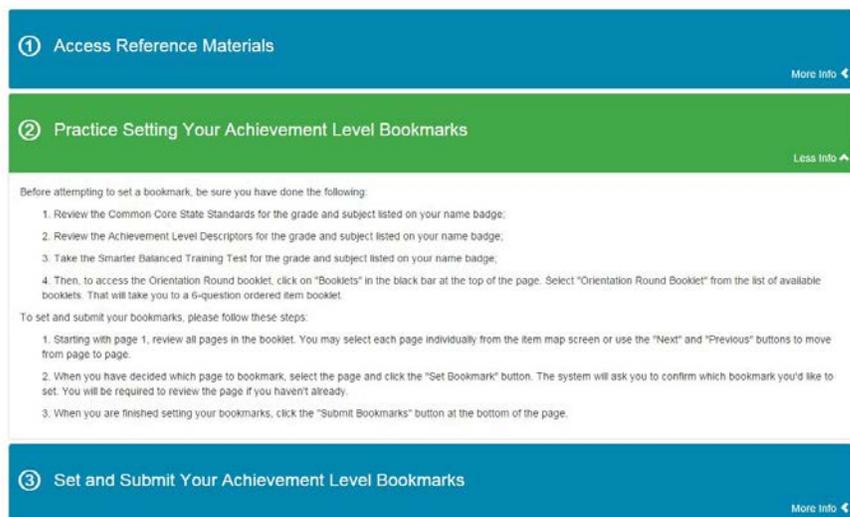
The home page contained all instructions plus links to additional resources. It consisted of four numbered, horizontal bars that could be expanded to reveal detailed information about each step of the process, as shown in Figure 1

Figure 1. Home Page With One Instruction Bar Expanded.

How to Use the Tool

During the In-Person Panel for Achievement Level Setting, you will need to complete the three steps listed below. To learn more about a step, click on the "More Info" button next to each one.

At certain points in this process, the system will present you with a short questionnaire. You must enter a response to all questions, and submit the questionnaire, before moving on to the next task. You may save your questionnaire at any time by clicking the "Save" button.



1 Access Reference Materials More Info

2 Practice Setting Your Achievement Level Bookmarks Less Info

Before attempting to set a bookmark, be sure you have done the following:

1. Review the Common Core State Standards for the grade and subject listed on your name badge;
2. Review the Achievement Level Descriptors for the grade and subject listed on your name badge;
3. Take the Smarter Balanced Training Test for the grade and subject listed on your name badge;
4. Then, to access the Orientation Round booklet, click on "Booklets" in the black bar at the top of the page. Select "Orientation Round Booklet" from the list of available booklets. That will take you to a 6-question ordered item booklet.

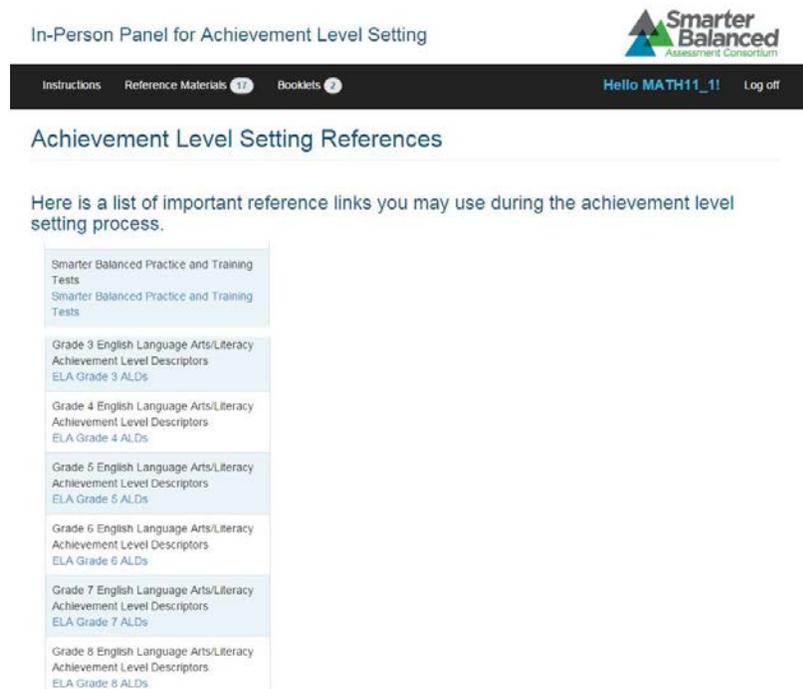
To set and submit your bookmarks, please follow these steps:

1. Starting with page 1, review all pages in the booklet. You may select each page individually from the item map screen or use the "Next" and "Previous" buttons to move from page to page.
2. When you have decided which page to bookmark, select the page and click the "Set Bookmark" button. The system will ask you to confirm which bookmark you'd like to set. You will be required to review the page if you haven't already.
3. When you are finished setting your bookmarks, click the "Submit Bookmarks" button at the bottom of the page.

3 Set and Submit Your Achievement Level Bookmarks More Info

The home page contained a list of all resource materials, accessible through hyperlinks, as shown in Figure 2.

Figure 2. List of Resources Accessible From Home Page.



In-Person Panel for Achievement Level Setting 

Instructions Reference Materials **17** Booklets **2** Hello MATH11_1! Log off

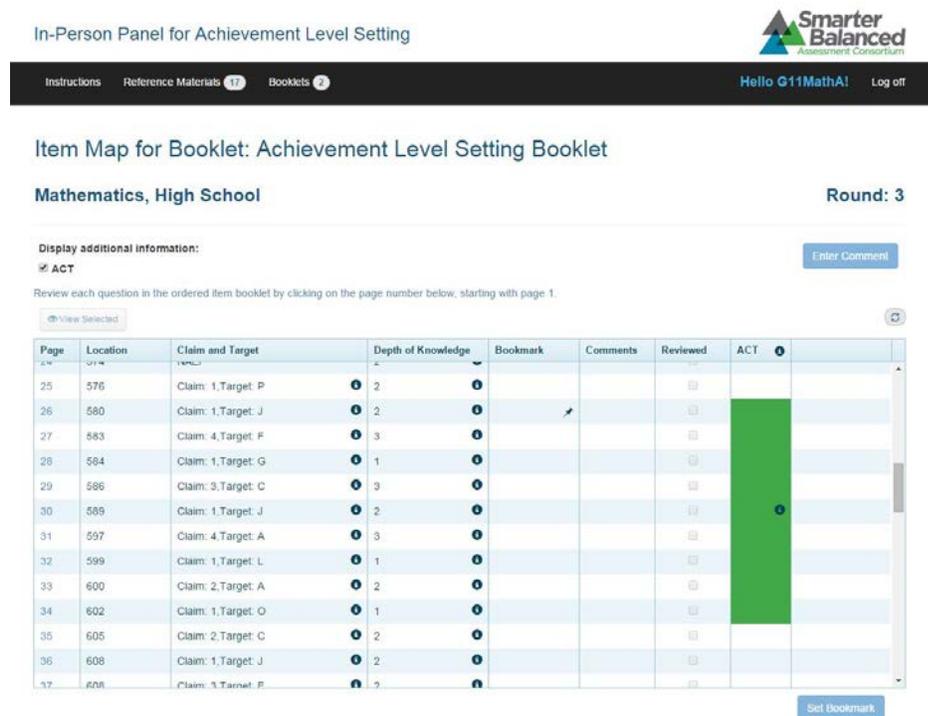
Achievement Level Setting References

Here is a list of important reference links you may use during the achievement level setting process.

- Smarter Balanced Practice and Training Tests
- Smarter Balanced Practice and Training Tests
- Grade 3 English Language Arts/Literacy Achievement Level Descriptors
ELA Grade 3 ALDs
- Grade 4 English Language Arts/Literacy Achievement Level Descriptors
ELA Grade 4 ALDs
- Grade 5 English Language Arts/Literacy Achievement Level Descriptors
ELA Grade 5 ALDs
- Grade 6 English Language Arts/Literacy Achievement Level Descriptors
ELA Grade 6 ALDs
- Grade 7 English Language Arts/Literacy Achievement Level Descriptors
ELA Grade 7 ALDs
- Grade 8 English Language Arts/Literacy Achievement Level Descriptors
ELA Grade 8 ALDs

The online item map page (see Figure 3) allowed panelists to review their progress, navigate through the ordered item booklet pages, access relevant item data, and submit their bookmarks. The Item Map drop-down menu allowed panelists to select and view their current results as well as the results from their previous round. Hovering over a comment indicator displayed the comments they submitted for a specific item during a round.

Figure 3. Sample Item Map.



Each OIB page displayed item-specific information including a preview of the item, item statistics, answer key(s), and associated passages and scoring rubrics. Additionally, the OIB page was designed to allow the panelist to make a comment about an item and store that comment for later review. The OIB page included a link to the Achievement Level Descriptor (ALD) for each test. Figure 4 shows a sample selected-response item, while Figure 5 shows the associated item information page, and Figure 6 shows a page for a constructed-response item (in this case, a performance task).

The item map and OIB pages were designed to allow panelists to toggle back and forth. Panelists could gain access to any page in the OIB by clicking that page number in the item map and return to the item map by clicking “Back to Item Map” at the top or bottom of the page. Each OIB page displayed the item, item statistics, rubrics, passages, and sample responses. Additionally, the OIB page was designed to allow the panelist to specify a cut score or navigate to the next or previous OIB page.

All items presented in the OIB were in static, portable data file (pdf) format rather than in interactive format as they had been in the practice tests on the Smarter Balanced website or as administered in the spring 2014 field test. The decision to render items in a static format was based on concerns about the rendering of the interactive versions of items on an uncontrollable array of online panelist devices and browsers. By displaying a static image or PDF of the item, it was possible to ensure that every panelist saw exactly the same rendering of the item for review independent of the platform used.

Figure 4. Sample OIB Page With Selected-Response Item.

Ordered Item Booklet: Achievement Level Setting Booklet

Mathematics, High School Page: 06 Round: 2

← Back to Item Map

Set Bookmark Enter Comment

← Previous Next →

Item Question Information Passages and Other Materials 0 Achievement Level Descriptors 1

This item permits calculator use.

12155

The formula for the rate at which water is flowing is $R = \frac{V}{t}$, where

- R is the rate,
- V is the volume of water measured in gallons (g), and
- t is the amount of time, in seconds (s), for which the water was measured.

Select an appropriate measurement unit for the rate.

Ⓐ gs

Ⓑ $\frac{g}{s}$

Ⓒ $\frac{s}{g}$

Ⓓ $\frac{1}{sg}$

← Back to Item Map

Figure 5. Item Information Page.

In-Person Panel for Achievement Level Setting 

Instructions Reference Materials **11** Booklets **2** Hello G11MathA! Log off

Ordered Item Booklet: Achievement Level Setting Booklet

Mathematics, High School Page: 31 Round: 2

[Back to Item Map](#)

Set Bookmark Enter Comment [Previous](#) [Next](#)

Item **Question Information** Passages and Other Materials **3** Achievement Level Descriptors **1**

Page	31
Location	597
Claim and Target	Claim: 4, Target: A i
Depth of Knowledge	3 i
Answer Key	See Passages and Other Material Tab

[Back to Item Map](#)

Figure 6. OIB Page For Constructed-Response Item.

In-Person Panel for Achievement Level Setting 

Instructions Reference Materials **17** Booklets **2** Hello G11ELAA! Log off

Ordered Item Booklet: Achievement Level Setting Booklet

English Language Arts/Literacy, High School Page: 03 Round: 1

[Back to Item Map](#)

Set Bookmark
 Enter Comment

[Previous](#) [Next](#)

Item
 Question Information
 Passages and Other Materials **3**
 Achievement Level Descriptors **1**

62019 ?

Student Directions for Part 2

You will now review your sources, take notes, and plan, draft, revise, and edit your article. You may use your notes and refer to the sources. Now read your assignment and the information about how your article will be scored; then begin your work.

Your assignment:
 After completing your research, you share your findings with your teacher. She is impressed with your work. As a final project for your psychology class, everyone must write an article for the Psychology Club's website. Your teacher suggests writing about malleable intelligence, and you decide this is a good idea. The audience for your article will be other students, teachers, and parents.

Using more than one source, craft a thesis to explain the concept of malleable intelligence. Once you have a thesis, select the most relevant information to support your thesis. Then, write a multi-paragraph explanatory article explaining your thesis. Clearly organize your article and elaborate on your ideas. Develop your ideas clearly and use your own words, except when quoting directly from the sources. Be sure to reference the source title or number when quoting or paraphrasing details or facts from the sources.

Explanatory Scoring
 Your explanatory article will be scored using the following:

1. **Organization/purpose:** How well did you state your thesis, and maintain your thesis with a logical progression of ideas from beginning to end? How well did you narrow your thesis so you can develop and elaborate the conclusion? How well did you consistently use a variety of transitions? How effective was your introduction and your conclusion?
2. **Elaboration/evidence:** How well did you integrate relevant and specific information from the sources? How effective were your elaborative techniques? How well did you clearly state ideas using precise language that is appropriate for your audience and purpose?
3. **Conventions:** How well did you follow the rules of grammar usage, punctuation, capitalization and spelling?

Now begin work on your article. Manage your time carefully so that you can:

- plan your multi-paragraph article
- write your multi-paragraph article
- revise and edit the final draft of your multi-paragraph article

Word-processing tools and spell check are available to you.

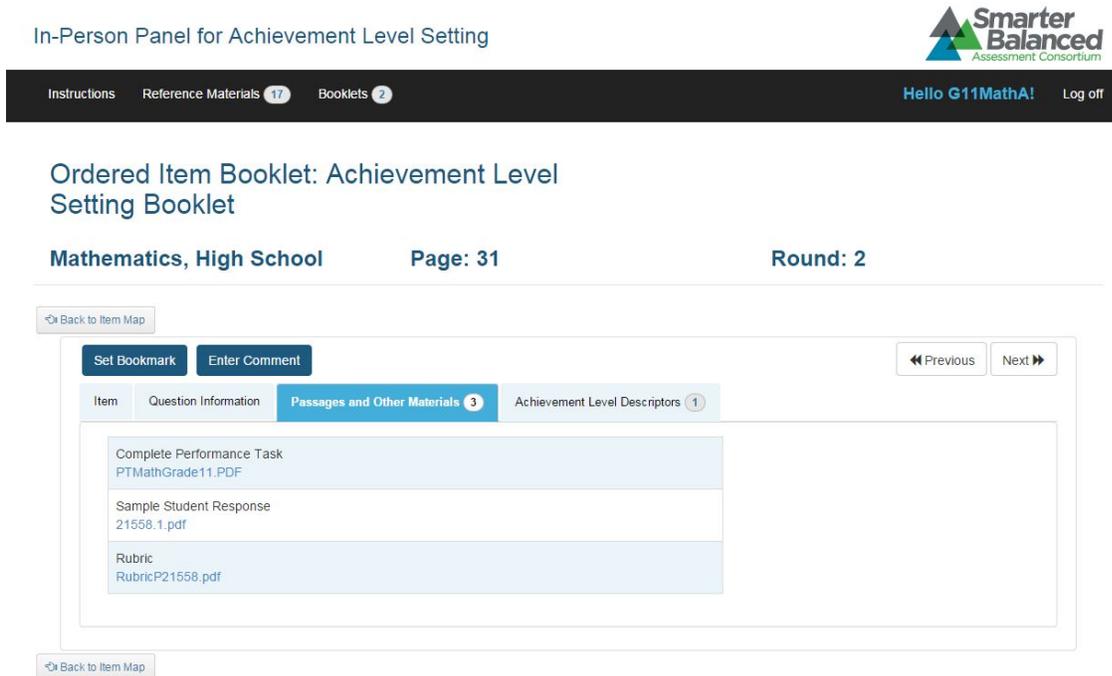
For Part 2, you are being asked to write a multi-paragraph article, so please be as thorough as possible. Type your response in the space provided. The box will expand as you type.

Remember to check your notes and your prewriting/planning as you write and then revise and edit your article.

B I U I_x | ¶ ☰ ☲ | ✂ 📄 ↶ ↷ ⌂ Ω

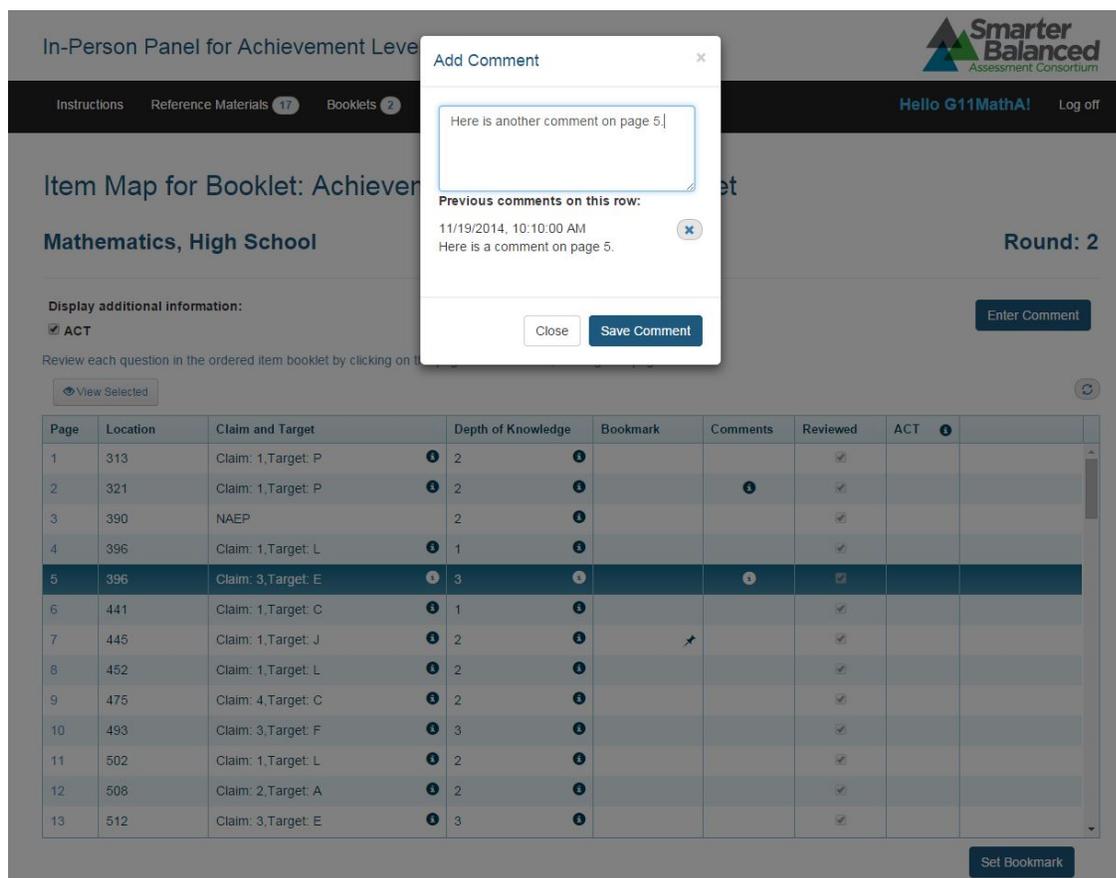
By clicking “Passages and Other Materials,” panelists could see resource materials such as reading or listening passages, sample student responses, and scoring rubrics, as shown in Figure 7.

Figure 7. OIB Page Showing Links to Performance Task, Sample Student Response, and Rubric.



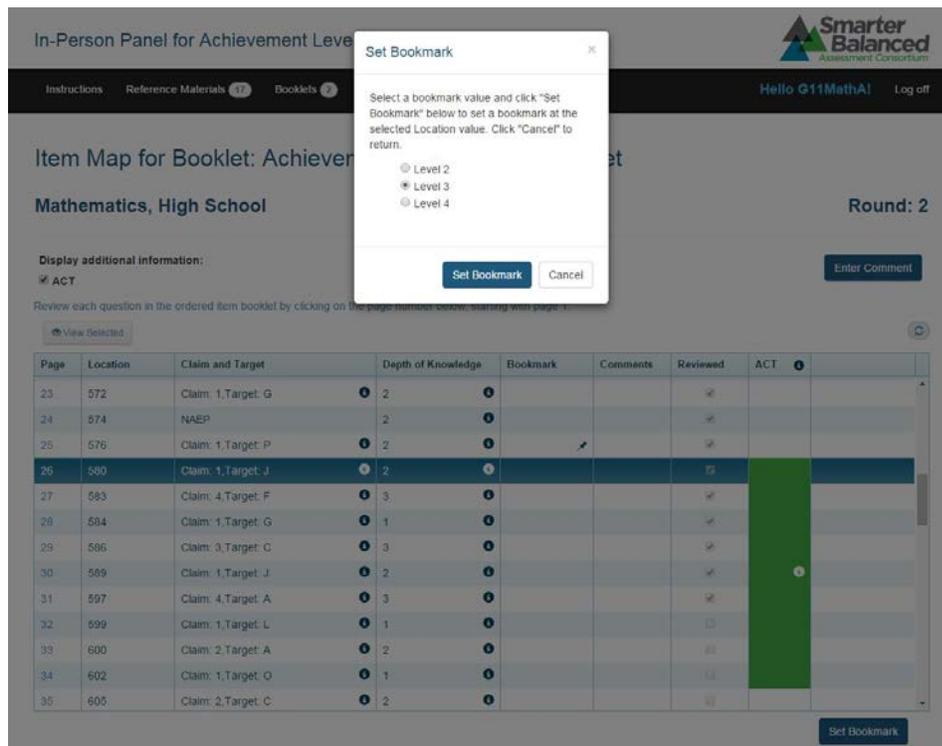
The system was designed to allow panelists to leave comments on any test item by clicking on “Comments” in the OIB or in the appropriate row of the item map. These comments were intended to be used during inter-round discussions of the items by the in-person panelists or for the online panelists if they needed to leave the task and resume it later. Figure 8 illustrates the “Comment” function.

Figure 8. Comment.



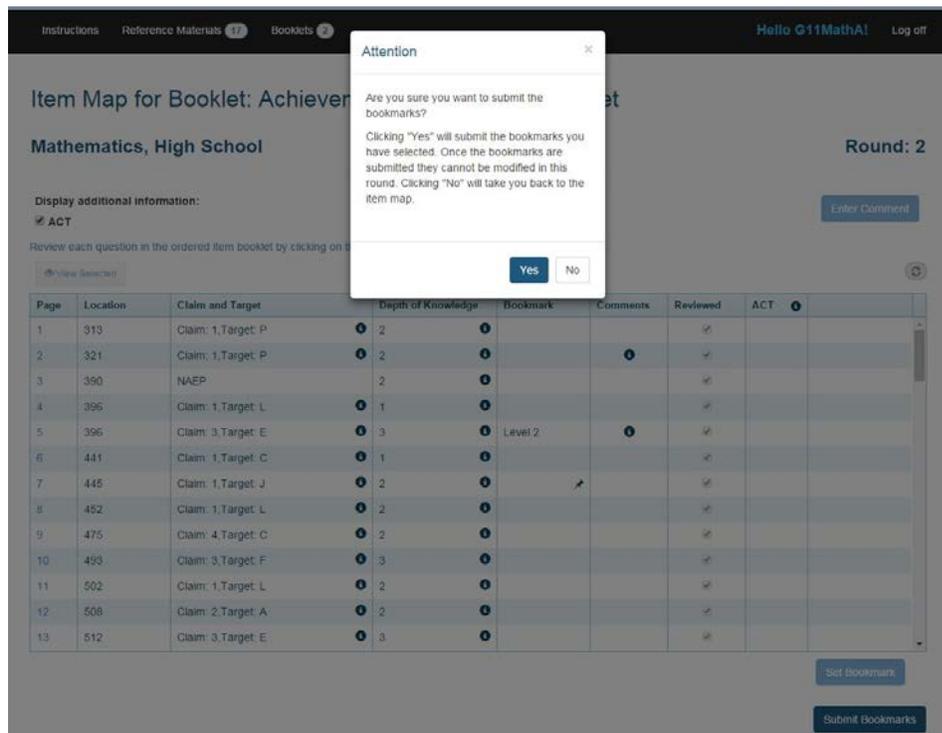
After reviewing items, panelists could enter a bookmark by clicking either on the page in the OIB or in the appropriate row of the item map. Figure 9 illustrates the “Enter Bookmark” function. After entering all bookmarks (a single bookmark for Level 3 for the online panel activity or bookmarks for Levels 2, 3, and 4 for the in-person workshop), panelists were prompted to review their work and make sure they were ready to submit their bookmark(s), as shown in Figure 10.

Figure 9. Set Bookmark Dropdown Box in the Item Map.



The screenshot shows the 'Item Map for Booklet: Achievement Mathematics, High School' interface. A 'Set Bookmark' dialog box is open in the center, containing the following text: 'Select a bookmark value and click "Set Bookmark" below to set a bookmark at the selected Location value. Click "Cancel" to return.' Below the text are three radio button options: 'Level 2', 'Level 3' (which is selected), and 'Level 4'. At the bottom of the dialog are 'Set Bookmark' and 'Cancel' buttons. The background interface shows a table with columns: Page, Location, Claim and Target, Depth of Knowledge, Bookmark, Comments, Reviewed, and ACT. A green vertical bar highlights the 'ACT' column for the selected row (Page 26, Location 580).

Figure 10. Submitting Bookmarks.



The screenshot shows the same 'Item Map for Booklet: Achievement Mathematics, High School' interface. An 'Attention' dialog box is open in the center, containing the following text: 'Are you sure you want to submit the bookmarks? Clicking "Yes" will submit the bookmarks you have selected. Once the bookmarks are submitted they cannot be modified in this round. Clicking "No" will take you back to the item map.' Below the text are 'Yes' and 'No' buttons. The background interface shows the same table as in Figure 9, but with a 'Submit Bookmarks' button visible at the bottom right.

Design and Implementation of the Online Panel

The purpose of the online panel was to broaden the input into the process of making decisions about cut scores. In addition, the online panel allowed thousands of people to examine the tests and to express their opinions. The original proposal called for an online panel of 840 individuals; subsequent negotiations increased that number significantly. The final plan called for the contractor to support up to 250,000 online panelists. The intent was to have these individuals review a single ordered item booklet (OIB) and place a bookmark to indicate the location of the Level 3 cut score. The OIBs and support materials were the same as those used in the in-person workshop, but without the extensive training, interaction, and support provided to in-person workshop panelists. The online ALS differed from the typical Bookmark application described above in that panelists did not meet or receive successive rounds of feedback.

Online panel activities.

Online panel activities commenced with recruitment (see Chapter 5), which began in April, 2014. Staff of McGraw-Hill Education (CTB), working in concert with Smarter Balanced staff and staff of Hager Sharp (H-S), crafted messages, first for educators and later for the general public, to alert them to the opportunity and explain the logistics.

Staff of Measurement Incorporated (MI) developed the software to support the online experience. That software included a home page, directions, links to reference materials, and digital OIBs, described below. Details of the software development and implementation are included in Appendix B.

Prior to the launch of the online panel on October 6, 2014, MI staff conducted a field test on August 14-15. That activity is described in Chapter 7 and summarized briefly here. Panelists for the online field test were 40 MI readers who logged in to a 30-minute webinar explaining the purpose of the activity and providing a brief introduction to the bookmark procedure. Panelists then had 48 hours to review an OIB for one of four ELA tests (grades 4, 6, 8, or 11) and enter a single bookmark, an activity that was estimated to take about three hours. Most who completed the activity took longer than three hours. Feedback from the panelists was collected via Survey Monkey, analyzed, and used to modify the process for October. Results are presented in Appendix D.

Conduct and results of the online panel.

Online panelists signed up for one of six 48-hour windows, the first of which started on October 6. Ultimately, all windows were extended, and the final date was moved to October 18. By October 6, 10,099 individuals had registered to participate. Of that number, 5,840 logged in, and 2,660 placed a bookmark. **Online materials and software were deployed successfully, and capacity was more than adequate for application use.**

Results for online panelists entering a bookmark are presented in Table 2. Impact (percent of students who would score at or above the Level 3 cut score) is presented in Table 3. Impact is not

reported for groups smaller than 25 online panelists. These results were also shared with the in-person workshop panelists and with the cross-grade review committee.

Table 2. Numbers of Online Panelists, by Role, Grade, and Subject

Grade	Teachers		Administrators		Higher Education		Other	
	ELA	Math	ELA	Math	ELA	Math	ELA	Math
3	151	167	67	37	9	5	31	30
4	89	124	31	28	2	4	16	22
5	96	114	31	35	5	5	12	21
6	66	91	11	22	4	8	9	17
7	70	100	12	22	4	5	6	8
8	87	115	27	39	4	7	11	22
11	193	267	55	64	60	83	13	26

Table 3. Impact of Online Panel Bookmark Placements: Percent of Students At or Above Level 3

Grade	Teachers		Administrators		Higher Education		Other	
	ELA	Math	ELA	Math	ELA	Math	ELA	Math
3	51%	54%	39%	50%			47%	45%
4	44%	43%	31%	52%				
5	61%	46%	65%	37%				
6	48%	38%						
7	57%	27%						
8	48%	18%	43%	18%				
11	55%	26%	48%	28%	56%	26%	58%	27%

The concept of an online panel is an innovation introduced to address the scale of the Smarter Balanced project and its number and variety of stakeholders. In addition to allowing wider achievement level setting participation, the online panel approach promotes deeper understanding of the content standards and of the tasks used in schools. It also provided in-person panelists with feedback from a broader perspective. Online values for the level 2/3 cut were very similar to those for initial in-person values. This suggests that the approach should be explored in future standard setting venues in a manner that could provide wider participation and save on travel costs.

Design and Implementation of the In-Person Workshop

As noted above, the bookmark procedure was used in the in-person workshop. The workshop took place at the Hilton Anatole in Dallas, Texas, on October 13-19, 2014. There were three waves of panels: the first wave, grade 11, began on Monday morning, October 13, and went through noon October 15; the second wave, grades 6–8, began on Wednesday morning, October 15, and went through noon October 17; the final wave, grades 3–5, began on Friday morning, October 17, and went through noon October 19. Table 4 summarizes the numbers of panelists by subject and grade. Table 5 summarizes the agenda for each 2.5-day session. Appendix D contains a detailed agenda for each day of the workshop.

Table 4. In-Person Workshop Panelists by Subject and Grade

Grade	English Language Arts/Literacy		Mathematics	
	Planned	Obtained	Planned	Obtained
3	1 panel of 30	1 panel of 26	1 panel of 30	1 panel of 30
4	1 panel of 30	1 panel of 27	1 panel of 30	1 panel of 29
5	1 panel of 30	1 panel of 27	1 panel of 30	1 panel of 29
6	1 panel of 30	1 panel of 30	1 panel of 30	1 panel of 30
7	1 panel of 30	1 panel of 27	1 panel of 30	1 panel of 30
8	1 panel of 30	1 panel of 30	1 panel of 30	1 panel of 29
11	2 panels of 36	2 panels of 34	2 panels of 36	2 panels of 35
Total	252	235	252	247
Grand Total	504	482 (95.6%)		

Table 5. High-Level Agenda for Each In-Person Workshop.

Day - Time	Event(s)
Day 1 A.M.	Welcome; overview, training on CCSS, ALDs, tests
Day 1 P.M.	Review of Ordered Item Booklet
Day 2 A.M.	Orientation to the Bookmark Procedure; complete Round 1
Day 2 P.M.	Review Round 1; complete Round 2
Day 3 A.M.	Review Round 2; complete Round 3; evaluate process

Recruitment and selection of panelists.

Recruitment of panelists for the In-Person Workshop began April 15. K-12 State Leads, Higher Education Leads, and Teacher Involvement Coordinators received communication tools developed by the contractor and approved by Smarter Balanced to enable them to recruit teachers (general as well as teachers of English language learners and students with disabilities), school administrators, higher education faculty, business and community leaders, and parents. Each Smarter Balanced state had 20–25 positions to fill, giving each state an opportunity to have at least one representative for each of the 14 tests.

Preparation of materials.

Staff of MI and CTB prepared the following training materials, all of which can be found in Appendix C:

- Introductory PowerPoint presentation to orient panelists to the goals and tasks of the workshop
- Common Core State Standards – up-to-date versions of the subject/grade-specific content standards as well as guidelines to their use in the achievement level setting activity
- Achievement Level Descriptors – up-to-date versions of the ALDs for the specific subject and grade for each panel
- Practice Test – using the online version of the Smarter Balanced practice tests for each grade and subject
- Orientation to the ordered item booklet – PowerPoint presentation designed to show panelists what to look for and questions to ask as they reviewed items in the OIB
- Orientation to the Bookmark procedure – PowerPoint presentation designed to show panelists how Bookmark works and specifically how panelists were to implement the procedure in a computer-based environment
- Bookmark Orientation Round – an exercise involving a 6-page OIB that panelists reviewed prior to entering a single bookmark and discussing their placements in a large-group setting.
- Readiness Form – a multipart form that asked panelists at several key points during the process how well they understood the process they were implementing and how ready they were to proceed to the next step
- Evaluation Form – a series of statements about the training, environment, and conduct of the workshop that the panelists responded to on a graded scale (such as Strongly Agree to Strongly Disagree)

MI and CTB staff drafted all training materials and submitted them to Smarter Balanced staff and the external auditor for review in advance of the workshop. Final versions of all training materials reflect the comments and recommendations of these reviews and were approved by Smarter Balanced leadership prior to use. All training materials are included in Appendix C.

Training of facilitators and table leaders.

In advance of the in-person workshop, staff of MI and CTB prepared a detailed facilitator script which was reviewed and approved by Smarter Balanced. Staff identified as facilitators studied the scripts and participated in in-house training sessions the week prior to the in-person workshop. In addition, Mr. Ricardo Mercado of CTB conducted a two-hour facilitator training session on Sunday night, October 12, on Tuesday night, October 14, and on Thursday night, October 16, as facilitators for each wave arrived in Dallas. At the same time, Dr. Jennifer Lord-

Bessen of CTB provided a two-hour orientation for table leaders who had been identified in advance by their State Leads. Training materials for those sessions are included in Appendix C.

Orientation and training.

Using the training materials approved by Smarter Balanced, MI and CTB staff provided large-group and small-group training. For the opening session, Dr. Joe Willhoft gave the welcome and charge. Dr. Michael Bunch of MI provided specific training on the content standards, ALDs, and practice tests. Dr. Daniel Lewis of CTB provided the orientation to the Bookmark procedure. At the end of each training session, panelists completed a portion of the Readiness Form (see Appendix C).

In-Person Workshop panelists were encouraged to review the appropriate ALDs and CCSS standards prior to coming to the workshop. However, it was not assumed that all had done so, and panelists were given an opportunity not only to review the materials on site but to discuss them in a large-group setting. They had an opportunity to indicate on the Readiness Form just how familiar they were with those materials. No panelist was permitted to advance to item review without indicating familiarity with the ALDs and content standards and indicating readiness to proceed.

The afternoon of Day 1 was devoted entirely to review of the OIB. In addition to being oriented to the software, panelists were introduced to the test items themselves. They spent the entire afternoon annotating items, using the Comments function of the software, and discussing items with others at their table in terms of the first two guiding questions. While this activity had been scheduled to end at 5 p.m. on Day 1, all panels required additional time and received from 30 to 60 minutes to complete the task at the beginning of Day 2, following orientation to the bookmark procedure.

At the beginning of Day 2, all panelists assembled in the ballroom for orientation to the Bookmark procedure. Dr. Daniel Lewis, Chief Research Advisor at CTB and co-creator of the Bookmark procedure, provided the orientation and answered questions. Following the orientation to the Bookmark procedure, panelists adjourned to their small groups to gain first-hand experience in setting a bookmark through a practice exercise. This exercise consisted of a 6-page OIB with items of varying difficulty. Each panel had access to two facilitators who oriented panelists to the computers and software and showed them how to navigate the OIB. Panelists then had several minutes to review the six items and enter a bookmark. The facilitator then led a discussion focusing on how many panelists chose each page to place their bookmarks. Following this discussion, panelists completed a section of their Readiness Forms, indicating their readiness to begin Round 1.

Round-by-round item review and discussion.

Panelists were invited to work through their on-screen OIBs and discuss the items with others at their table. They were able to discuss their opinions with one another at their table as much as they wished, but when they entered a bookmark, it was to be their bookmark, not that of the table. They started by placing a bookmark for Level 3, then Level 4, and finally, Level 2. After placing three bookmarks, panelists were dismissed for lunch, during which time CTB staff tallied

bookmarks but did not provide reports to the panelists. Results are shown in Table 6 in terms of median bookmark placement for each subject, grade, and level. Complete results, including distributions of bookmark placements, are included in Appendix D.

Table 6. Results of Round 1 of Bookmark Placement (Entries are Median Page Numbers).

Subject/Grade	ELA			Math		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
ELA 3	16.0	38.0	58.5	22.0	47.0	69.5
ELA 4	20.0	42.0	60.0	12.0	33.0	69.0
ELA 5	13.0	27.0	63.0	21.5	50.0	65.5
ELA 6	15.0	35.0	63.0	18.0	37.5	61.0
ELA 7	16.0	41.0	69.0	21.5	42.5	63.0
ELA 8	19.0	39.5	68.0	18.0	39.0	58.0
EALA 11	21.5	45.0	66.0	19.0	48.5	69.0

Panelists, upon returning from lunch, were directed to share their Round 1 bookmark placements with others at their table, discuss their rationales for placing those bookmarks, and compare approaches as well as comments they had left on the item map. The facilitator then introduced and led a discussion on the bookmark placements of the online panel. Once they completed their discussions, panelists completed the portion of the Readiness Form that indicated they were ready to begin Round 2.

In Round 2, panelists proceeded as in Round 1, conferring with others at their table but entering their own bookmarks. When they entered three bookmarks and submitted them, they were free to log out for the day. Results of Round 2 are shown in Table 7. Complete results, including bookmark distributions and interquartile ranges, are shown in Appendix D.

Table 7. Results of Round 2 of Bookmark Placement (Entries are Median Page Numbers).

Grade	ELA			Math		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
3	19.0	38.0	57.5	28.0	49.0	70.0
4	20.0	44.0	63.0	9.0	32.0	71.0
5	14.0	27.0	61.0	20.0	50.5	64.0
6	15.0	36.5	63.0	16.5	37.0	60.0
7	16.0	37.5	69.5	18.0	46.0	61.0
8	17.0	40.5	66.5	17.0	40.0	60.0

11	22.0	42.0	65.0	20.0	50.0	69.0
----	------	------	------	------	------	------

Panelists returned the morning of the third day to see the results of Round 2. The facilitator led a discussion of the range of bookmark placements, corresponding cut scores, and percentages of students classified at each level, based on the Round 2 cut scores. Once again, the facilitator showed the online panel results, this time in terms of percentages of students at or above Level 3, based on online panelists' bookmark placements. A room-wide discussion ensued. Finally, facilitators revealed the impact for the next grade up; i.e., panelists in grade 8 were able to see the final impact of the cut scores set by grade 11 panels, panelists in grade 7 were able to see the Round 2 results for grade 8, and so on down to grade 3. By virtue of being first, grade 11 panelists did not get to see results of any other in-person workshop panels.

After review and discussion of all results, panelists completed the final section of their Readiness Forms and began Round 3. They completed Round 3 as they had Round 2, bypassing many pages which no one had recommended in previous rounds and keeping or changing their bookmark placements depending on their response to the discussion. Each panelist entered three bookmarks and then submitted those bookmarks for analysis. Results of Round 3 are shown in Tables 8 (bookmark placement) and 9 (scale score cuts and percentages of students at or above each level).

Table 8. Results of Round 3 of Bookmark Placement (Entries are Median Page Numbers).

Subject/Grade	ELA			Math		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
3	13.0	33.0	54.0	27.0	47.0	70.0
4	19.0	43.0	62.0	15.0	39.0	71.0
5	11.0	27.0	63.0	19.0	50.0	64.0
6	14.5	34.5	60.5	18.0	45.5	61.5
7	16.0	38.0	66.0	17.0	45.0	61.0
8	18.0	39.5	68.0	16.0	40.0	60.0
11	19.0	42.0	65.0	19.5	48.0	68.0

Table 9. Round 3 Cut Score Recommendations: Scale Score Cuts and % At or Above

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 3 English Language Arts/Literacy	362	66.5%	427	40.1%	485	19.1%
Grade 3 Mathematics	383	67.3%	436	38.9%	506	10.8%
Grade 4 English Language Arts/Literacy	413	64.4%	470	42.0%	530	18.9%
Grade 4 Mathematics	400	77.6%	470	44.7%	541	15.6%
Grade 5 English Language Arts/Literacy	406	78.7%	450	64.0%	574	16.9%
Grade 5 Mathematics	459	63.5%	532	31.4%	583	13.8%
Grade 6 English Language Arts/Literacy	466	66.6%	527	42.2%	614	12.2%
Grade 6 Mathematics	491	58.3%	561	29.4%	603	15.6%
Grade 7 English Language Arts/Literacy	474	68.2%	547	40.1%	660	6.6%
Grade 7 Mathematics	513	53.1%	609	19.3%	674	5.8%

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 8 English Language Arts/Literacy	471	76.4%	543	50.9%	663	10.2%
Grade 8 Mathematics	534	51.3%	605	25.6%	683	7.4%
Grade 11 English Language Arts/Literacy	490	72.9%	565	47.6%	677	12.1%
Grade 11 Mathematics	533	62.6%	644	28.0%	740	8.0%

After entering their Round 3 bookmarks, panelists took a short break and returned to review the final cut scores, complete a final questionnaire, and then evaluate the process, using online evaluation forms. Results of those questionnaires and evaluation forms are summarized in Tables 10 and 11. Questionnaire and evaluation results for individual panels are included in Appendix D.

Table 10. Round 3 Questionnaire Results: Confidence in Cut Scores Recommended (Discounting Blanks)

How confident are you about the three bookmarks you just entered?

Bookmark	Very Confident		Uncertain	Very Uncertain		Total
	Confident	Confident		Uncertain	Very Uncertain	
Level 2	222 (47%)	237 (51%)	10 (2%)	0 (0%)	469	
Level 3	234 (50%)	220 (47%)	15 (3%)	0 (0%)	469	
Level 4	245 (52%)	217 (46%)	7 (1%)	0 (0%)	469	

Table 11. Summary of Round 3 Evaluation Responses (Discounting Blanks)

Evaluation Statement	Strongly Agree	Agree	Disagree	Strongly Disagree	Total
The orientation provided me with a clear understanding of the purpose of the meeting.	253 (58%)	170 (39%)	13 (3%)	2 (0%)	438
The workshop leaders clearly explained the task.	245 (56%)	161 (37%)	25 (6%)	7 (2%)	438
The training and practice exercises helped me understand how to perform the task.	247 (56%)	174 (40%)	16 (4%)	1 (0%)	438

Evaluation Statement	Strongly Agree	Agree	Dis-agree	Strongly Disagree	Total
Taking the practice test helped me to understand the assessment.	231 (53%)	192 (44%)	14 (3%)	1 (0%)	438
The Achievement Level Descriptions were clear and useful.	199 (45%)	216 (49%)	21 (5%)	2 (0%)	438
The large and small group discussions aided my understanding of the process.	300 (68%)	132 (30%)	4 (1%)	2 (0%)	438
The time provided for discussions was appropriate.	230 (53%)	184 (42%)	23 (5%)	1 (0%)	438
There was an equal opportunity for everyone in my group to contribute his/her ideas and opinions.	292 (67%)	135 (31%)	8 (2%)	3 (1%)	438
I was able to follow the instructions and complete the rating tasks accurately.	284 (65%)	151 (34%)	1 (0%)	2 (0%)	438
The discussions after the first round of ratings were helpful to me.	273 (62%)	151 (34%)	12 (3%)	2 (0%)	438
The discussions after the second round of ratings were helpful to me	270 (62%)	156 (36%)	11 (3%)	1 (0%)	438
The information showing the distribution of student scores was helpful to me.	220 (50%)	200 (46%)	13 (3%)	4 (1%)	437
I am confident about the defensibility and appropriateness of the final recommended cut scores.	203 (46%)	202 (46%)	27 (6%)	6 (1%)	438
The facilities and food service helped create a productive and efficient working environment.	324 (74%)	104 (24%)	10 (2%)	0 (0%)	438

Data analysis and reporting.

As panelists entered and submitted bookmarks, the data flowed directly from their computers to servers MI had set up prior to the start of the workshop. Staff from CTB, using BookmarkPro software, received the data, analyzed them, and produced reports that facilitators shared at the beginning of the next round. A full set of reports is included in Appendix D.

Design and Implementation of the Cross-Grade Review Committee

The vertical articulation committee was renamed the cross-grade review committee to reflect more clearly the nature of their task, which was to review all cut scores and impact across all grades within a given subject and make adjustments where necessary to prevent or minimize large discontinuities in impact across grades. For example, if 50 percent of students in grades 5,

6, and 8 were at or above Level 3, but only 40 percent of grade 7 students were above Level 3, such a discrepancy would need to be examined.

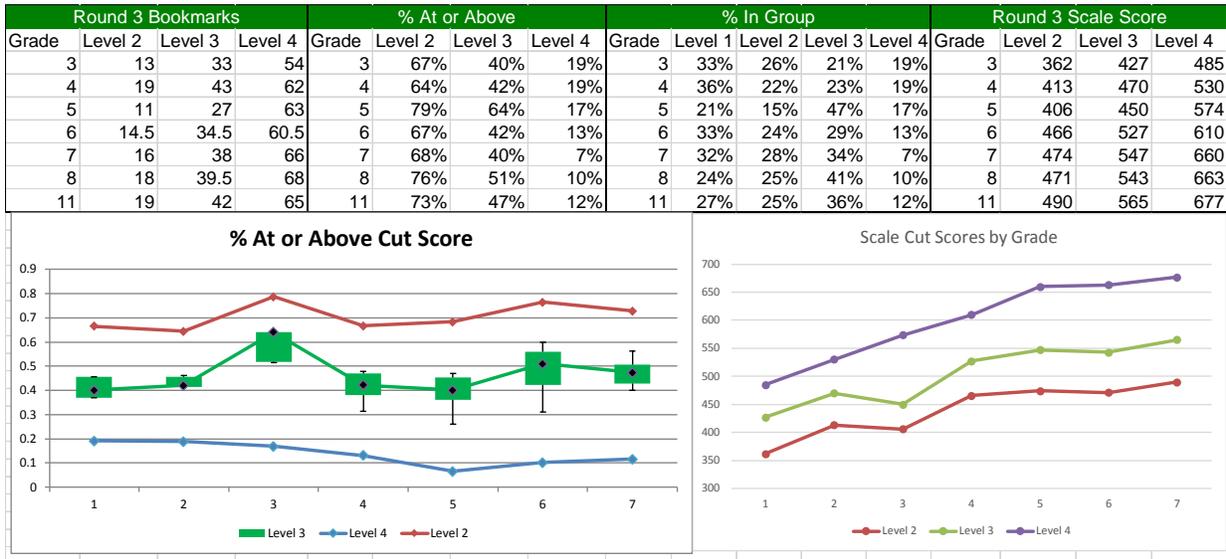
The committees (32 members each for ELA and mathematics) met on October 20, 2014. Dr. Bunch provided an introduction to the tasks and ground rules. The complete presentation is included in Appendix B and is summarized here.

Trends in percentages of students at or above a given level tend to be of one of three types: either more and more students reach a given level over time or across grades (i.e., generally increasing), the same percentages of students reach a given level year after year or from one grade to the next (level), or fewer and fewer students reach a given level over time or across years (generally decreasing). Trends that go up one year and down the next, or up for one grade and down for the next, are much more difficult to explain (though there may be legitimate reasons for such trends). The task of the cross-grade review committee was to investigate any discontinuities and determine whether they were accurate reflections of reality or indications that one or more panels had been overly stringent or overly lenient.

Dr. Bunch explained that the process would include review of actual OIBs and impact data, starting with grades 8 and 11. Any panelist would be welcome to recommend changing any cut score, although a panelist from the grade directly involved or from an adjacent grade would be the preferred initiator of any recommended change. He explained the process for introducing and seconding a motion to change a cut score, to be followed by discussion and a vote. Given that any change would alter the work of a panel of 30 to 36 people, a 2/3 super majority was required to pass any recommended change.

After the orientation, 32 mathematics panelists reconvened in an adjacent room, while the 32 ELA panelists remained in the room in which the orientation had taken place. In both rooms, computers from the previous week's in-person workshop were still in place with all software still loaded. For each subject, all seven OIBs and all support materials used by in-person workshop panelists were available. Whenever anyone suggested a change, the facilitator (Dr. Bunch for ELA and Dr. Lewis for mathematics) was able to show on a large screen in the front of the room a projected image of how that change would affect impact. An example of the on-screen graphic is shown in Figure 11.

Figure 11. Cross-Grade Review Graphic.



In Figure 11, the four tables at the top represent the Round 3 median bookmark placements, the percentages at or above Levels 2-4 based on those bookmark placements, the resulting percentages of students classified into each level, and the Round 3 bookmark placements translated into temporary scale scores. The graph on the bottom left reflects the impact in the second table, with the black dots representing the medians, the green boxes representing the interquartile ranges, and the black vertical lines representing the 10th and 90th percentiles. Panelists could recommend changing any bookmark placement in the first table, and all other tables, as well as the two graphs at the bottom, would immediately change accordingly.

Panelists began by reviewing cut scores for grades 8 and 11 and then worked their way down through the middle and elementary grades. By the end of the day, the ELA committee had made 8 changes, and the mathematics committee had made 11. Final results for the two committees are shown in Table 12, with changes from Round 3 highlighted in yellow.

Table 12. Cross-Grade Review Results

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 3 English Language Arts/Literacy	362	66.5%	427	40.1%	485	19.1%
Grade 3 Mathematics	381	68.3%	436	38.9%	501	12.1%
Grade 4 English Language Arts/Literacy	413	64.4%	470	42.0%	530	18.9%

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 4 Mathematics	413	72.3%	487	36.5%	551	12.6%
Grade 5 English Language Arts/Literacy	434	69.7%	494	47.1%	574	16.9%
Grade 5 Mathematics	459	63.5%	532	31.4%	583	13.8%
Grade 6 English Language Arts/Literacy	453	71.3%	527	42.2%	614	12.2%
Grade 6 Mathematics	491	58.3%	570	26.1%	609	14.0%
Grade 7 English Language Arts/Literacy	474	68.2%	547	40.1%	644	9.5%
Grade 7 Mathematics	513	53.1%	596	23.2%	674	5.8%
Grade 8 English Language Arts/Literacy	482	73.1%	562	43.3%	663	10.2%
Grade 8 Mathematics	534	51.3%	616	22.1%	683	7.4%
Grade 11 English Language Arts/Literacy	488	73.3%	578	42.8%	677	11.6%
Grade 11 Mathematics	565	48.3%	650	26.4%	740	5.8%

Approval by Chiefs

Subsequent to the completion of the cross-grade review, Smarter Balanced and contractor staff prepared to present results to the Chiefs for review and approval. On November 6, Chiefs met in Chicago to review the results. While endorsing the work of the panels, the Chiefs did not vote on the cut scores. A second meeting was scheduled for November 14, in conjunction with the meeting of the Council of Chief State School Officers (CCSSO) in San Diego. Meanwhile, Smarter Balanced staff prepared options to present to the Chiefs at the November 14 meeting, incorporating evidence from recent studies conducted by the National Assessment Governing Board (NAGB). In addition, Smarter Balanced staff created a new reporting scale, replacing the temporary scale used throughout achievement level setting and cross-grade review. While the temporary scale had a range of 100 to 900, the final scale had a range of 2000 to 3000 and can be easily derived from the temporary scale by adding 2000 to the original scale. Thus, for example, the grade 11 mathematics Level 2 cut score of 565 would translate to a final score of 2565.

The two options presented to the Chiefs at the November 14 meeting consisted of the results shown in Table 12 and those same results moderated in the direction of the NAGB results. Specifically, while working within a range of plus-or-minus one standard error of measurement of the cut scores recommended by the cross-grade review committee, Smarter Balanced staff recommended ELA cut scores that were higher and mathematics cut scores that were lower than those recommended by the cross-grade review committee. These modifications kept recommended cut scores within or very close to the one SEM range, approximated NAGB results, and brought ELA and mathematics impacts into closer alignment with each other. The Chiefs voted unanimously (with two abstentions) on November 14 to approve the modified cut scores, presented in Table 13.

Table 13. Final Cut Scores Approved By Chiefs, With Impact Data.

Test	Level 2 Cut	% At or Above	Level 3 Cut	% At or Above	Level 4 Cut	% At or Above
Grade 3 English Language Arts/Literacy	2367	65%	2432	38%	2490	18%
Grade 3 Mathematics	2381	68%	2436	39%	2501	12%
Grade 4 English Language Arts/Literacy	2416	63%	2473	41%	2533	18%
Grade 4 Mathematics	2411	73%	2485	37%	2549	13%
Grade 5 English Language Arts/Literacy	2442	67%	2502	44%	2582	15%
Grade 5 Mathematics	2455	65%	2528	33%	2579	15%
Grade 6 English Language Arts/Literacy	2457	70%	2531	41%	2618	11%
Grade 6 Mathematics	2473	65%	2552	33%	2610	14%
Grade 7 English Language Arts/Literacy	2479	66%	2552	38%	2649	8%
Grade 7 Mathematics	2484	64%	2567	33%	2635	13%
Grade 8 English Language Arts/Literacy	2487	72%	2567	41%	2668	9%
Grade 8 Mathematics	2504	62%	2586	32%	2653	13%
Grade 11 English Language Arts/Literacy	2493	72%	2583	41%	2682	11%
Grade 11 Mathematics	2543	60%	2628	33%	2718	11%

Achievement Level Descriptors

Prior to the awarding of a contract for achievement level setting, Smarter Balanced had awarded several other contracts for program management, test development, and development of achievement level descriptors (ALDs). There are, or will be by spring 2015, four sets of ALDs (from Egan, Schneider, & Ferrara, 2012):

- Policy – brief statements that articulate policy makers’ vision of goals and rigor for the final performance standards;
- Range – guidelines created by test developers to identify which aspects of items align to a particular performance level with regard to the cognitive and content rigor that has been defined;
- Threshold (Target) – detailed statements created in conjunction with the Range ALDs and used by achievement level setting panelists to represent the knowledge and skills of a student just at the threshold of a given level;
- Reporting – relatively brief statements developed by a sponsoring agency once cut scores are finalized, to define the appropriate and intended interpretations of test scores.

Policy ALDs allowed Smarter Balanced to communicate to the educational community its intentions for development and implementation of rigorous assessments. The Range ALDs were used to guide item developers. Threshold ALDs were used to guide online and in-person achievement level setting panelists in the placement of bookmarks to recommend cut scores. In the spring of 2015, Smarter Balanced will use the reporting ALDs to describe the achievement of millions of students to parents, schools, districts, and states.

Once final cut scores are set, it is advisable to review ALDs to make sure that they are aligned to the cut scores. This section documents revisions to the Threshold ALDs in light of modifications to cut scores recommended by panelists, subsequent review of the Range ALDs, and development of the Reporting ALDs.

Threshold ALDs

Threshold ALDs were a central part of the training of online and in-person achievement level setting panelists. Through three rounds of achievement level setting, which included review of recommendations of Online Panelists, In-Person Workshop panelists justified each cut score on the basis of the content alignment of a test item on a given page of an ordered item booklet (OIB) with the description in the Threshold ALD. Subsequent changes to those cut scores by the Cross-Grade Review Committees (formerly known as the Vertical Articulation Committees) were also grounded in the Threshold ALDs. All cut score recommendations going forward to the Chiefs were thus firmly grounded in the language of the Threshold ALDs.

The final cut scores, however, were not always the same as those emerging from the Cross-Grade Review Committees. In some instances, they went up; in others, they went down. Thus, a

review of the alignment of the final cut scores and the Threshold ALDs was in order. The process, findings, and recommendations are detailed below.

Comparison of final cuts to recommended cuts.

Table 14 compares final cut scores to those recommended by the Cross-Grade Review Committees. In each instance, cut scores have been translated into page numbers in the OIBs, since these had been the focus of the initial recommendations. The plausible range of page numbers indicates the interquartile range of bookmark placements from Round 3 of the In-Person Workshop, as augmented by the Cross-Grade Review Committee. For example, if the middle 50 percent of the range of bookmarks for Level 3 placed by panelists in the In-Person Workshop for grade 4 Mathematics was 38 to 48, but the Cross-Grade Review Committee moved the bookmark to page 49, the plausible range was from pages 38 to 49. If the Cross-Grade Review Committee did not alter the cut score or moved it within the Round 3 range, the plausible range was whatever it had been at the end of Round 3. In 14, E and M refer to English language arts/literacy and Mathematics, and L2 – L4 refer to Levels 2 – 4. Cell entries are OIB page numbers. For the Plausible Range, medians and interquartile ranges can include page numbers that are not whole numbers; thus, they are reported in quarter-page increments.

Table 14. Comparison of Final Cuts to Those Recommended by the Cross-Grade Review Committees.

		OIB Page #			Plausible Range						In Range?			Out of Range by ___ Pages		
		L2	L3	L4	L2		L3		L4		L2	L3	L4	L2	L3	L4
Subject	Grade	L2	L3	L4	From	To	From	To	From	To						
E	3	14	36	57	11.75	15.50	28.00	38.00	53.00	61.75	Yes	Yes	Yes			
E	4	21	44	63	15.00	20.00	37.00	44.00	60.00	63.00	No	Yes	Yes	1		
E	5	21	38	65	10.00	18.00	27.00	37.00	61.00	65.00	No	No	Yes	3	1	
E	6	11	35	61	7.75	19.00	29.00	40.00	52.00	63.25	Yes	Yes	Yes			
E	7	16	39	65	8.00	16.00	34.50	43.50	64.00	74.00	Yes	Yes	Yes			
E	8	22	46	68	14.00	21.50	34.00	46.50	60.00	70.00	No	Yes	Yes	0.5		
E	11	19	46	65	15.25	23.00	40.00	45.00	63.00	66.00	Yes	No	Yes		1	
M	3	26	46	69	26.00	28.00	44.50	53.00	66.00	72.25	Yes	Yes	Yes			
M	4	17	46	72	12.00	18.00	38.00	49.00	71.00	73.00	Yes	Yes	Yes			
M	5	17	49	61	18.25	21.00	50.00	51.00	62.00	64.00	No	No	No	-1.25	-1	-1
M	6	15	40	63	13.00	20.00	32.50	53.50	59.00	63.00	Yes	Yes	Yes			
M	7	9	30	53	13.25	21.00	40.00	51.00	58.75	64.00	No	No	No	-4.25	-10	-5.75
M	8	8	33	50	15.00	18.00	36.50	48.00	57.50	63.00	No	No	No	-7	-3.5	-7.5
M	11	20	44	63	17.50	27.00	44.00	55.25	63.75	69.00	Yes	Yes	No			-0.75

As can be seen, most final cuts are either in range or very close to the plausible range. For those not within the Plausible Range, the final three columns indicate the distance from the edge of the range. Those out-of-range cuts were the primary focus of the ALD review.

ALD review.

MI staff drafted a plan, based on Table 14, and presented it to Smarter Balanced staff on November 24. Smarter Balanced approved the plan, and MI set it into motion. MI staff reviewed threshold ALDs, test blueprints, panelist and facilitator notes from the In-Person Workshop Panel and Cross-Grade Review Committees, and comments from the Online Panel for all cut scores in Table 14 that were out of range. They then shared their finding with Smarter Balanced staff, who reviewed them and provided feedback. MI staff then submitted final recommendations to Smarter Balanced staff for review and approval. Those findings and recommendations are detailed in the next subsection.

Findings and recommendations.

MI content specialists reviewed five modified cut scores for English language arts/literacy and ten for mathematics. In each instance, the content specialists were able to justify the new cut score in terms of the threshold ALDs. In several instances, the new cut score was only a page or two away from the plausible range established by the cross-grade review committee. However, even when the new cut score was as much as 10 pages below the range, the content specialists found that the content of the item associated with the new cut score met the threshold ALD criteria; i.e., that the item just below the bookmark presented a student at the threshold with about a 50% chance of answering correctly and that the item at the bookmark presented the student at the threshold with less than a 50% chance of answering correctly. Thus, in 15 out of 15 instances, the content specialists concluded that the final cut scores aligned to the threshold ALDs. An item-by-item account of the findings and recommendations is included in a separate report.

Range ALDs

As noted above, the purpose of Range ALDs is to guide test item developers. Specifically, item developers need to know what is to be expected of students at Levels 1, 2, 3, and 4. If those expectations change, item-development guidance also needs to change. As item development will continue into the foreseeable future, any change in expectations of students at various levels, as reflected in the Threshold ALDs, needs to be reflected in the Range ALDs. However, given that there were no changes to the Threshold ALDs, no changes are recommended for the Range ALDs.

Reporting ALDs

As noted above, reporting ALDs are relatively brief statements developed by a sponsoring agency once cut scores are finalized, to define the appropriate and intended interpretations of test scores. Ideally, they should reflect the specific knowledge, skills, and processes embodied in the tests. However, in the case of computer adaptive tests, those sets may vary from student to student;

therefore, the reporting ALDs for Smarter Balanced will need to be more generic, reflecting a range of knowledge, skills, and processes.

MI staff gathered requirements and recommendations from Smarter Balanced staff and others and drafted a matrix of policy ALDs and reporting ALDs for high school, grades 6-8, and grades 3-5. Staff of MI, CTB, and Hager Sharp met on December 1 to review the matrix and make revisions. This revised matrix was presented to Smarter Balanced leadership on December 2 for further review and revision. The results of that presentation were forwarded to Smarter Balanced for further revision. Draft reporting ALDs are included in a separate report.

Long Range Validity Agenda for Performance Level Cut Scores

As Smarter Balanced shifts from a developmental to an operational mode, additional research on cut scores is planned. The final task under Contract 21 is to prepare a long-range research agenda that will allow Smarter Balanced to test the validity of the cut scores against various external criteria.

In 2012, Smarter Balanced commissioned Stephen Sireci to prepare a comprehensive research “to demonstrate that the assessment system adheres to professional and Federal guidelines for fair and high quality assessment...to provide a comprehensive and detailed research agenda for the Consortium that includes suggestions and guidance for both short- and long-term research activities that will support Consortium goals” (Sireci, 2012, p. 5). The current report has a much more narrow focus: validation of cut scores established in the fall of 2014. However, the Sireci (2012) research agenda provides a solid foundation on which to build the plan for cut score validation.

The present plan is further guided by the 2014 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), just as the Sireci agenda was guided by the 1999 *Standards*, which are similar in many ways to the 2014 *Standards*. The present proposal is also guided by *Peer Review Guidance* (U. S. Department of Education, 2009), principles of Universal Design (Johnson, Altman, & Thurlow, 2006), Michael Kane’s recent essays on validation (Kane, 2001, 2006), and similar work by Susan Loomis (2011). In particular, this paper (as does the Sireci paper) uses the theoretical framework and terminology employed by Kane (2001, 2006) and reflected in the 2014 *Standards*; i.e., “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA, APA, NCME, 2014, p. 11). With specific reference to achievement level setting, Kane (2001, p. 54) notes, “To set a standard is to develop policy, and policy decisions are not right or wrong. They can be wise or unwise, effective or ineffective, but they cannot be validated by comparing them to some external criterion. The

argument for validity, or appropriateness, of a standard is necessarily extended, complex, and circumstantial.”

The validation studies described in this research plan focus principally on the summative assessments (i.e., those for which cut scores are to be set) and represent the perspective of the contractor for Smarter Balanced Contract 21 (Achievement Level Setting). They are based, in large measure, on previous large-scale cut-score validation efforts that involved collection and analysis of a wide range of data external to the assessment programs in question; i.e., the American Diploma Project (Miles, Beimers, & Way, 2010).

There will be opportunities following the 2015 operational test administration and beyond to examine the cut scores with respect to **internal** variables from the Smarter Balanced assessment system and targeted **external** variables. Appraising the cut scores from these various perspectives will yield important information as to the appropriateness of the interpretations and uses of these scores. The remainder of this section will therefore be devoted to a description of validation studies with internal variables and validation studies with external variables.

Validation Studies with Internal Variables

Internal variables are those specific to the tests themselves, their blueprints, their internal structures, and their score distributions. Specific variables and uses are outlined below.

- *Interim assessments*: States and districts will have administered the interim assessments and collected test score data based on the performance of their students. Because these assessments are on the same scale as the summative test, it will be possible to calculate the agreement level of students' achievement levels that emerge from the interim and summative scores.
- *Formative assessments*: Using samples of students matched on propensity scores, it will be possible to examine the performance of students on the interim and summative assessments based on their use of the formative assessments.
- *Digital Library*: Using samples of students matched on propensity scores, it will be possible to examine the performance of students on the interim and summative assessments based on their use of the Digital Library materials.
- *Expert studies*: Blueprints for the 2015 assessments were approved at the Collaborative Conference on April 30. Given the computer-adaptive nature of the 2015 assessments, there will not be a static form whose alignment to a blueprint can be readily evaluated. However, it will be possible to assemble proto-tests (i.e., collections of items for students at specific hypothetical ability levels) that will closely resemble tests administered to actual students. Those proto-tests should be reviewed by content experts. Specifically, higher education faculty should evaluate the high school tests, high school faculty will evaluate middle school tests, and middle school faculty should evaluate elementary school tests for alignment to the blueprints as well as depth of knowledge and rigor.
- *Level on subsequent tests*: Just as one would expect some degree of consistency of impact across grades, one would also expect individual students to perform consistently from one grade to the next. Specifically, most students would be expected to progress from their initial level to higher levels over time. Failure of large numbers of students to progress as expected

would call into question the appropriateness of the cut scores, whether instruction and motivation were adequate, or whether some combination of factors caused the deviation from expectation. A longitudinal study in which a sample of students in grades 3-8, over a three-year period (until a majority of the first year's eighth graders take the high school tests) is recommended. Students should have their scores and levels tracked from grade to grade. Given that implementation of the Common Core is expected to strengthen year by year, percentages of students in this cohort scoring at or above Level 3 should either increase each year or at least hold true to the expected percentages established through vertical articulation in October 2014. Thus, the focus should be on deviations from these two possible patterns.

- *Consistency of cut scores across grades:* The results of the cross-grade review committee and subsequent actions by the Governing States provided a progression of cut scores rather than a single cut score. The reasonableness of the final distribution of cut scores and percentages of students at or above any given level can be compared to teachers' evaluations of their own students in the spring of 2015. This study will be carried out in conjunction with the teacher rating study in the spring of 2015.
- *Differential outcomes:* The Smarter Balanced assessment design was based on the principles of universal design. Therefore, one would expect equal access to all assessments and minimal differential item functioning (DIF). The DIF analyses proposed by Sireci (2012) for test development have been carried out. Similar analyses should be conducted in the spring of 2015.

Validation Studies with External Variables

External variables are those outside of the tests themselves, in terms of how they relate to performance on the tests. These include teacher ratings, student grades, scores on other tests, employer ratings, and other variables. Specific examples are outlined below.

- *Teacher ratings in 2015:* Teachers in selected schools will provide ratings of samples of students, using the threshold ALDs. It will then be possible to cross-tabulate those ratings with Smarter Balanced test scores and level designations.
- *Teacher ratings in subsequent years:* If a Grade 5 student is deemed ready to move on to Grade 6 and perform adequately, the Grade 6 teacher should find that student ready. To the extent that faculty in subsequent years find students not to be as prepared as the level designations they received the previous year, either the cut scores were invalid, or there was a mismatch between prior-year scholastic preparation and subsequent-year requirements; i.e., misalignment of curriculum and instruction. Starting in the fall of 2015, selected teachers in Smarter Balanced states will be asked to categorize their students, using the Smarter Balanced ALDs (for the previous grade). The level designations they assign to their incoming students will be compared to the level designations those students earned on the previous spring's tests. This study will be repeated in the fall of 2016 and fall of 2017 with different schools, teachers, and students.
- *Student grades in 2015:* Other selected schools will provide class grades or course grades of samples of students who also take the Smarter Balanced tests. It will then be possible to cross-tabulate those course grades with Smarter Balanced test level designations.
- *Course grades in subsequent years:* A parallel study will focus on course grades in 2016 and 2017. Students rated at the top level in Grade 6 should outperform students rated at lower levels when they are evaluated on Grade 7 work. College- and career-ready high school students should perform better in college algebra and freshman English than those who are

not considered college and career ready. Starting in the 2015-16 school year, selected schools in Smarter Balanced states will be asked to supply course grades for their students. These grades will be compared to the level designations those students earned on the previous spring's tests. This study will be repeated in the fall of the 2016-17 and 2018-19 school years, by which time two cohorts of high school students will have entered postsecondary education and can supply college course grades.

- *NAEP scores:* Many of the schools testing in the spring of 2015 will also administer the National Assessment of Educational Progress (NAEP). For students taking those tests, level designations and/or scale scores may be available. In the event that they are, those scale scores and level designations can be compared to Smarter Balanced scale scores and level designations. This is essentially the approach taken by Gary Phillips (2012) in comparing NAEP scores to state achievement test scores.
- *Scores on other tests:* Many states, districts, and individual schools will continue to administer other standardized assessments, either commercial off-the-shelf tests or additional state-sponsored tests. Samples of students who take both a Smarter Balanced assessment and an additional standardized test in 2014-15 will provide the data for these studies. It will be necessary to obtain not only scale scores but also percentile ranks, proficiency levels, or other derived scores and scales for the external tests.
- *Scores of college students on the Grade 11 tests:* Although higher education faculty will be involved in the setting of achievement levels for the high school tests, there is no substitute for administration of the Grade 11 tests to college freshmen during the 2014-15 school years and cross-tabulation of level designation with their course grades. Similar studies have been conducted as part of the American Diploma Project (Miles, Beimers, & Way, 2010). Samples of college students, drawn from a variety of institution types, should take Smarter Balanced high school tests and also report their course grades in freshman English or mathematics. Similarly, samples of high school freshmen should take the grade 8 tests, and samples of grade 6 students should take the grade 5 tests.
- *Differential prediction:* Prediction of future outcomes is but a first step; comparing predictability across subgroups is the second step. In particular, it is advisable to compare the predictive power of Smarter Balanced assessments for students in general with their predictive power for specific target groups.
- *Opportunity to learn:* Curricular and instructional validation must be considered, especially over time. The first opportunity-to-learn (OTL) survey should be conducted in the spring of 2015 concurrent with an operational administration of Smarter Balanced tests. No matter how well the tests are constructed, no matter how well they are aligned with the Common Core, and no matter how carefully cut scores are derived, if large numbers of students have not had the opportunity to learn the content of the tests, no cut score will be meaningful. Students in states adopting and implementing the Common Core early would be expected to perform better on Smarter Balanced tests than students in later adopting and implementing states. These studies should provide concrete evidence to support or dispel those expectations.
- *Employer evaluations:* The study of course grades in subsequent years will cover the college half of "college and career ready." Surveys of employers should cover the career half. During the 2014-15 school year, businesses and industries that hire large numbers of students right out of high school should be identified. Representatives of those businesses should identify minimum academic skill requirements of entry-level employees. By the 2016-17 school year, many of the high school students who took Smarter Balanced tests in 2014-15 will have entered the work force. In the fall of 2016 and again in the fall of 2017, selected employers of those students who have entered the work force in those years (having taken the grade 11 tests the previous year) should receive survey forms to complete regarding the readiness of

those young people for the jobs they have taken. Responses from those employers would be matched with the Smarter Balanced scores and level designations of their employees.

Organization and Implementation of Studies

Table 15 shows a proposed organization and implementation timeline for the various studies outlined above. It is intentionally general in nature, prescribing neither sample size nor specific analytic tools or procedures.

Table 15. Validation Study Implementation Timeline.

School Year	Internal Validation Studies	External Validation Studies
2014-15	Interim assessments – collect data from selected schools and districts showing the relationships between interim assessments and subsequent summative assessments, particularly the relationship between the summative assessments and those interim assessments taken shortly before them.	Teacher ratings – collect teacher ratings from selected schools and districts, and compare their evaluations of student levels to those obtained on Smarter Balanced tests.
	Formative assessments – collect data from users and non-users of formative assessments to compare summative performance.	Student grades – collect student grades from selected schools and districts, and compare those grades to student scale scores and levels on Smarter Balanced tests.
	Digital library – collect data from users and non-users of the digital library to compare summative performance.	NAEP scores – identify students who will participate in the National Assessment and arrange to match their NAEP scores to their Smarter Balanced scores to solidify the link between the two scales.
	Expert studies – recruit higher education faculty and teachers from each grade to evaluate proto-tests or take a computer adaptive version of a specific test and provide feedback to Smarter Balanced with regard to alignment with the Common Core, rigor, and difficulty.	Scores on other tests – identify schools and districts administering at least two commercially available tests and at least two state- or locally-administered tests; match student scores, and report correlations, scale equations, and differential scores on other tests by Smarter Balanced level.
	Consistency of cut scores – examine percentages of students at each grade level to determine whether the pattern of percentages in 2014 holds up.	Scores of college students on grade 11 tests – identify 4-year and 2-year colleges to participate; administer Smarter Balanced grade 11 ELA tests to selected students, and report distributions of scale scores and levels.

	Differential outcomes – perform DIF analyses on all items by race, gender, and program.	Opportunity to learn – identify a sample of districts and schools in all states administering Smarter Balanced states, and administer an OTL survey to teachers and administrators. Compare test results to OTL rates.
2015-16	Interim assessments – repeat 2014-15 study. Report cumulative results as well as trends.	Teacher ratings – collect teacher ratings from receiving teachers and compare to level designations from 2015 Smarter Balanced assessments.
	Formative assessments – repeat 2014-15 study. Report cumulative results as well as trends.	Student grades – compare course grades of students in each grade to their previous year’s Smarter Balanced level designation. Tabulate average grades by Smarter Balanced level.
	Digital library – repeat 2014-15 study. Report cumulative results as well as trends.	Differential prediction – carry out studies of teacher ratings and student grades by subgroup (race, gender other reporting categories).
	Level on subsequent tests – for selected districts and schools, merge 2015 and 2016 assessment data and compare level designations. Report percentages of students moving up, down, or remaining in the same level.	Opportunity to learn -- repeat 2014-15 study. Report cumulative results as well as trends.
	Consistency of cut scores across grades – repeat 2014-15 study. Report cumulative results as well as trends.	Employer evaluations – for students tested in grade 11 in 2015, collect employer evaluations of new 2016 graduates by October 2016.
	Differential outcomes - repeat 2014-15 study. Report cumulative results as well as trends.	
	Differential outcomes- repeat 2014-15 study. Report cumulative results as well as trends.	
2016-17	Interim assessments – repeat 2015-16 study. Report cumulative results as well as trends.	Teacher ratings – repeat 2015-16 study. Report cumulative results as well as trends.
	Formative assessments – repeat 2015-16 study. Report cumulative results as well as trends.	Student grades – repeat 2015-16 study. Report cumulative results as well as trends.

	Digital library – repeat 2015-16 study. Report cumulative results as well as trends.	Differential prediction – repeat 2015-16 study. Report cumulative results as well as trends.
	Level on subsequent tests – for selected districts and schools, merge 2015 and 2016 assessment data and compare level designations. Report percentages of students moving up, down, or remaining in the same level.	Opportunity to learn – repeat 2015-16 study. Report cumulative results as well as trends.
	Consistency of cut scores across grades – repeat 2015-16 study. Report cumulative results as well as trends.	Employer evaluations – for students tested in grade 11 in 2016, collect employer evaluations of new 2017 graduates by October 2017.
	Differential outcomes - repeat 2015-16 study. Report cumulative results as well as trends.	
	Differential outcomes- repeat 2015-16 study. Report cumulative results as well as trends.	
	Convene K-12 educators, higher education faculty and administrators, general public, and other key stakeholders to review cut scores set in 2014 in light of validation data collected to date. Recommend changes if necessary.	

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Author.
- Cizek, G. J. & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.
- Egan, K. A., Schneider, C., & Ferrara, S. (2012). Performance level descriptors. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd Ed.), New York: Routledge.
- International Test Commission (2010). *Guidelines for translating and adapting tests*. Downloaded from the world wide web at <http://www.intestcom.org> on September 1, 2014.
- Johnstone, C. J.; Altman, J.; & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Available from <http://www.cehd.umn.edu/nceo/OnlinePubs/StateGuideUD/default.htm>
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational Measurement* (4th ed., pp. 17-64). Washington, DC: American Council on Education/Praeger.
- Lewis, D. M., Mitzel, H. C., Mercado, R. L, & Schultz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations*. New York: Routledge.
- Lewis, D. M., & Mitzel, H. C. (1995). An item response theory based standard setting procedure. In D. C. Green (Chair), *Some uses of item response theory in standard setting setting*. Symposium conducted at the annual meeting of the California Educational Research Association, Lake Tahoe, NV.
- Lewis D.M., Mitzel, H. C., Green, D. R. (1996). *Standard Setting: A Bookmark Approach*. In D. R. Green (Chair), *IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring*. Symposium presented at the 1996 Council of Chief State School Officers 1996 National Conference on Large Scale Assessment, Phoenix, AZ.
- Loomis, S. C. (2011). *Toward a validity framework for reporting preparedness of 12th graders for college-level course placement and entry to job training programs*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
- Miles, J. A., Beimers, J. N., & Way, W. D. (2010). *The Modified Briefing Book Standard Setting Method: Using Validity Data as a Basis for Setting Cut Scores*. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Denver, CO.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum.

- Phillips, G. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations*. New York: Routledge.
- Sireci, S. G. (2012). *Smarter Balanced Assessment Consortium: Comprehensive Research Agenda Report Prepared for the Smarter Balanced Assessment Consortium*.
- Smarter Balanced Assessment Consortium (2010). *Theory of Action: An Excerpt from the Smarter Balanced Race to the Top Application*. Tacoma, WA: Author.
- U.S. Department of Education (2009, January). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. [Revised December 21, 2007 to include Modified academic achievement standards. Revised with technical edits, January 12, 2009] Washington, DC: Author.

Change Log

Section	Page	Description of Changes	Version Date
Chapter 8	352	“Test reliability” replaced with “performance task reliability.”	01/14/16
Chapter 8, Table 19	353	Removed the term “Test” from the table caption.	01/14/16

Appendices

[Appendices are available upon request.]