

Chapter 6 Pilot Test and Special Studies (Dimensionality Analysis and IRT Model Choice) 7

Pilot Data Collection Design 8

 Table 1. Total Number of CAT Components and Performance Tasks (PT). 9

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

 Table 3. ELA/literacy Grade 11 Pilot Test CAT Component Blueprint..... 10

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

Vertical Linking Item Assignment 11

 Figure 1. Summary of Vertical Articulation of Test Content by Grade..... 12

Pilot Test Sampling Procedures 12

 Sampling Consideration for the Pilot Test. 12

 Test Administration and Sample Size Requirements 13

 Table 5. Targeted Student Sample Size by Content Area and Grade for the Pilot Test. 14

 Pilot Sampling Considerations 14

 Sampling Procedures..... 17

 Table 6. Approximate Sample Sizes by Content Area and State, the Sample Target and the Number Obtained for the Pilot Test..... 19

 Table 7. ELA/literacy Student Population and Sample Characteristics (Percentages)..... 20

 Table 8. Mathematics Student Population and Sample Characteristics (Percentages) 21

Pilot Classical Test Results..... 22

 Pilot Classical Item Flagging Criteria 23

 Description of Pilot Classical Statistics Evaluated 24

 Pilot Results 26

 Table 9. Summary of Number of Pilot Test Items and Students Obtained..... 26

 Table 10. Overview of ELA/literacy CAT Component Statistics. 27

 Table 11. Overview of Mathematics Component Statistics. 27

 Table 12. Student Testing Durations in Days (Percentage Completion). 28

 Table 13. Summary of Reliability and Difficulty for CAT Administrations. 29

 Table 14. Description of Item Flagging for Selected-response Items..... 30

 Table 15. Description of Item Flagging for Constructed-response Items. 31

 Table 16. Number of Items Flagged for ELA/literacy by Selected- and Constructed-response..... 32

 Table 17. Number of Items Flagged for Mathematics by Selected- and Constructed-response..... 32

 Table 18. Definition of Focal and Reference Groups. 33

Table 19. DIF Categories for Selected-Response Items.....34

Table 20. DIF Categories for Constructed-Response Items.34

Table 21. Number of DIF Items Flagged by Item Type and Subgroup (ELA/literacy, Grades 3 to 7).35

Table 22. Number of DIF Items Flagged by Item Type and Subgroup (ELA/literacy, Grades 8 to 11).36

Table 23. Number of C DIF Items Flagged by Item Type and Subgroup (Mathematics, Grades 3 to 7).37

Table 24. Number of C DIF Items Flagged by Item Type and Subgroup (Mathematics, Grades 8 to 11).38

Dimensionality Study39

 Rationale and Approach39

 Factor Models39

 Figure 2. An Example of the Bifactor Model with Four Minor Factors Corresponding to Claims.40

 Table 25. Summary of MIRT Analysis Configuration Showing Number of Content, Grades and MIRT Models.41

 MIRT Scaling Models41

 Software and System Requirements42

 Evaluation of the Number and Types of Dimensions and MIRT Item Statistics.....42

 Table 26. Models and Fit Measures for ELA/literacy Within Grade.44

 Table 27. Models and Fit Measures for Mathematics Within Grade.46

 Table 28. Models and Fit Measures for ELA/literacy Across Adjacent Grades.48

 Table 29. Models and Fit Measures for Mathematics Across Adjacent Grades.52

 MIRT Item Statistics and Graphs56

 Discussion and Conclusion57

 Figure 3. Item Vector Plot for ELA/literacy Grade 3 (Within Grade).....59

 Figure 4. Item Vector Plot for ELA/literacy Grade 4 (Within Grade).....59

 Figure 5. Item Vector Plot for ELA/literacy Grade 5 (Within Grade).....60

 Figure 6. Item Vector Plot for ELA/literacy Grade 6 (Within Grade).....60

 Figure 7. Item Vector Plot for ELA/literacy Grade 7 (Within Grade).....61

 Figure 8. Item Vector Plot for ELA/literacy Grade 8 (Within Grade).....61

 Figure 9. Item Vector Plot for ELA/literacy Grade 9 (Within Grade).....62

 Figure 10. Item Vector Plot for ELA/literacy Grade 10 (Within Grade)62

 Figure 11. Item Vector Plot for ELA/literacy Grade 11 (Within Grade)63

Figure 12. Item Vector Plot for ELA/literacy Grades 3 and 4 (Across Grades).....	63
Figure 13. Item Vector Plot for ELA/literacy Grades 4 and 5 (Across Grades).....	64
Figure 14. Item Vector Plot for ELA/literacy Grades 5 and 6 (Across Grades).....	64
Figure 15. Item Vector Plot for ELA/literacy Grades 6 and 7 (Across Grades).....	65
Figure 16. Item Vector Plot for ELA/literacy Grades 7 and 8 (Across Grades).....	65
Figure 17. Item Vector Plot for ELA/literacy Grades 8 and 9 (Across Grades).....	66
Figure 18. Item Vector Plot for ELA/literacy Grades 9 and 10 (Across Grades).....	66
Figure 19. Item Vector Plot for ELA/literacy Grades 10 and 11 (Across Grades)	67
Figure 20. Item Vector Plots for the Subset of ELA/literacy Grades 3 and 4 Vertical Linking Items.....	67
Figure 21. Item Vector Plots for the Subset of ELA/literacy Grades 4 and 5 Vertical Linking Items.....	68
Figure 22. Item Vector Plots for the Subset of ELA/literacy Grades 5 and 6 Vertical Linking Items.....	68
Figure 23. Item Vector Plots for the Subset of ELA/literacy Grades 6 and 7 Vertical Linking Items.....	69
Figure 24. Item Vector Plots for the Subset of ELA/literacy Grades 7 and 8 Vertical Linking Items.....	69
Figure 25. Item Vector Plots for the Subset of ELA/literacy Grades 8 and 9 Vertical Linking Items.....	70
Figure 26. Item Vector Plots for the Subset of ELA/literacy Grades 9 and 10 Vertical Linking Items.....	70
Figure 27. Item Vector Plots for the Subset of ELA/literacy Grades 10 and 11 Vertical Linking Items.....	71
Figure 28. Item Vector Plot for Mathematics Grade 3 (Within Grade).....	71
Figure 29. Item Vector Plot for Mathematics Grade 4 (Within Grade).....	72
Figure 30. Item Vector Plot for Mathematics Grade 5 (Within Grade).....	72
Figure 31. Item Vector Plot for Mathematics Grade 6 (Within Grade).....	73
Figure 32. Item Vector Plot for Mathematics Grade 7 (Within Grade).....	73
Figure 33. Item Vector Plot for Mathematics Grade 8 (Within Grade).....	74
Figure 34. Item Vector Plot for Mathematics Grade 9 (Within Grade).....	74
Figure 35. Item Vector Plot for Mathematics Grade 10 (Within Grade)	75
Figure 36. Item Vector Plot for Mathematics Grade 11 (Within Grade)	75
Figure 37. Item Vector Plot for Mathematics Grades 3 and 4 (Across Grades)	76
Figure 38. Item Vector Plot for Mathematics Grades 4 and 5 (Across Grades).....	76

Figure 39. Item Vector Plot for Mathematics Grades 5 and 6 (Across Grades) 77

Figure 40. Item Vector Plot for Mathematics Grades 6 and 7 (Across Grades) 77

Figure 41. Item Vector Plot for Mathematics Grades 7 and 8 (Across Grades) 78

Figure 42. Item Vector Plot for Mathematics Grades 8 and 9 (Across Grades) 78

Figure 43. Item Vector Plot for Mathematics Grades 9 and 10 (Across Grades)..... 79

Figure 44. Item Vector Plot for Mathematics Grades 10 and 11 (Across Grades) 79

Figure 45. Item Vector Plot for the Subset of Mathematics Grades 3 and 4 (Vertical Linking Items)..... 80

Figure 46. Item Vector Plot for the Subset of Mathematics Grades 4 and 5 (Vertical Linking Items)..... 80

Figure 47. Item Vector Plot for the Subset of Mathematics Grades 5 and 6 (Vertical Linking Items)..... 81

Figure 48. Item Vector Plot for the Subset of Mathematics Grades 6 and 7 (Vertical Linking Items)..... 81

Figure 49. Item Vector Plot for the Subset of Mathematics Grades 7 and 8 (Vertical Linking Items)..... 82

Figure 50. Item Vector Plot for the Subset of Mathematics Grades 8 and 9 (Vertical Linking Items)..... 82

Figure 51. Item Vector Plot for the Subset of Mathematics Grades 9 and 10 (Vertical Linking Items)..... 83

Figure 52. Item Vector Plot for the Subset of Mathematics Grades 10 and 11 (Vertical Linking Items)..... 83

Figure 53. Clustering of Item Angle Measures for Grades 3 to 5, ELA/literacy (within grade)..... 84

Figure 54. Clustering of Item Angle Measures for Grades 6 to 8, ELA/literacy (within grade)..... 85

Figure 55. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (within grade)..... 86

Figure 56. Clustering of Item Angle Measures for Grades 3 to 6, ELA/literacy (across grades)..... 87

Figure 57. Clustering of Item Angle Measures for Grades 6 to 9, ELA/literacy (across grades)..... 88

Figure 58. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (across grades)..... 89

Figure 59. Clustering of Item Angle Measures for Grades 3 to 6, ELA/literacy (vertical linking)..... 90

Figure 60. Clustering of Item Angle Measures for Grades 6 to 9, ELA/literacy (vertical linking) 91

Figure 61. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (vertical linking) 92

Figure 62. Clustering of Item Angle Measures for Grades 3 to 5, Mathematics (within grade)..... 93

Figure 63. Clustering of Item Angle Measures for Grades 6 to 8, Mathematics (within grade)..... 94

Figure 64. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (within grade)..... 95

Figure 65. Clustering of Item Angle Measures for Grades 3 to 6, Mathematics (across grades)..... 96

Figure 66. Clustering of Item Angle Measures for Grades 6 to 9, Mathematics (across grades)..... 97

Figure 67. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (across grades)..... 98

Figure 68. Clustering of Item Angle Measures for Grades 3 to 6, Mathematics (vertical linking) 99

Figure 69. Clustering of Item Angle Measures for Grades 6 to 9, Mathematics (vertical linking) 100

Figure 70. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (vertical linking) 101

Item Response Theory (IRT) Model Comparison 102

 IRT Data Step 104

 Table 30. Number of Items Dropped from the Calibration (On-grade). 106

 Table 31. Number of Constructed-response and Performance tasks with Collapsed Score Levels (On-grade). 106

 Table 32. Number of Constructed-response and Performance tasks with Collapsed Score Levels for ELA/literacy (Detail). 107

 Table 33. Number of Constructed-response with Collapsed Score Levels for Mathematics (Detail). 108

 Table 34. Number of ELA/literacy and Mathematics Items in the IRT Calibration. 109

 Table 35. Descriptive Statistics for Number of Students per Item for ELA/literacy and Mathematics. 110

 IRT Model Calibration 110

 IRT Model Fit Comparison 112

Table 36. Summary of G^2 Statistics of On-Grade ELA/literacy Items across 1PL, 2PL, and 3PL IRT Models.	113
Table 37. Summary of G^2 Statistics of On-Grade Mathematics Items across 1PL, 2PL, and 3PL IRT Models.	113
Guessing Evaluation	114
Table 38. Summary of Guessing Parameter Estimates for On-Grade ELA/literacy Items.	114
Table 39. Summary of Guessing Parameter Estimates for On-Grade Mathematics Items.....	115
Common Discrimination Evaluation.....	116
Table 40. Summary of 2PL/GPC Slope and Difficulty Estimates and Correlations for ELA/literacy.	117
Table 41. Summary of 2PL/GPC Slope and Difficulty Estimates and Correlations for Mathematics.	122
Evaluation of Ability Estimates	126
Table 42. ELA/literacy Correlations of Ability Estimates across Different Model Combinations.	126
Table 43. Mathematics Correlations of Ability Estimates across Different Model Combinations.	128
IRT Model Recommendations	130
Figure 71. Scatter Plot of ELA/literacy 2PL/GPC Slope and Difficulty Estimates by Item Type, Score Category and Claim.....	131
Figure 72. Scatter Plot of Mathematics 2PL/GPC Slope and Difficulty Estimates by Item Type, Score Category, and Claim.....	136
Figure 73. ELA/literacy Scatter Plots of Theta Estimates across Different Model Combinations	141
Figure 74. Mathematics Scatter Plots of Theta Estimates Across Different Model Combinations	146
References	151

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.

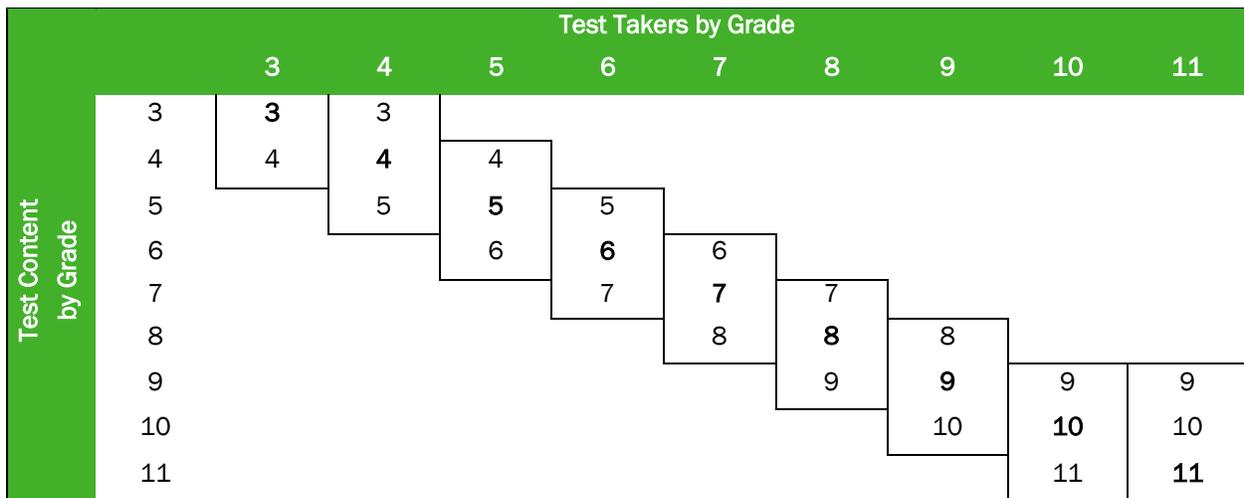


Figure 1. Summary of Vertical Articulation of Test Content by Grade.

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

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 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.

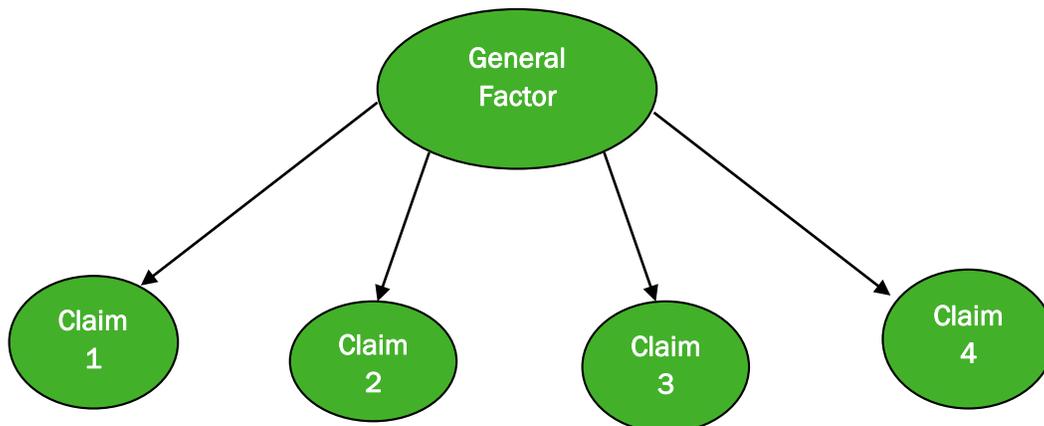


Figure 2. An Example of the Bifactor Model with Four Minor Factors Corresponding to Claims.

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

$$P_{ij} = 1 - \frac{1}{1 + e^{\tilde{\beta}_{2j} \square \tilde{\theta}_i + \beta_{\delta j}}} = \frac{1}{1 + e^{-\tilde{\beta}_{2j} \square \tilde{\theta}_i - \beta_{\delta j}}}$$

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):

$$P_{ijk} = P(X_{ij} = k-1 | \vec{\theta}_i, \vec{\beta}_{2j}) = \frac{e^{(k-1)\vec{\beta}_{2j} \square \vec{\theta}_i - \sum_{t=1}^k \beta_{\delta jt}}}{\sum_{m=1}^{K_j} e^{(m-1)\vec{\beta}_{2j} \square \vec{\theta}_i - \sum_{t=1}^m \beta_{\delta jt}}},$$

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., theta). 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		Overall	2,106,595	3,330,799	-918,914	134,384	
		8	1,346,541	2,059,674	-590,583	82,687	
		9	760,054	1,181,582	-328,330	51,697	
Claim Scores		Overall	2,355,982	4,796,591	-910,079	267,912	
		8	1,454,408	2,876,536	-562,310	164,894	
		9	901,573	1,741,563	-347,769	103,018	
Bifactor		Overall	2,592,106	5,640,955	-961,373	334,680	
		8	1,564,631	3,341,268	-576,317	205,999	
		9	1,027,475	2,076,717	-385,057	128,681	
9 to 10	Unidimensional	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	2,946,770	6,944,012	-1,046,057	427,328
			7	1,424,973	3,179,747	-508,754	203,732
Exploratory		Overall	1,702,770	2,288,505	-786,775	64,610	
		8	1,184,158	1,576,296	-547,039	45,040	
		9	518,613	672,584	-239,736	19,570	
Claim Scores		Overall	1,785,655	2,951,024	-764,280	128,547	
		8	1,245,487	2,026,316	-533,059	89,684	
		9	540,168	845,931	-231,221	38,863	
Bifactor		Overall	2,027,808	4,352,371	-757,491	256,413	
		8	1,401,321	2,959,559	-521,686	178,975	
		9	626,486	1,235,746	-235,805	77,438	
	Bifactor	Overall	2,160,865	5,065,062	-760,082	320,350	
		8	1,504,595	3,451,549	-528,675	223,622	
		9	656,270	1,417,297	-231,407	96,728	
9 to 10	Unidimensional	Overall	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.

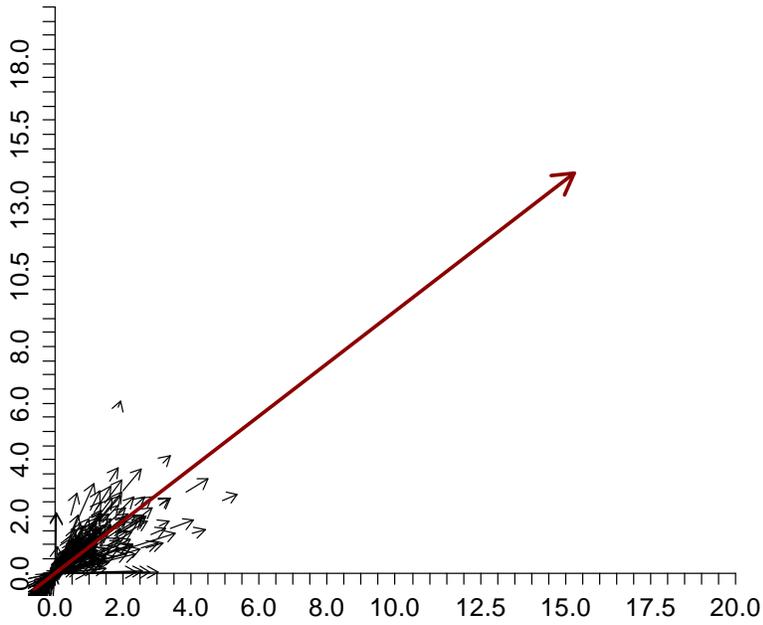


Figure 3. Item Vector Plot for ELA/literacy Grade 3 (Within Grade)

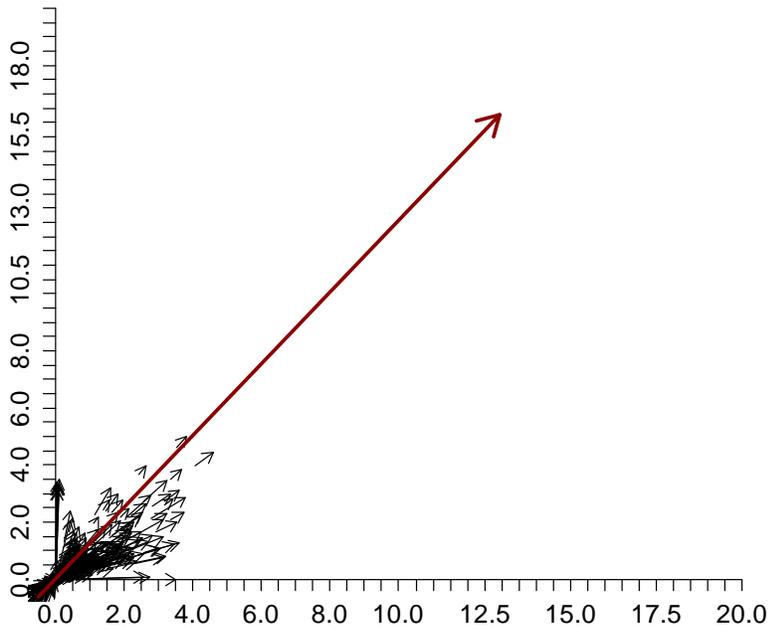


Figure 4. Item Vector Plot for ELA/literacy Grade 4 (Within Grade)

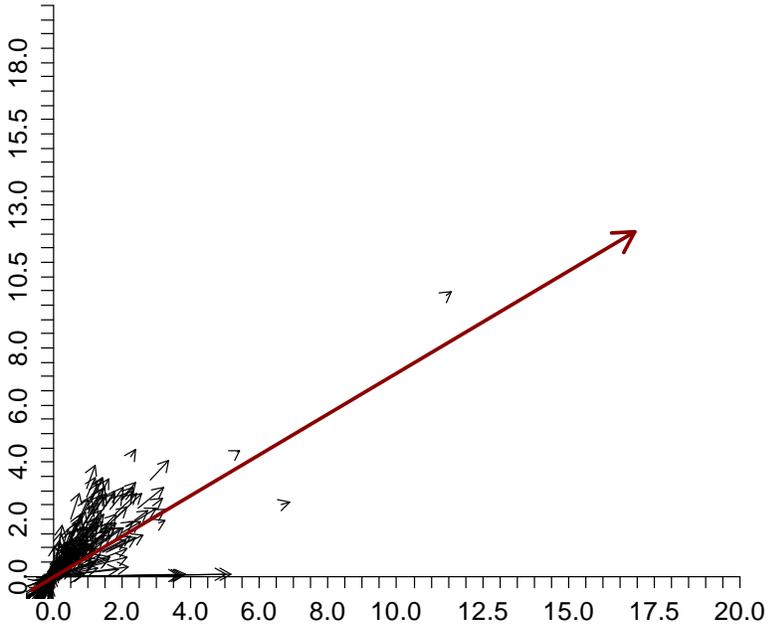


Figure 5. Item Vector Plot for ELA/literacy Grade 5 (Within Grade)

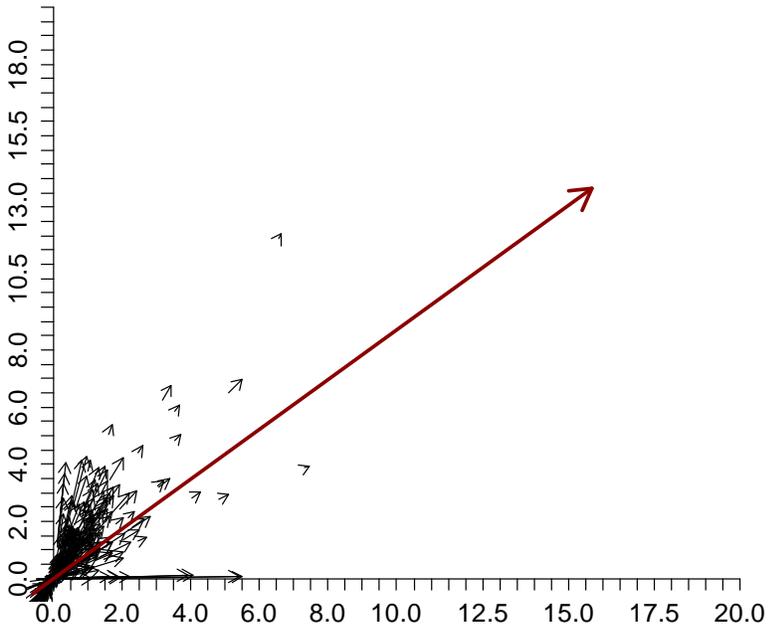


Figure 6. Item Vector Plot for ELA/literacy Grade 6 (Within Grade)

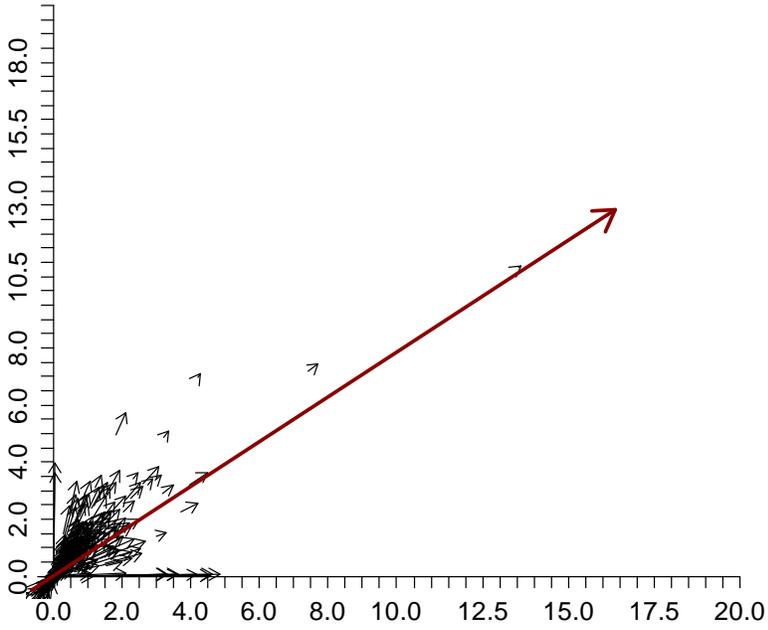


Figure 7. Item Vector Plot for ELA/literacy Grade 7 (Within Grade)

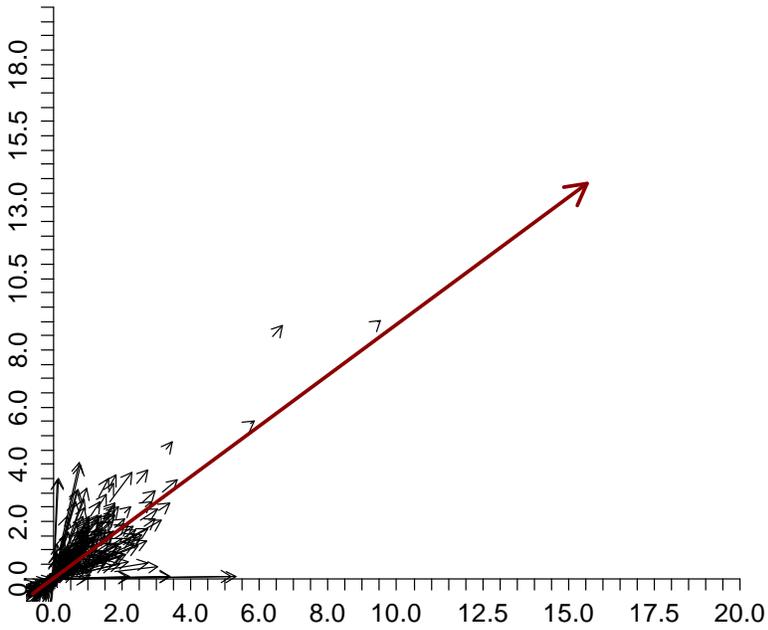


Figure 8. Item Vector Plot for ELA/literacy Grade 8 (Within Grade)

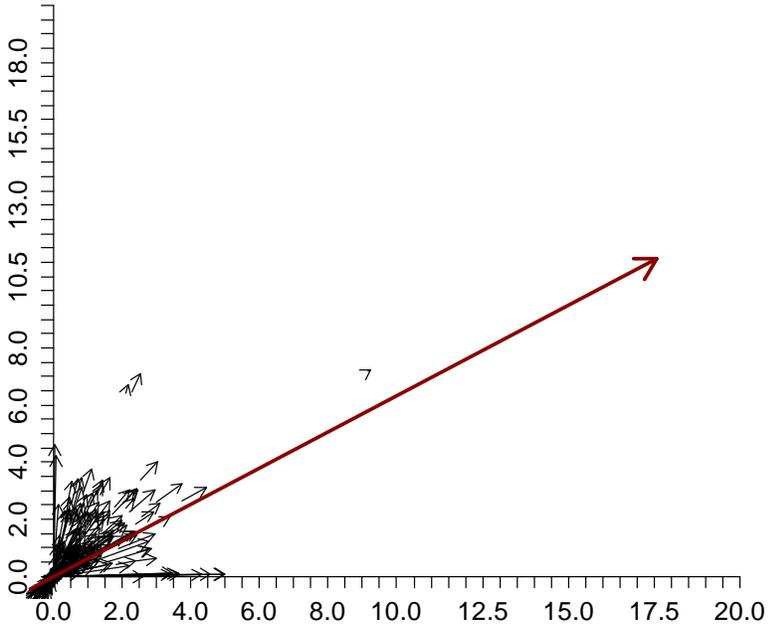


Figure 9. Item Vector Plot for ELA/literacy Grade 9 (Within Grade)

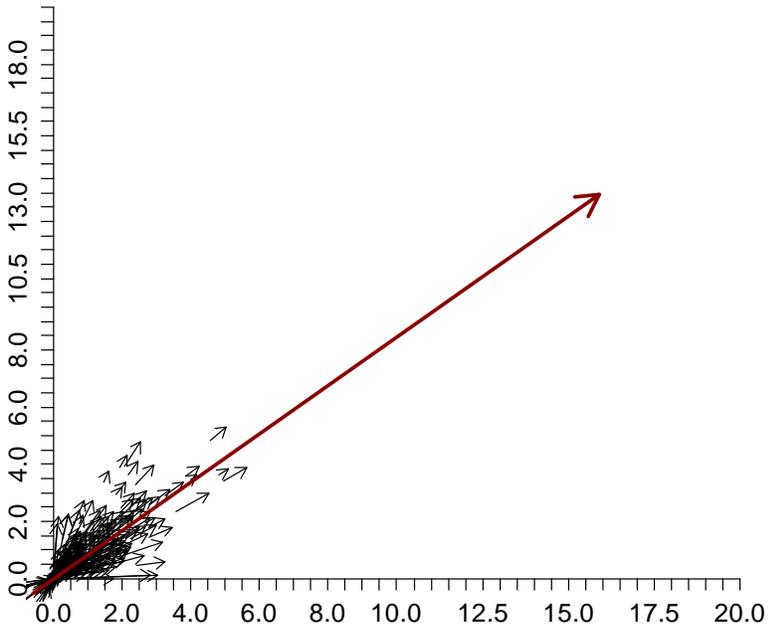


Figure 10. Item Vector Plot for ELA/literacy Grade 10 (Within Grade)

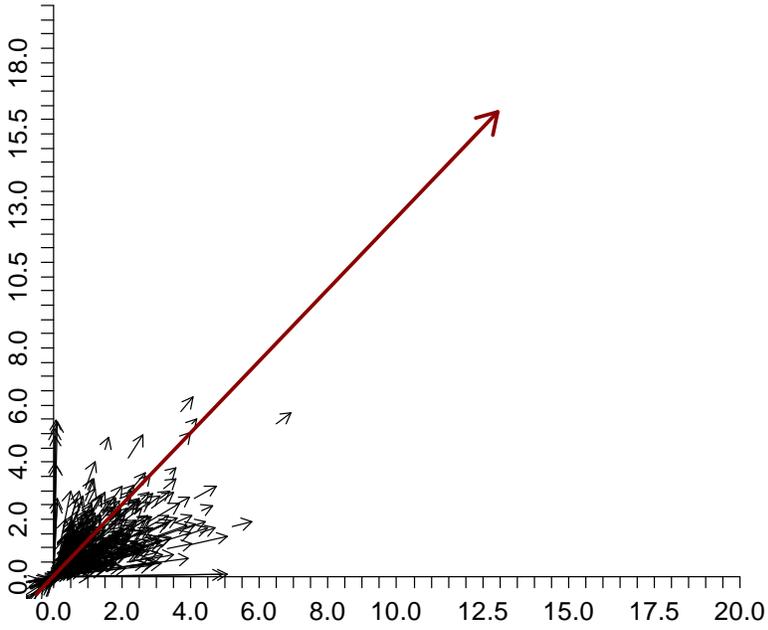


Figure 11. Item Vector Plot for ELA/literacy Grade 11 (Within Grade)

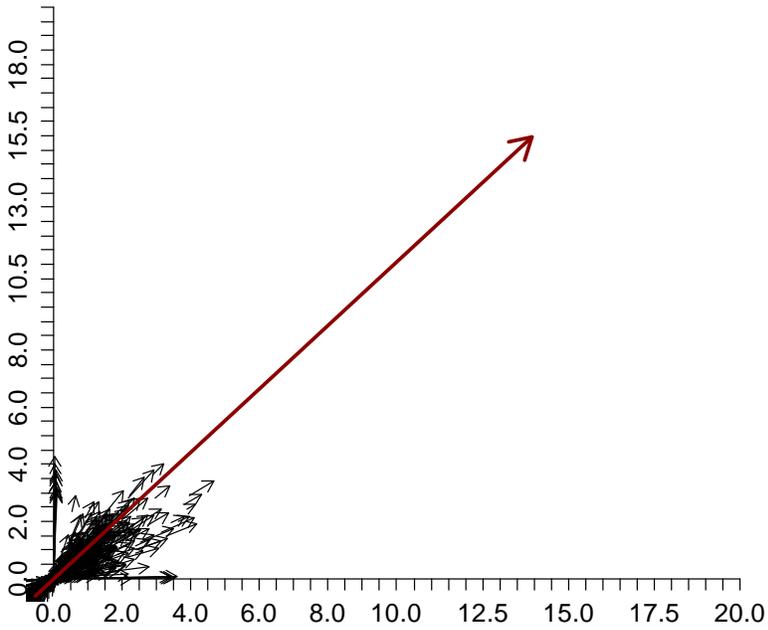


Figure 12. Item Vector Plot for ELA/literacy Grades 3 and 4 (Across Grades)

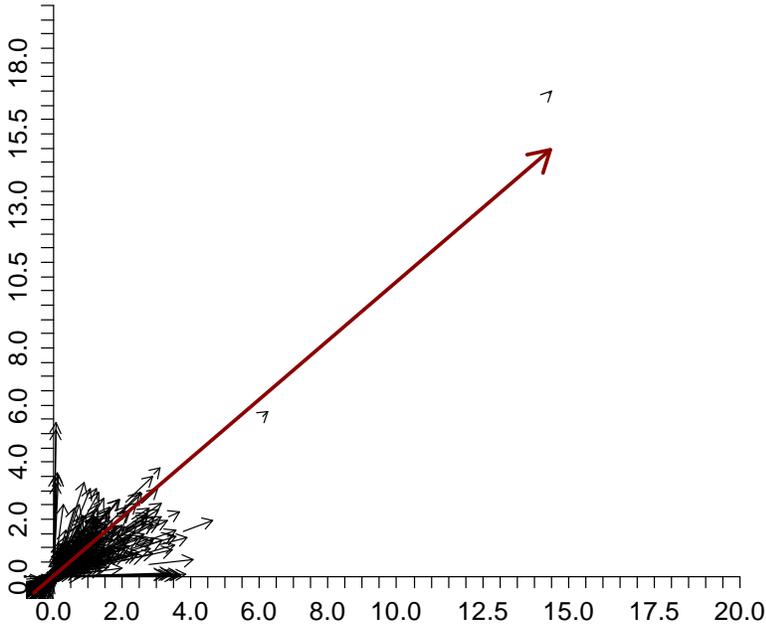


Figure 13. Item Vector Plot for ELA/literacy Grades 4 and 5 (Across Grades)

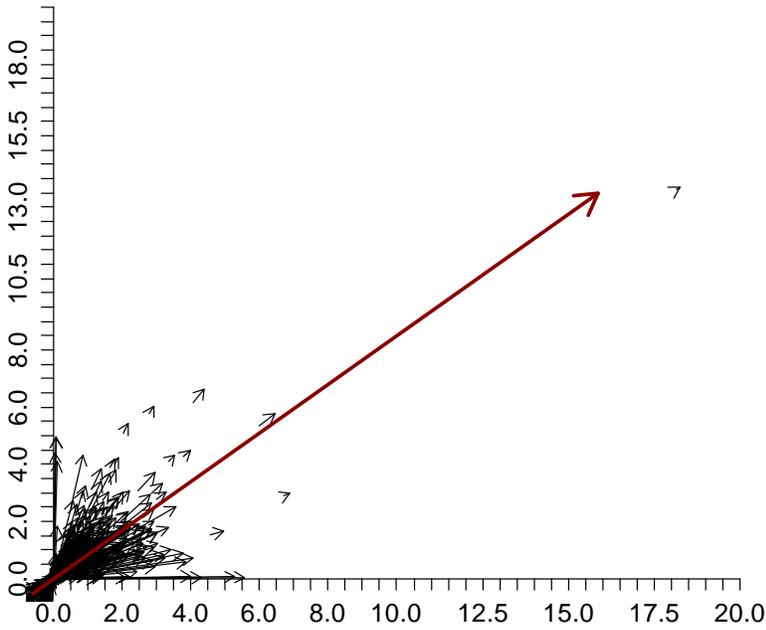


Figure 14. Item Vector Plot for ELA/literacy Grades 5 and 6 (Across Grades)

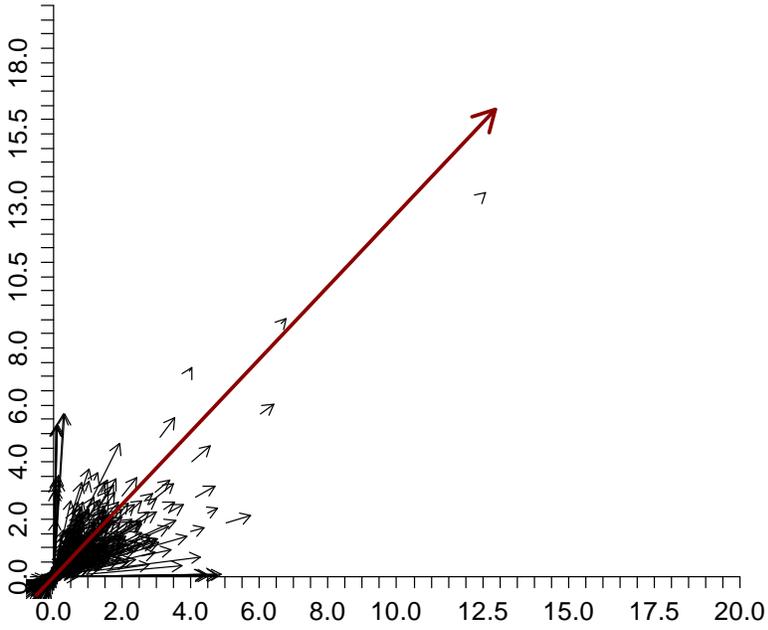


Figure 15. Item Vector Plot for ELA/literacy Grades 6 and 7 (Across Grades)

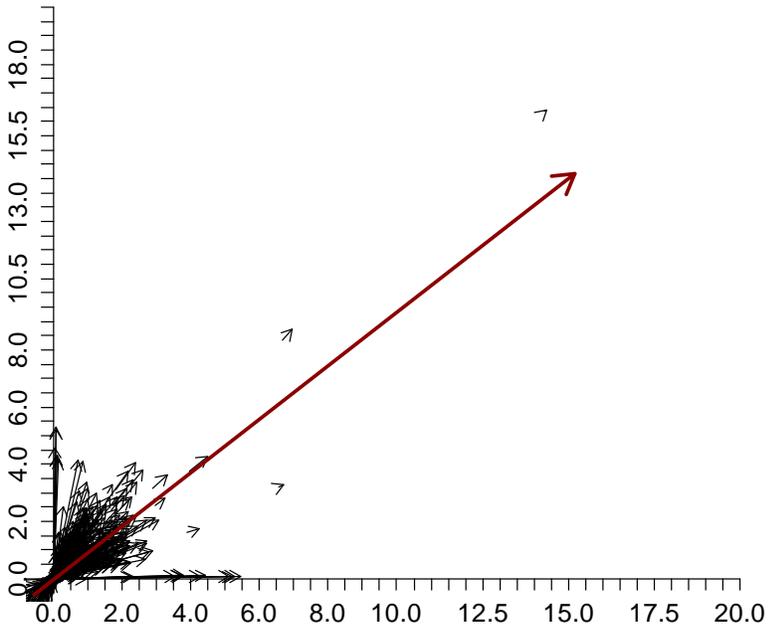


Figure 16. Item Vector Plot for ELA/literacy Grades 7 and 8 (Across Grades)

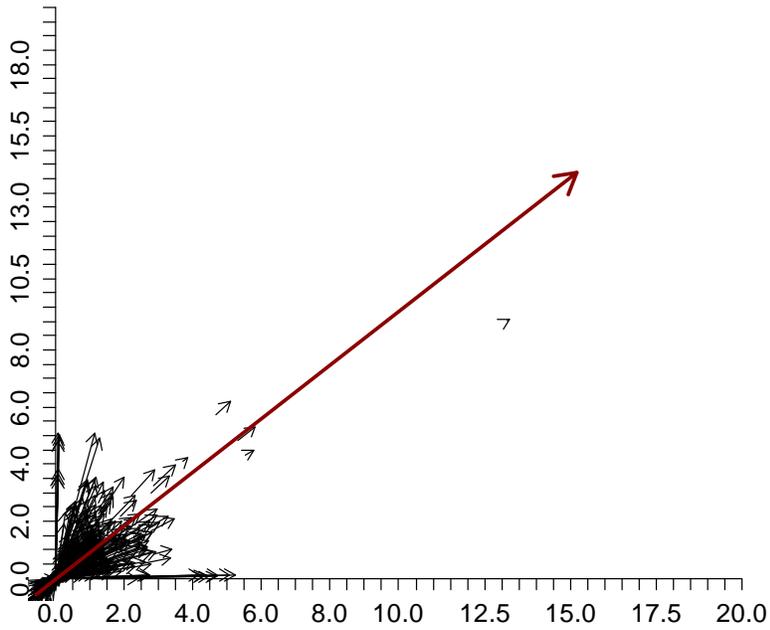


Figure 17. Item Vector Plot for ELA/literacy Grades 8 and 9 (Across Grades)

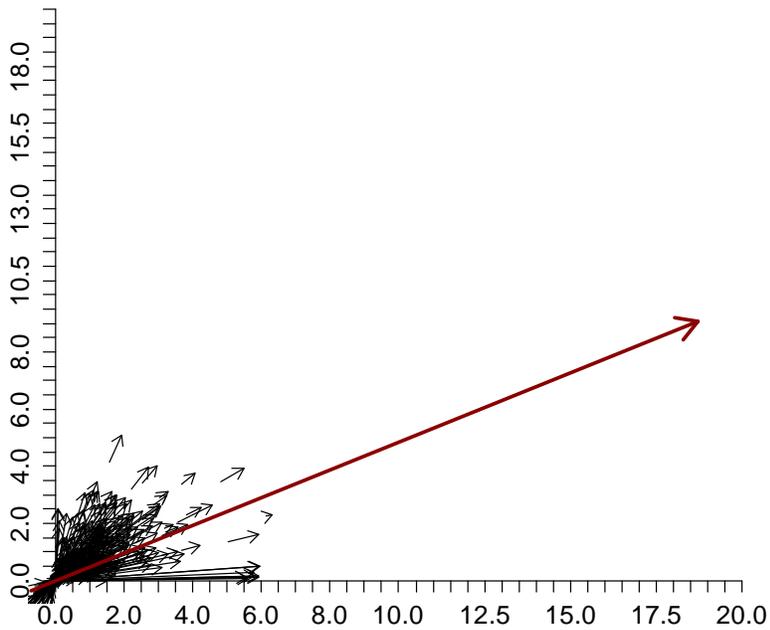


Figure 18. Item Vector Plot for ELA/literacy Grades 9 and 10 (Across Grades)

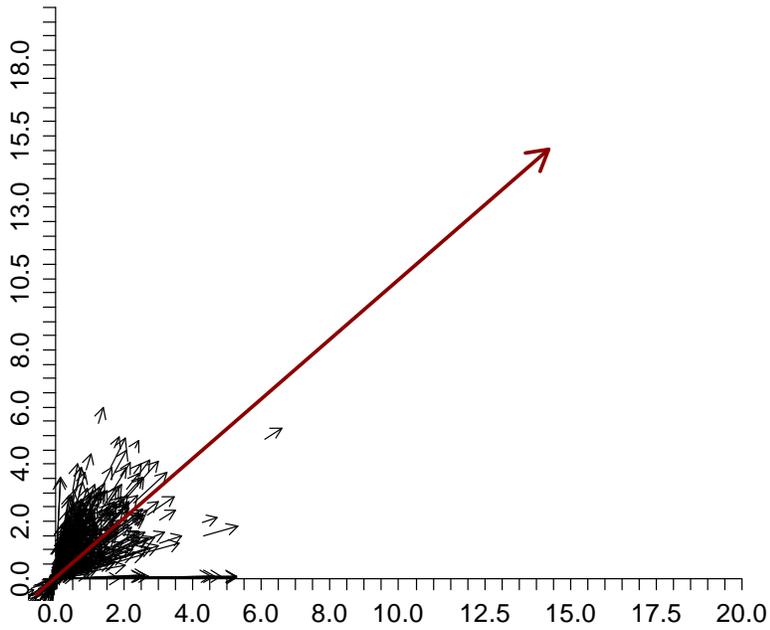


Figure 19. Item Vector Plot for ELA/literacy Grades 10 and 11 (Across Grades)

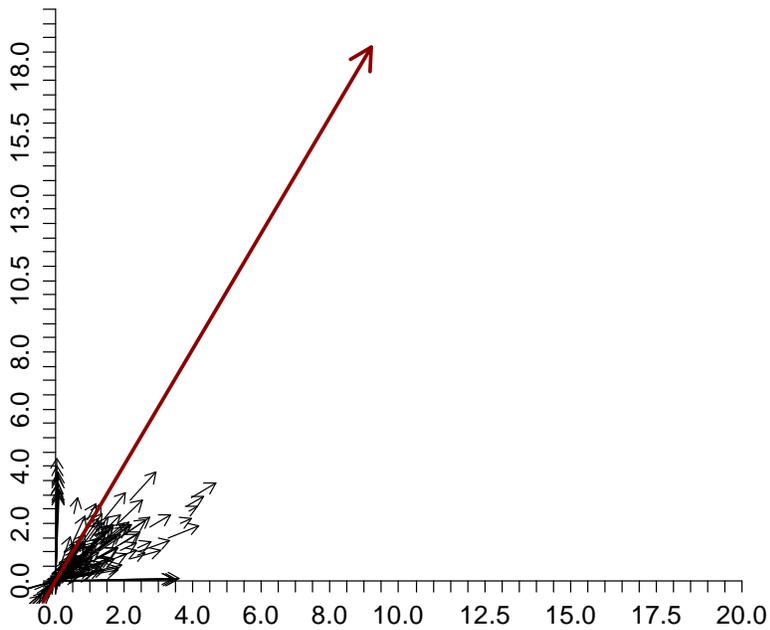


Figure 20. Item Vector Plots for the Subset of ELA/literacy Grades 3 and 4 Vertical Linking Items

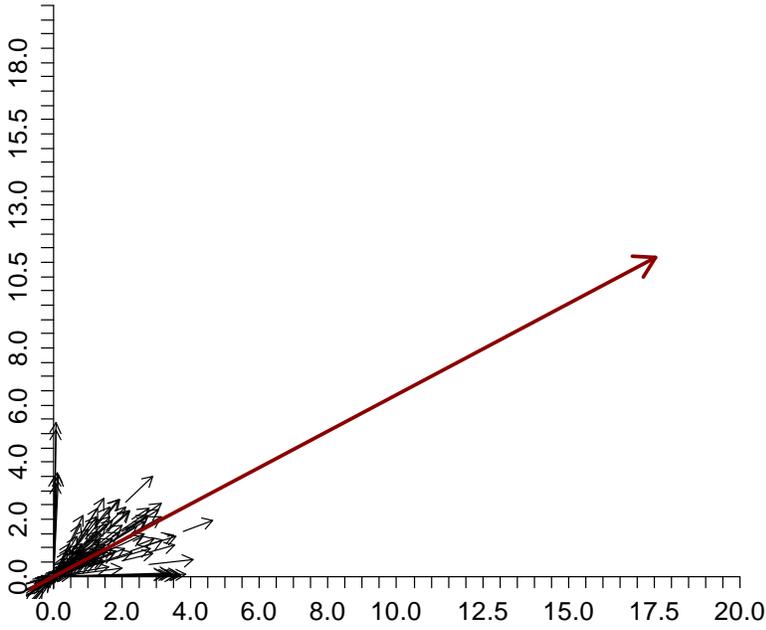


Figure 21. Item Vector Plots for the Subset of ELA/literacy Grades 4 and 5 Vertical Linking Items

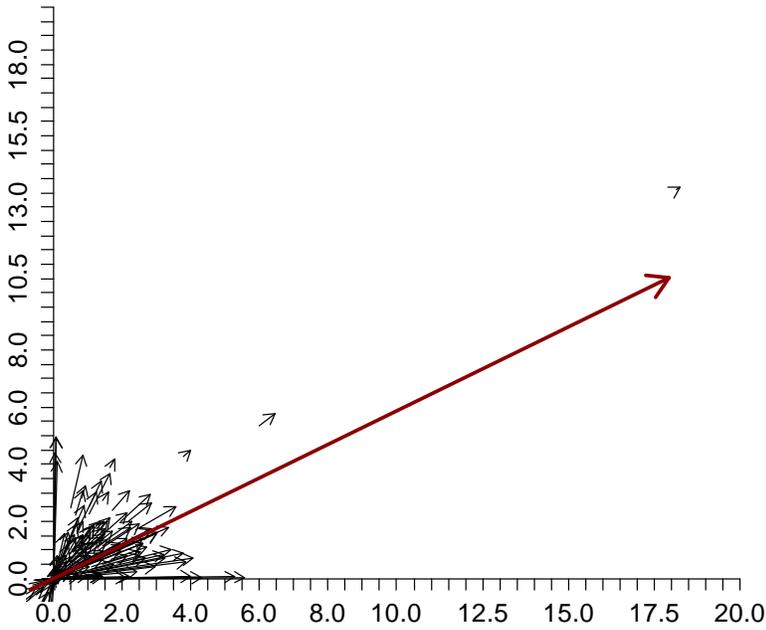


Figure 22. Item Vector Plots for the Subset of ELA/literacy Grades 5 and 6 Vertical Linking Items

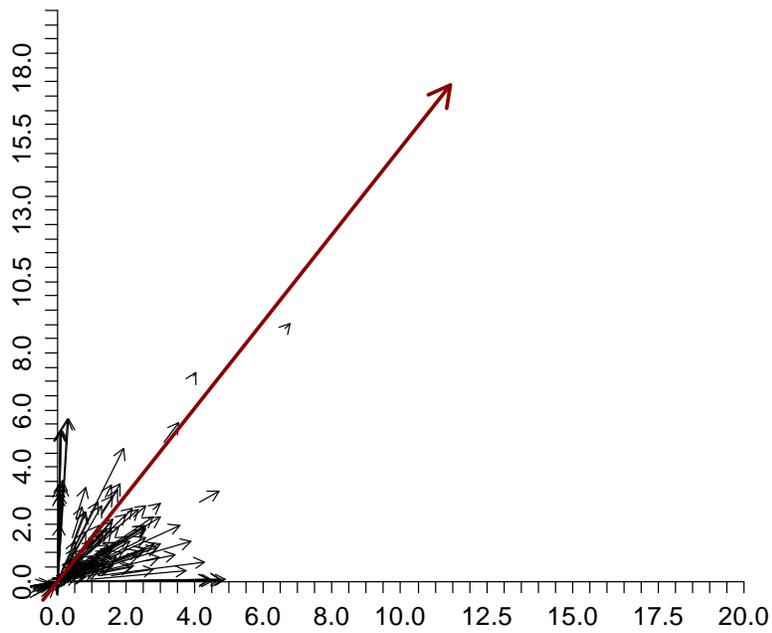


Figure 23. Item Vector Plots for the Subset of ELA/literacy Grades 6 and 7 Vertical Linking Items

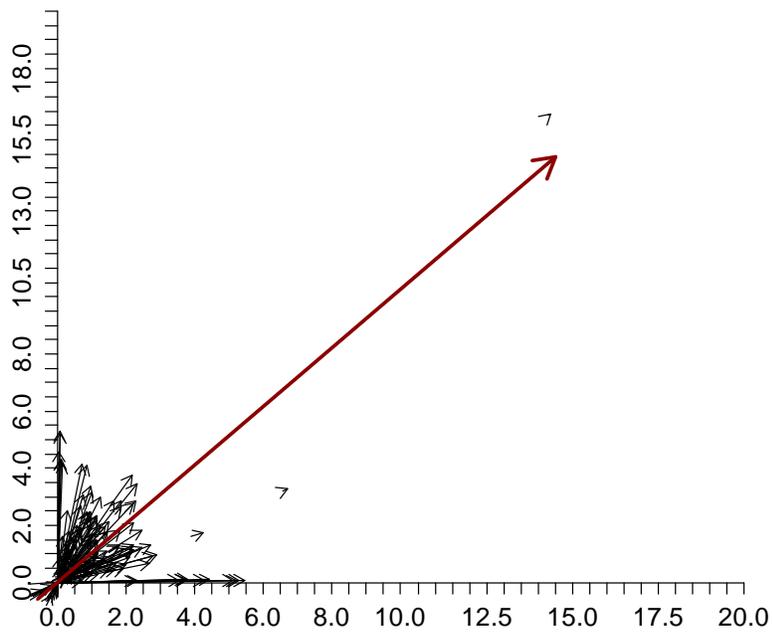


Figure 24. Item Vector Plots for the Subset of ELA/literacy Grades 7 and 8 Vertical Linking Items

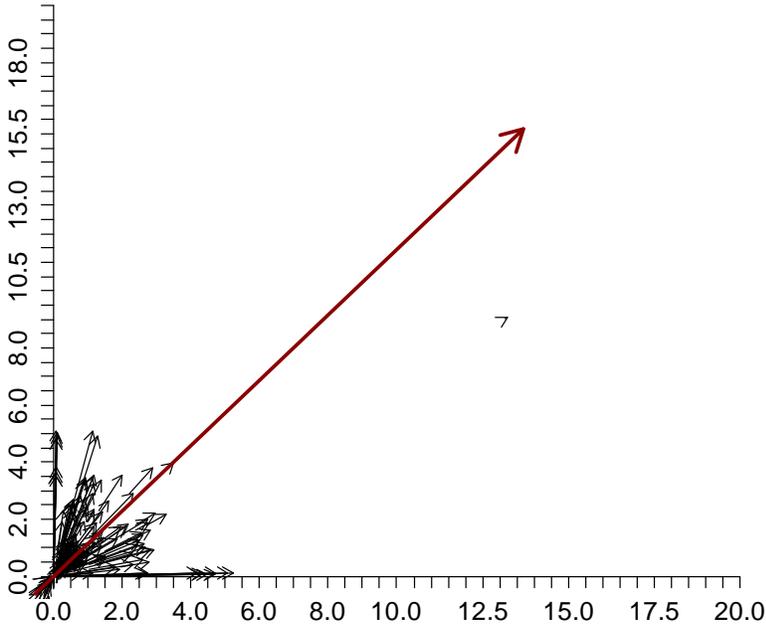


Figure 25. Item Vector Plots for the Subset of ELA/literacy Grades 8 and 9 Vertical Linking Items

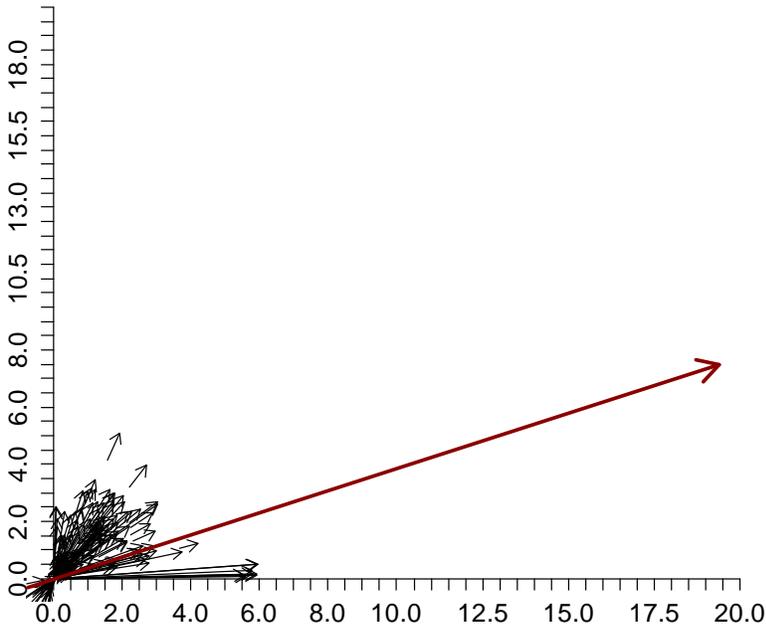


Figure 26. Item Vector Plots for the Subset of ELA/literacy Grades 9 and 10 Vertical Linking Items

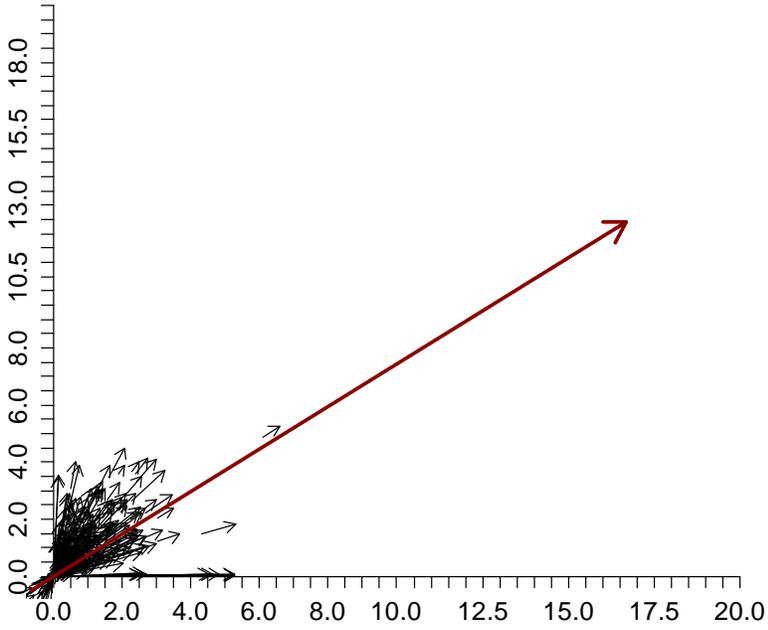


Figure 27. Item Vector Plots for the Subset of ELA/literacy Grades 10 and 11 Vertical Linking Items

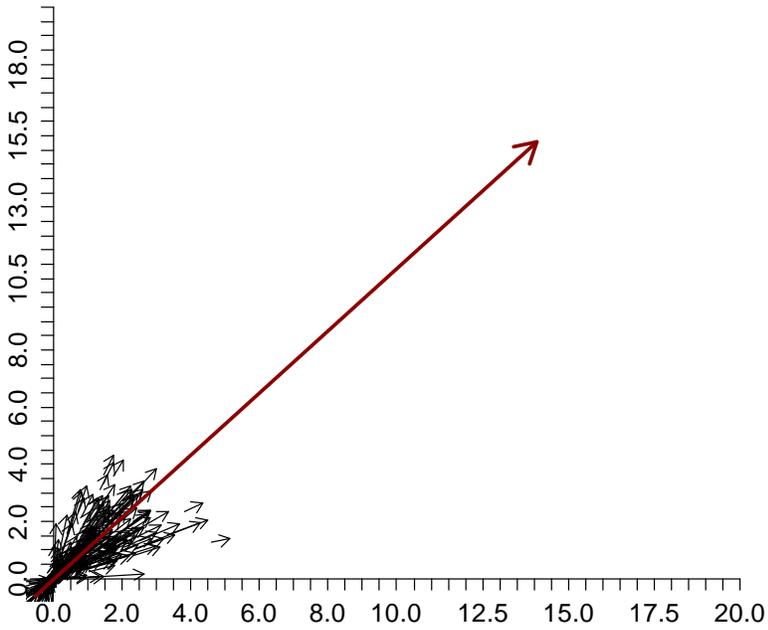


Figure 28. Item Vector Plot for Mathematics Grade 3 (Within Grade)

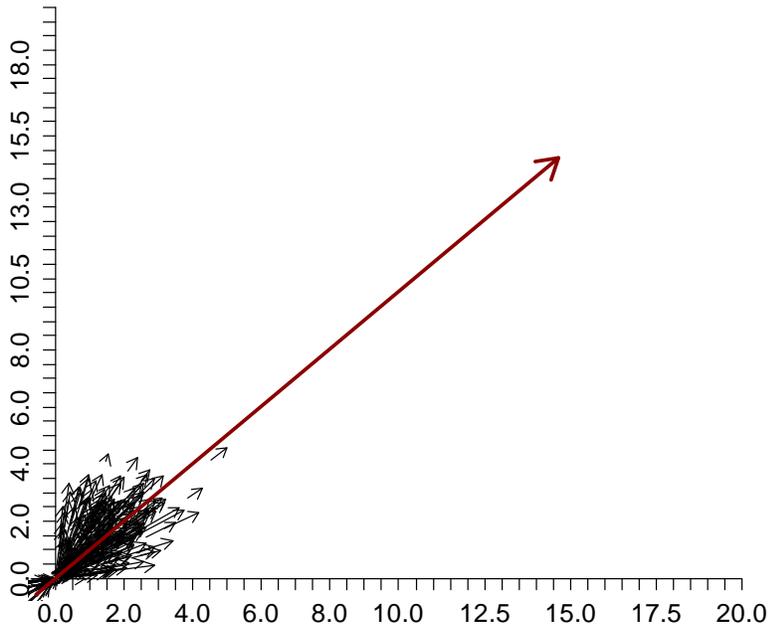


Figure 29. Item Vector Plot for Mathematics Grade 4 (Within Grade)

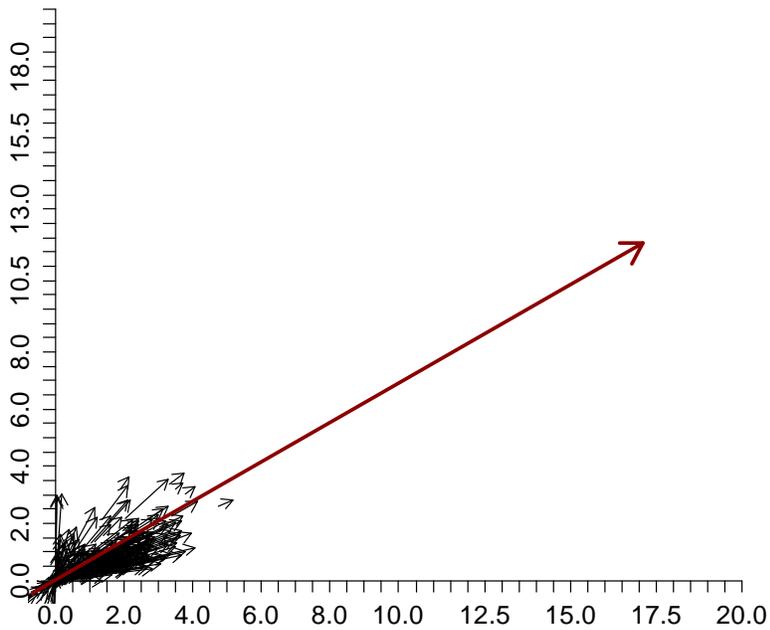


Figure 30. Item Vector Plot for Mathematics Grade 5 (Within Grade)

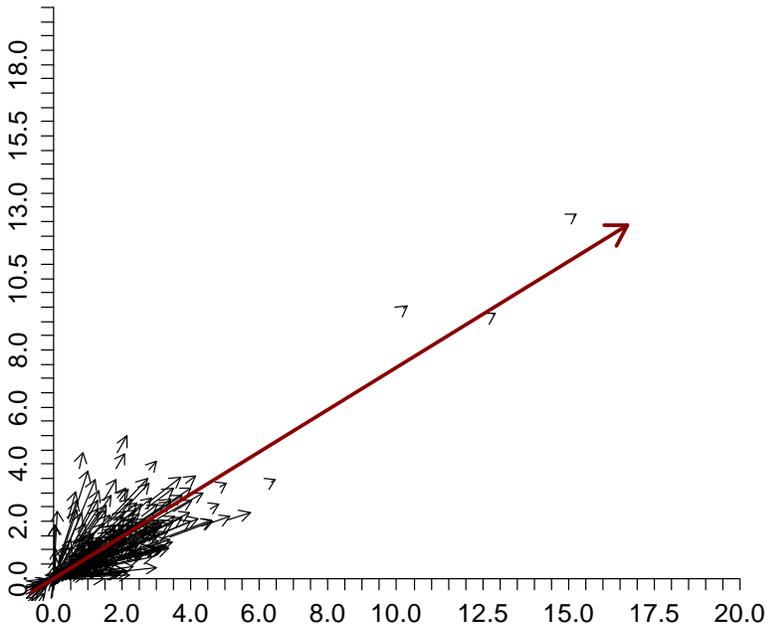


Figure 31. Item Vector Plot for Mathematics Grade 6 (Within Grade)

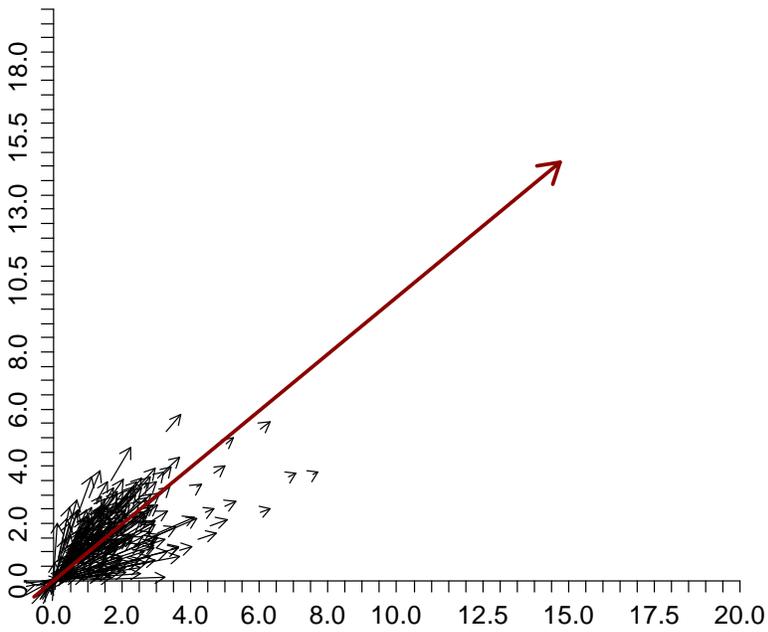


Figure 32. Item Vector Plot for Mathematics Grade 7 (Within Grade)

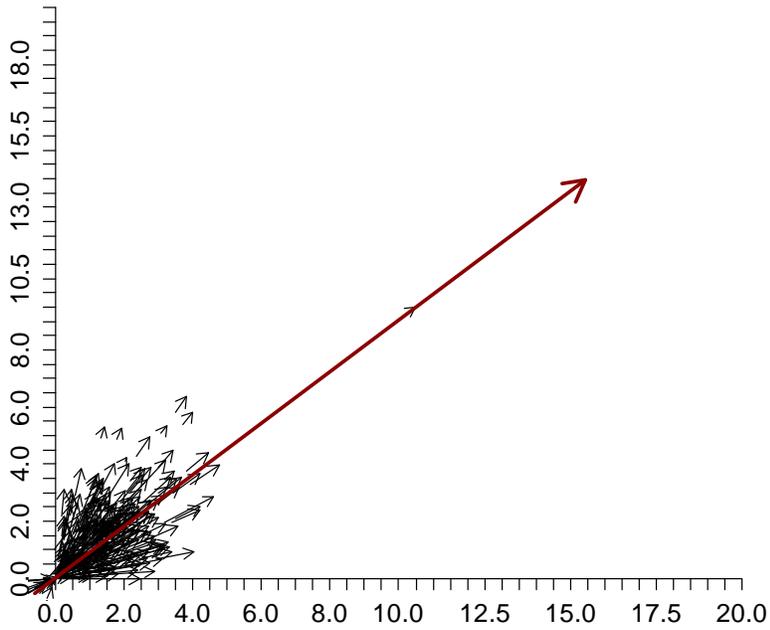


Figure 33. Item Vector Plot for Mathematics Grade 8 (Within Grade)

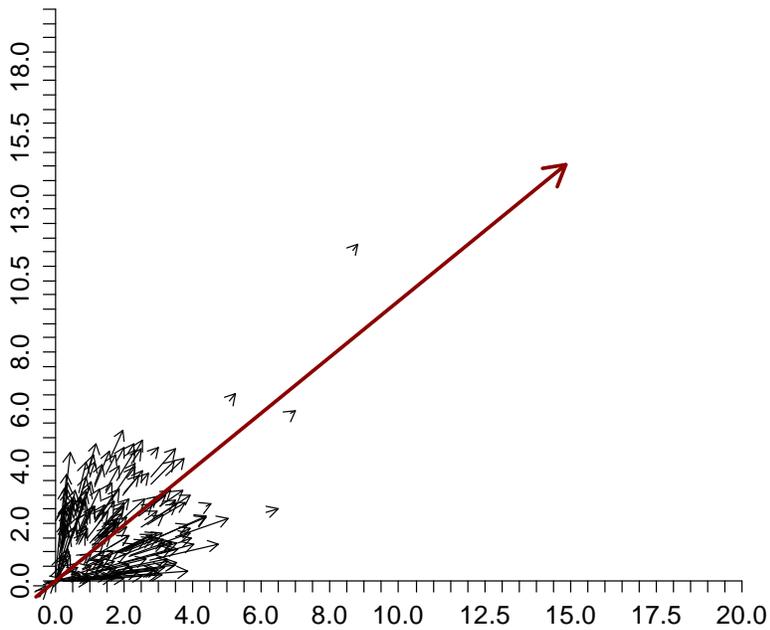


Figure 34. Item Vector Plot for Mathematics Grade 9 (Within Grade)

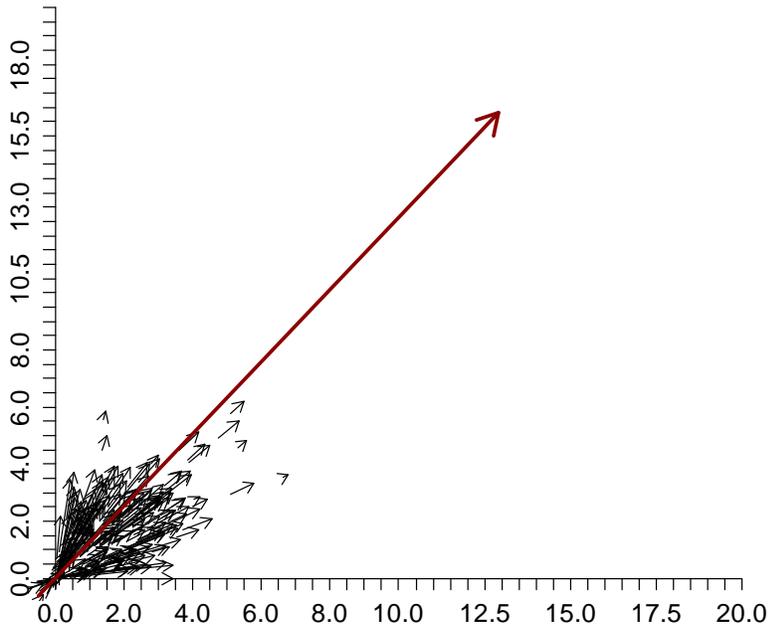


Figure 35. Item Vector Plot for Mathematics Grade 10 (Within Grade)

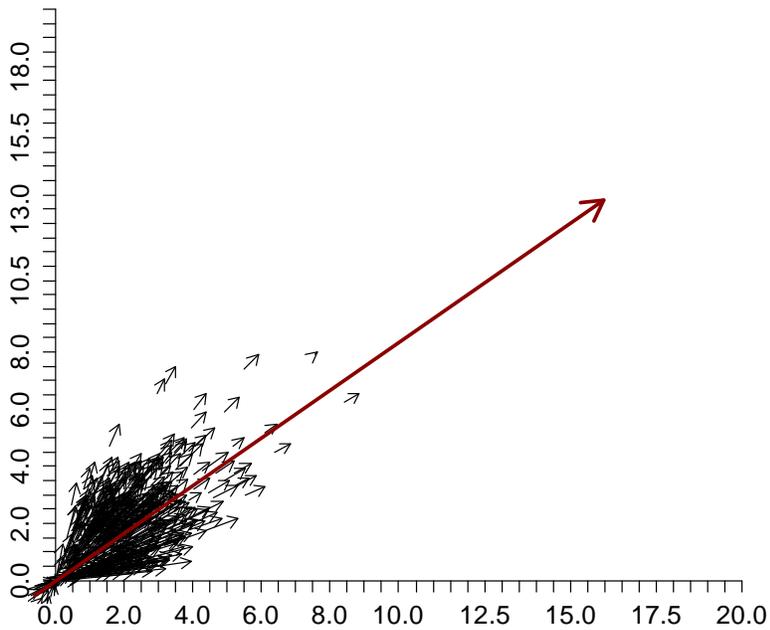


Figure 36. Item Vector Plot for Mathematics Grade 11 (Within Grade)

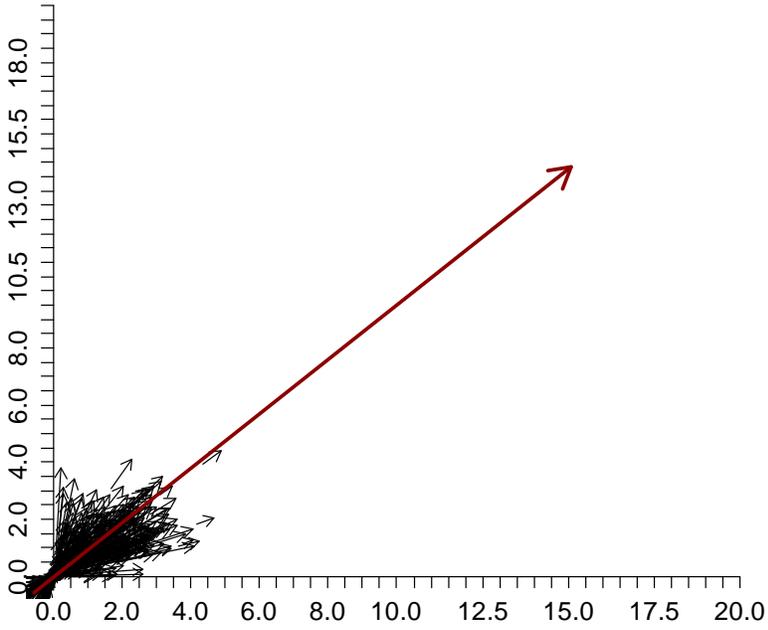


Figure 37. Item Vector Plot for Mathematics Grades 3 and 4 (Across Grades)

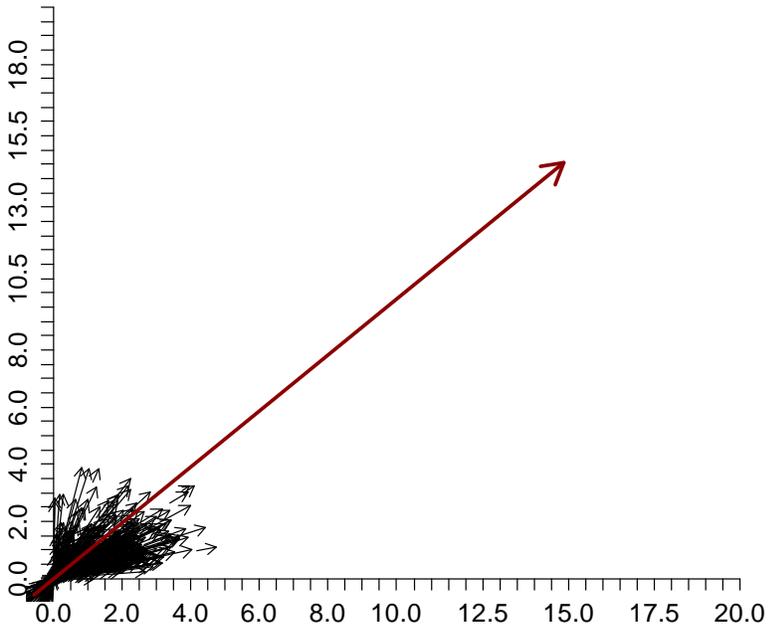


Figure 38. Item Vector Plot for Mathematics Grades 4 and 5 (Across Grades)

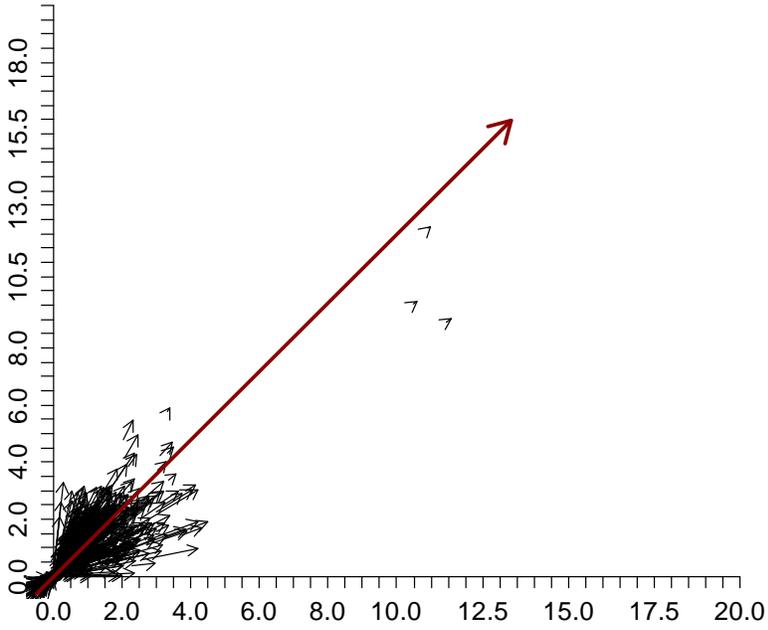


Figure 39. Item Vector Plot for Mathematics Grades 5 and 6 (Across Grades)

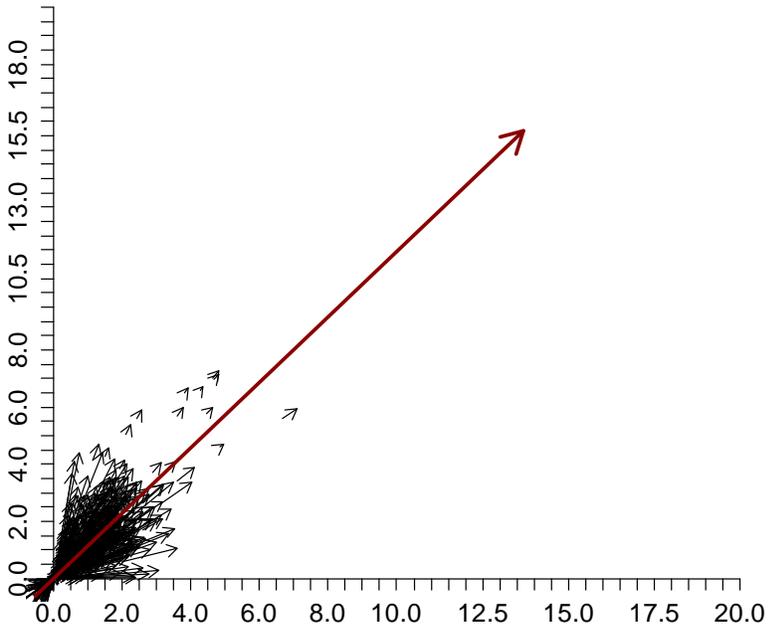


Figure 40. Item Vector Plot for Mathematics Grades 6 and 7 (Across Grades)

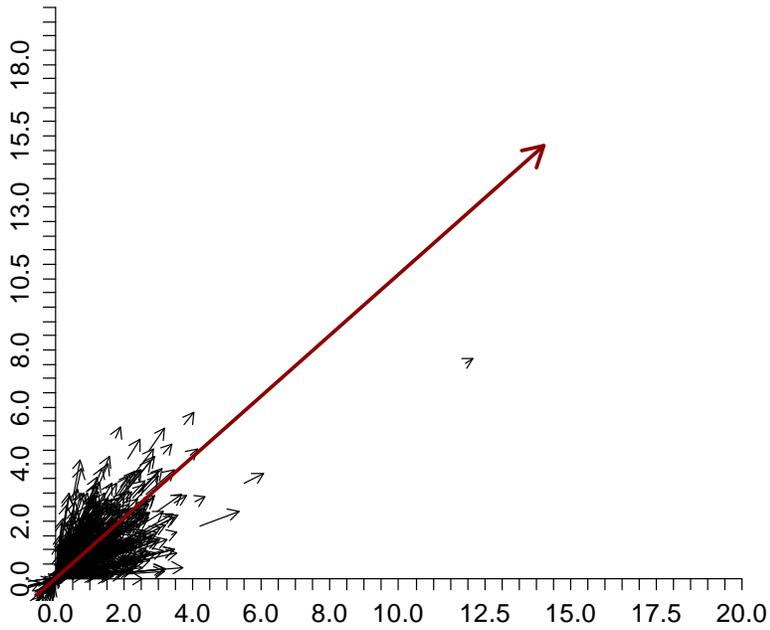


Figure 41. Item Vector Plot for Mathematics Grades 7 and 8 (Across Grades)

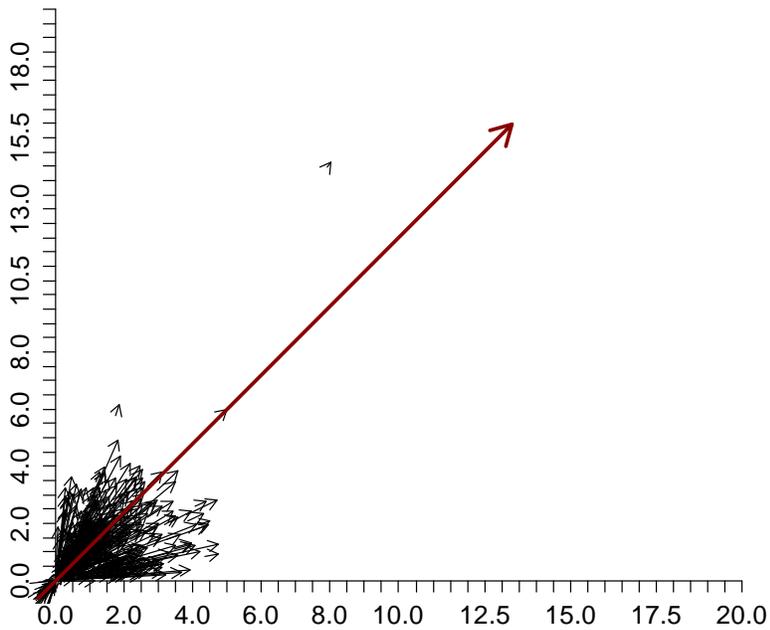


Figure 42. Item Vector Plot for Mathematics Grades 8 and 9 (Across Grades)

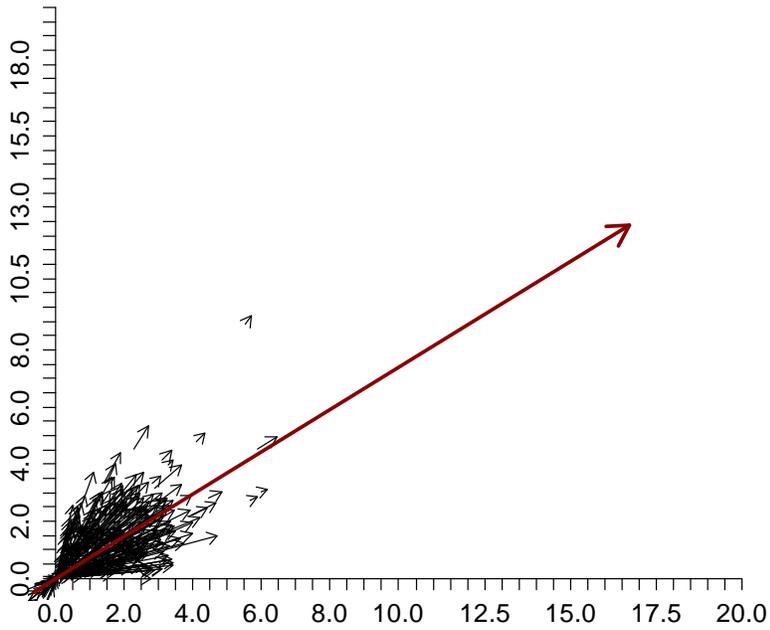


Figure 43. Item Vector Plot for Mathematics Grades 9 and 10 (Across Grades)

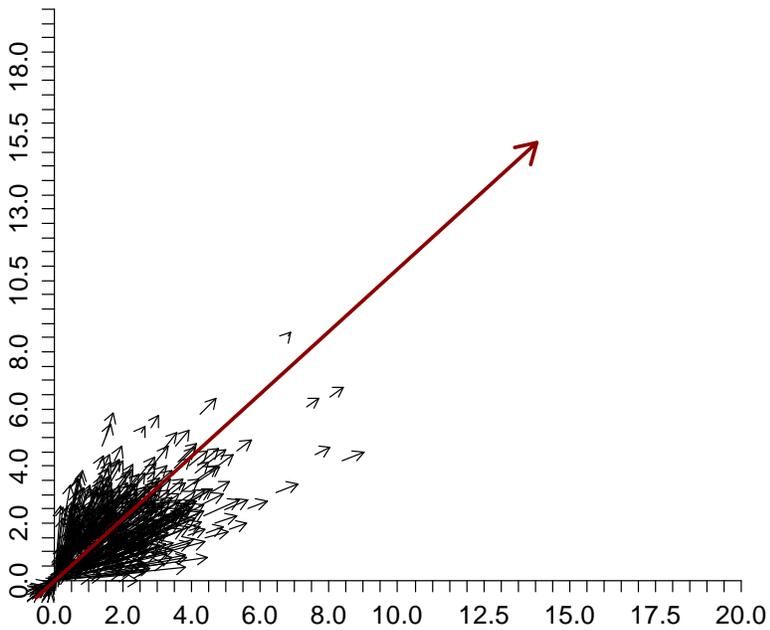


Figure 44. Item Vector Plot for Mathematics Grades 10 and 11 (Across Grades)

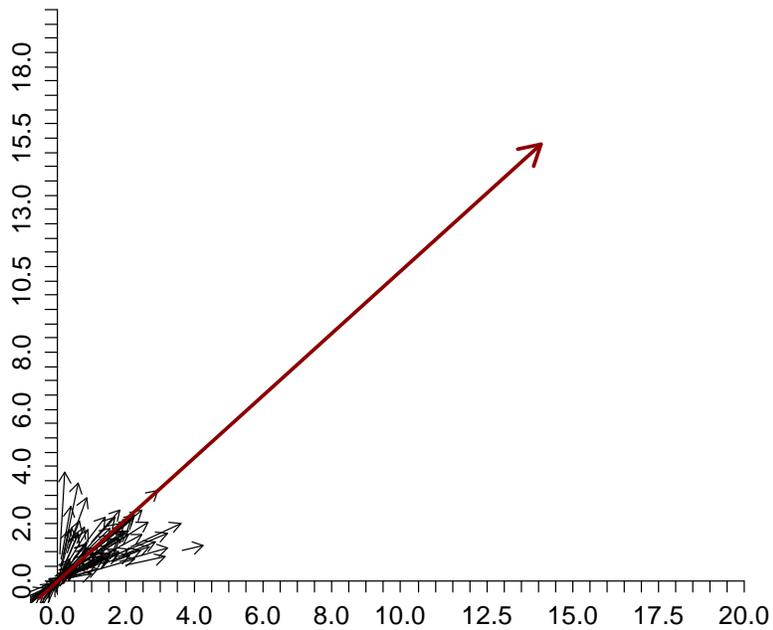


Figure 45. Item Vector Plot for the Subset of Mathematics Grades 3 and 4 (Vertical Linking Items)

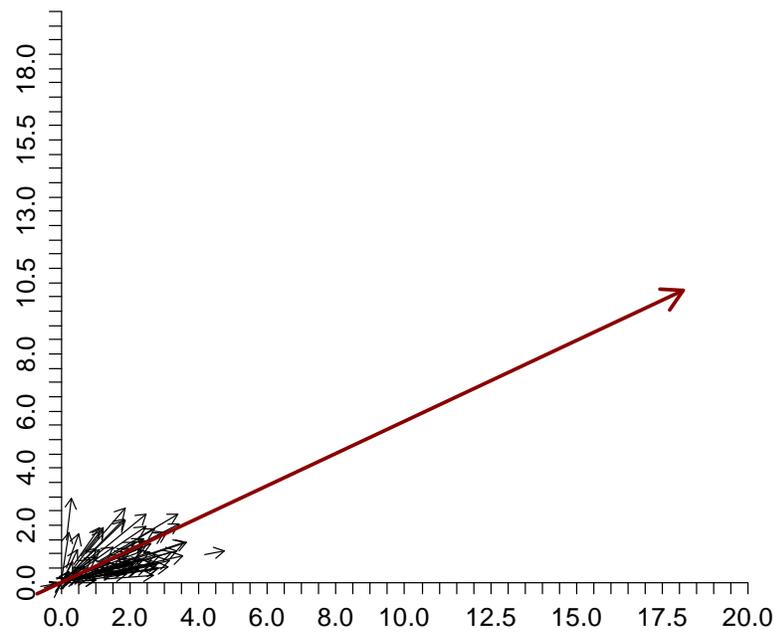


Figure 46. Item Vector Plot for the Subset of Mathematics Grades 4 and 5 (Vertical Linking Items)

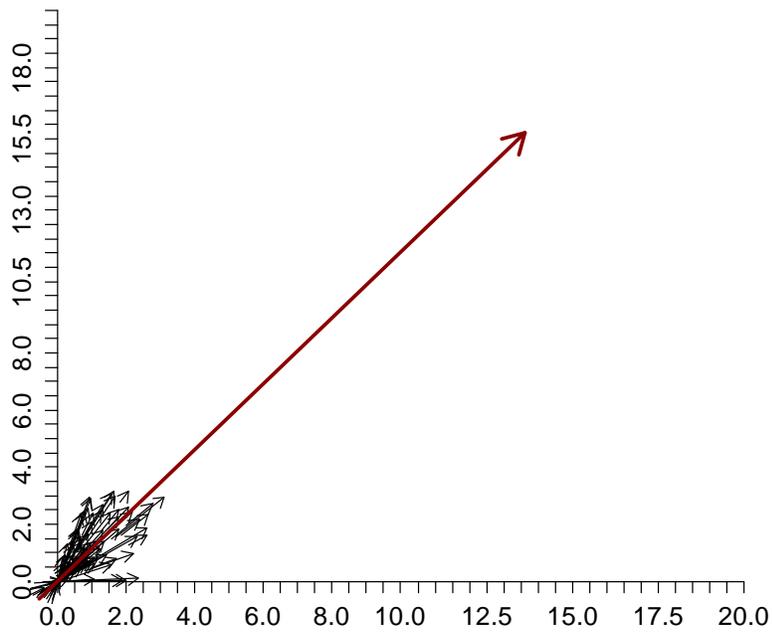


Figure 47. Item Vector Plot for the Subset of Mathematics Grades 5 and 6 (Vertical Linking Items)

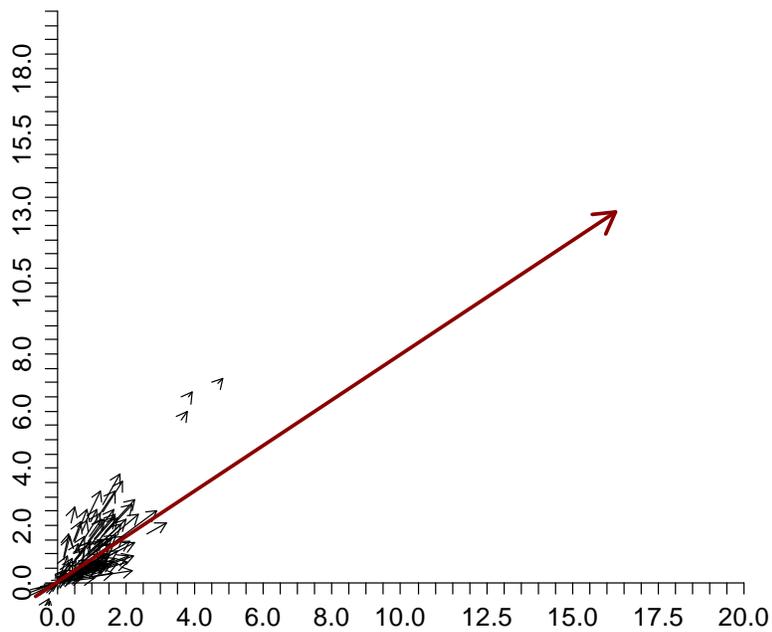


Figure 48. Item Vector Plot for the Subset of Mathematics Grades 6 and 7 (Vertical Linking Items)

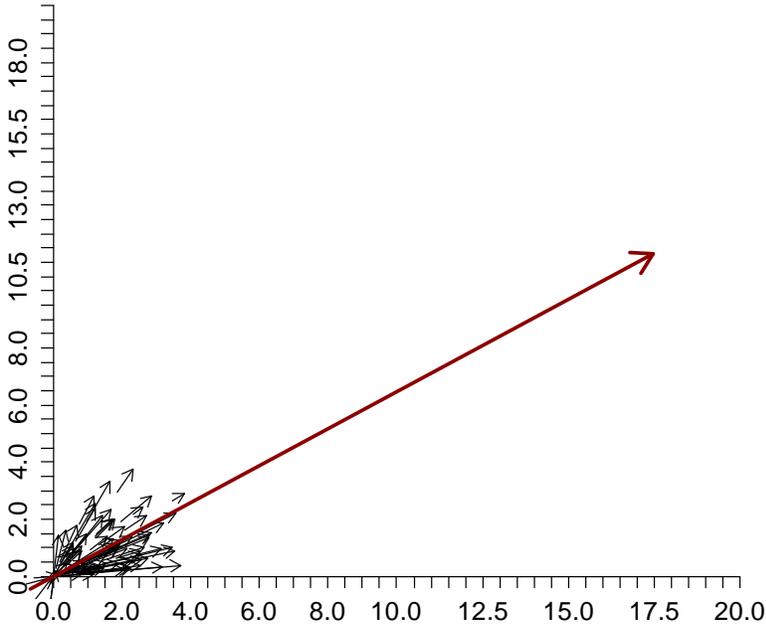


Figure 49. Item Vector Plot for the Subset of Mathematics Grades 7 and 8 (Vertical Linking Items)

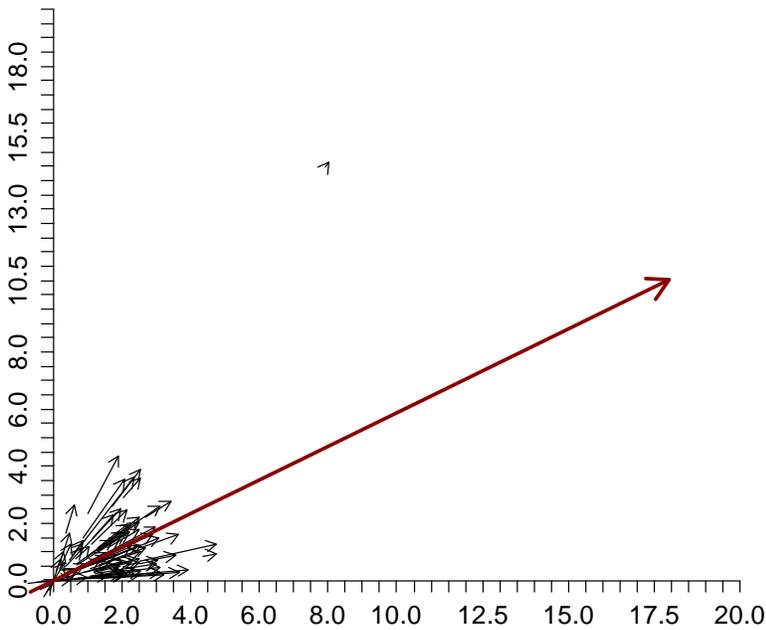


Figure 50. Item Vector Plot for the Subset of Mathematics Grades 8 and 9 (Vertical Linking Items)

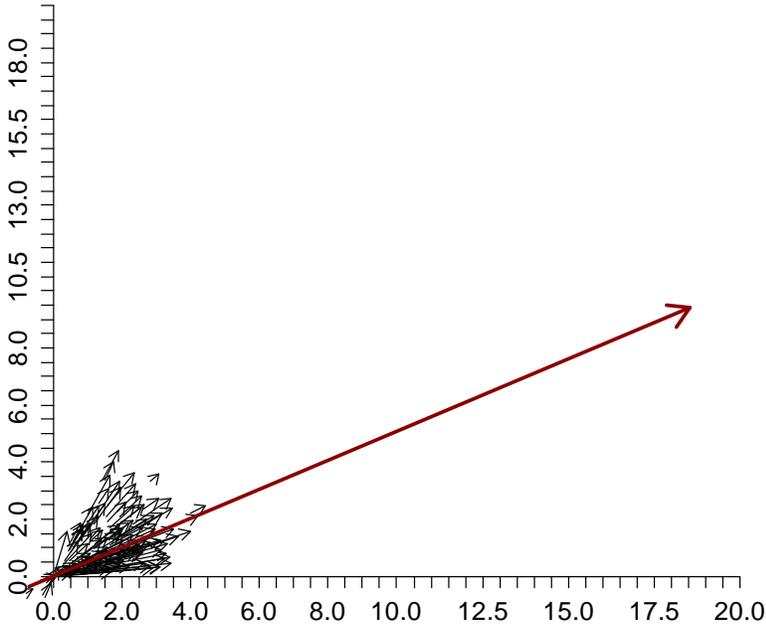


Figure 51. Item Vector Plot for the Subset of Mathematics Grades 9 and 10 (Vertical Linking Items)

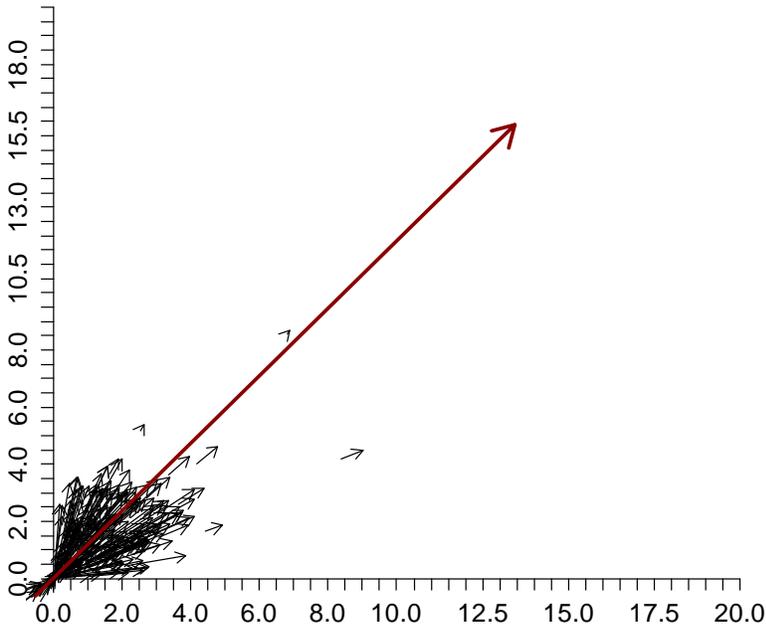


Figure 52. Item Vector Plot for the Subset of Mathematics Grades 10 and 11 (Vertical Linking Items)

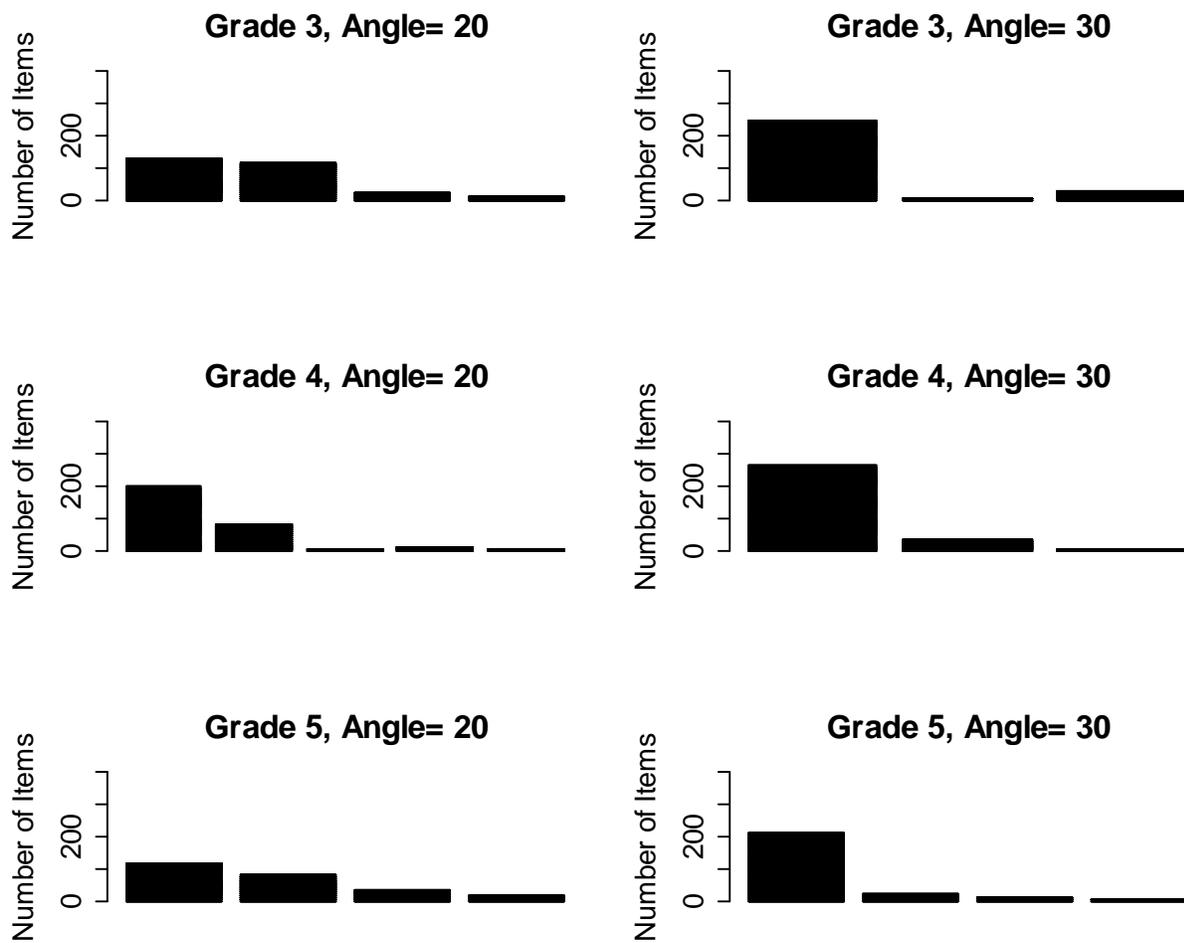


Figure 53. Clustering of Item Angle Measures for Grades 3 to 5, ELA/literacy (within grade)

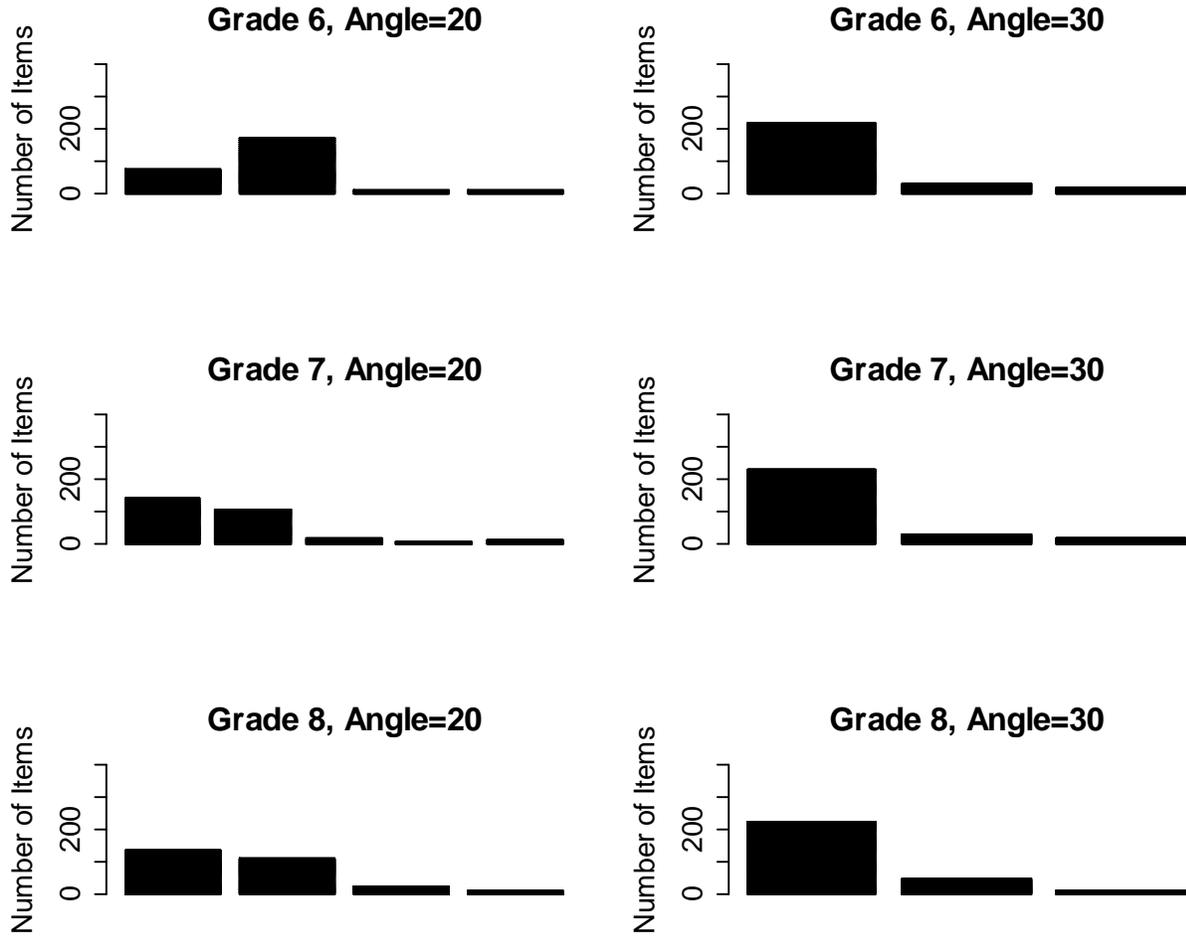


Figure 54. Clustering of Item Angle Measures for Grades 6 to 8, ELA/literacy (within grade)

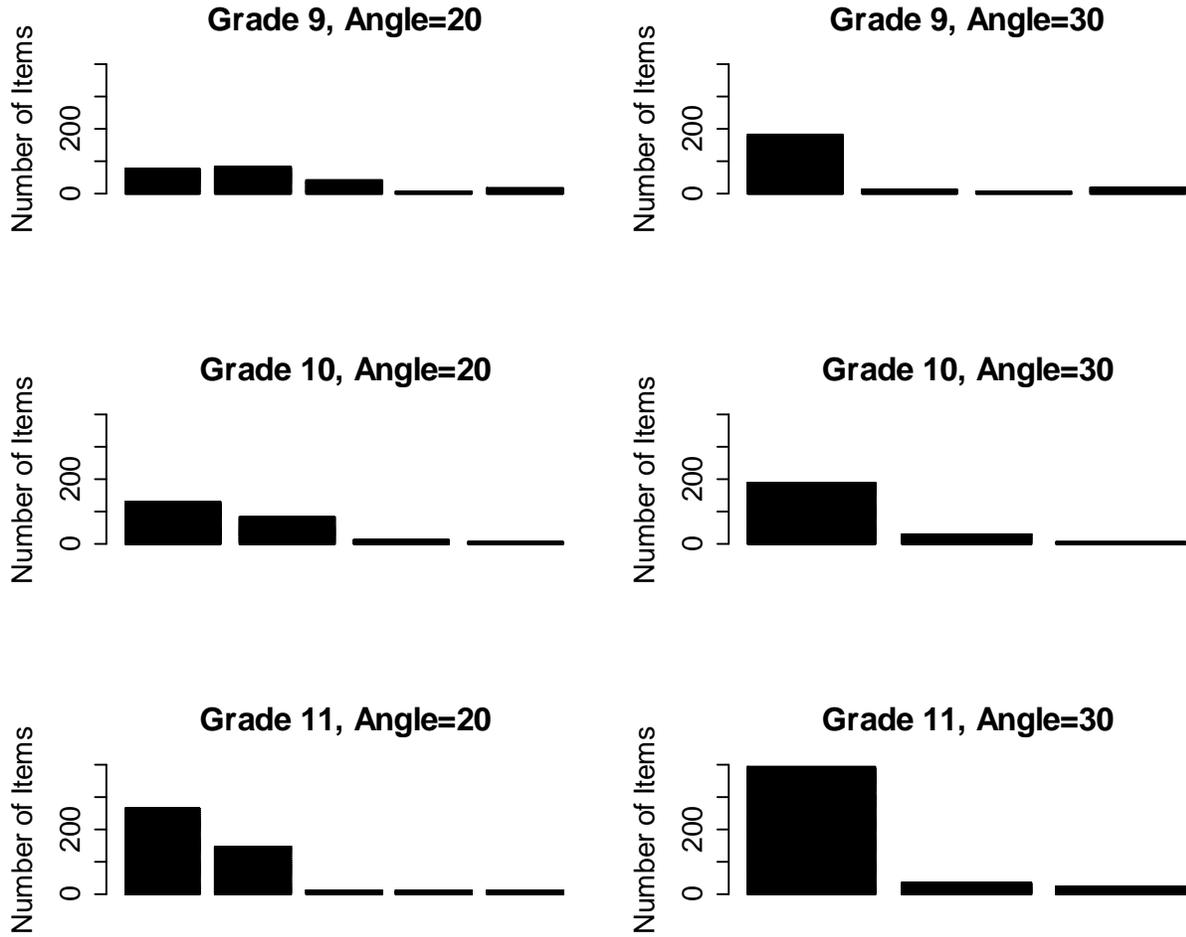


Figure 55. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (within grade)

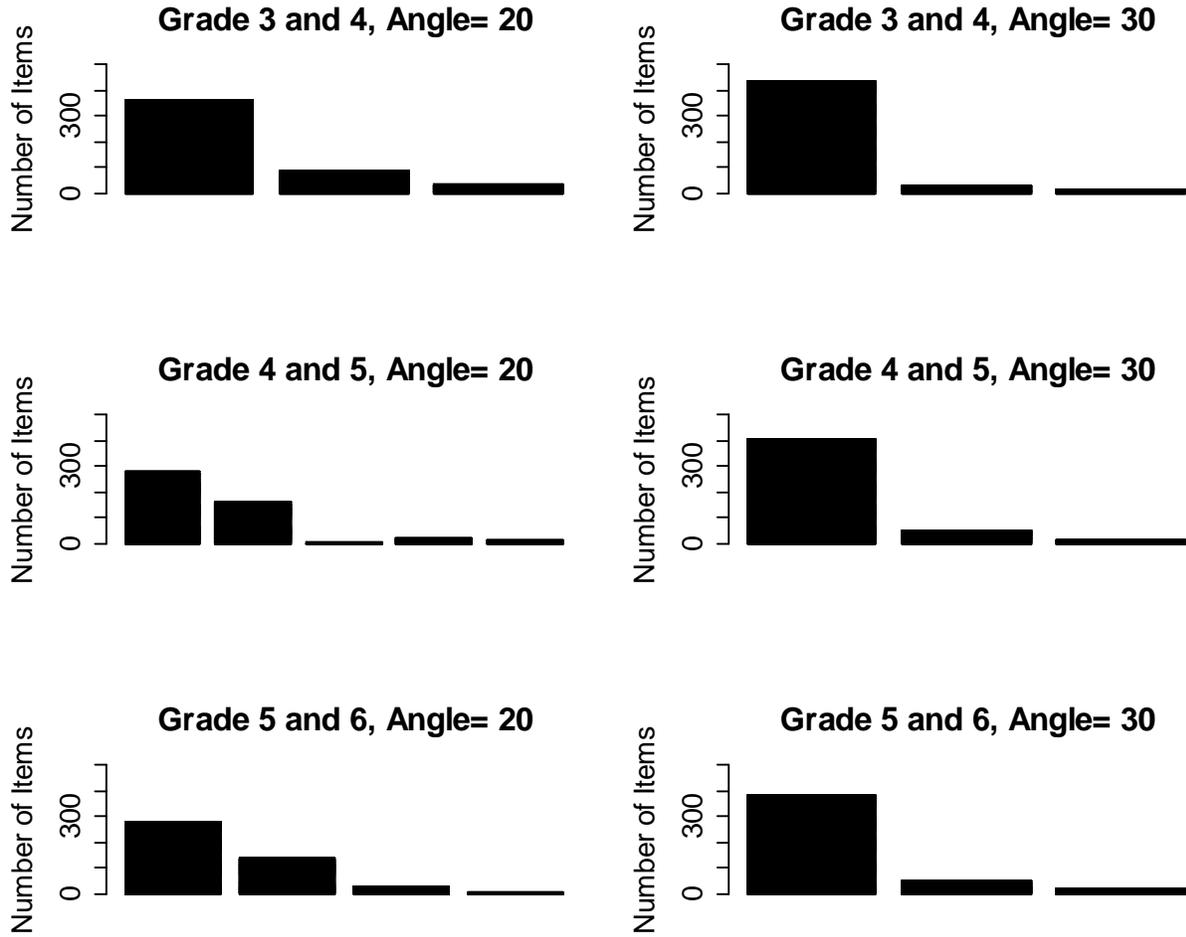


Figure 56. Clustering of Item Angle Measures for Grades 3 to 6, ELA/literacy (across grades)

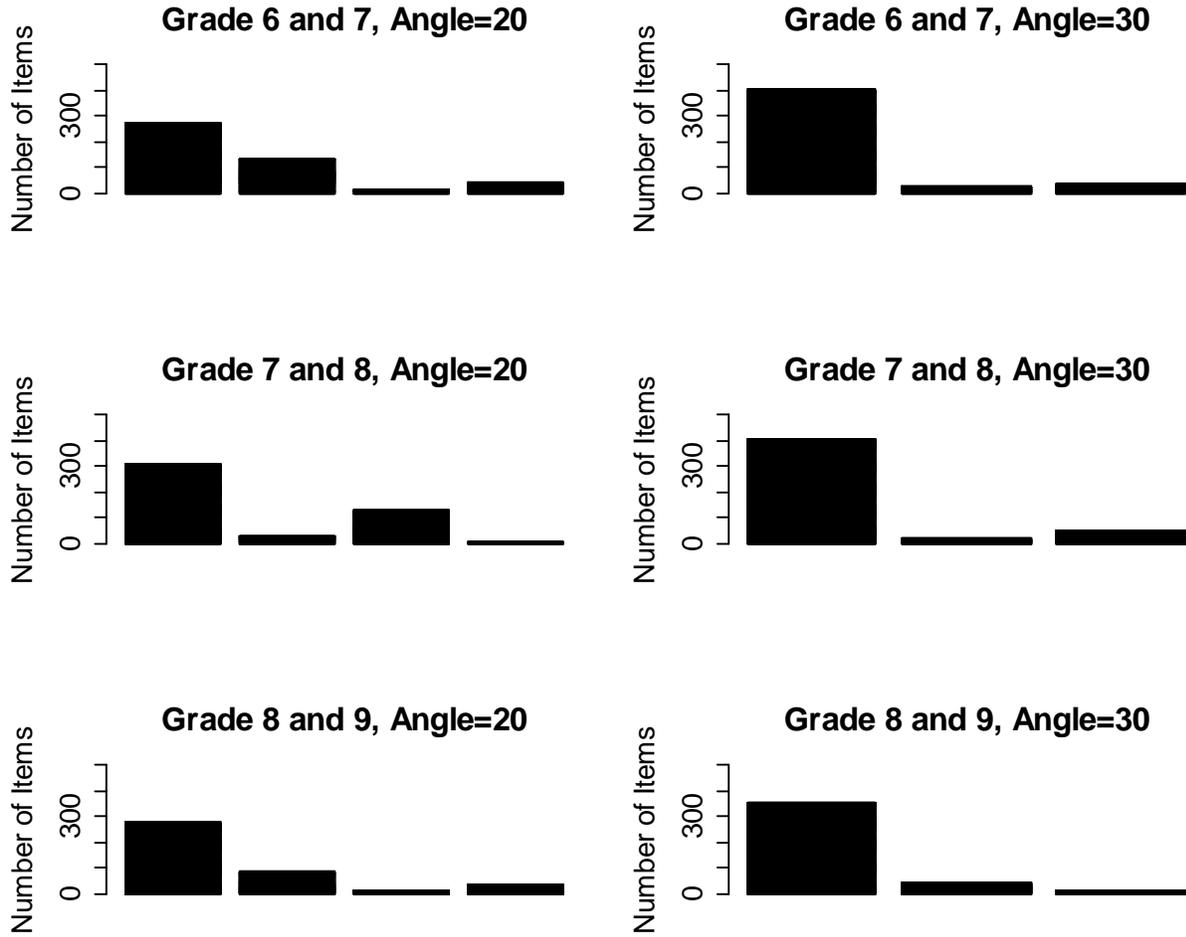


Figure 57. Clustering of Item Angle Measures for Grades 6 to 9, ELA/literacy (across grades)

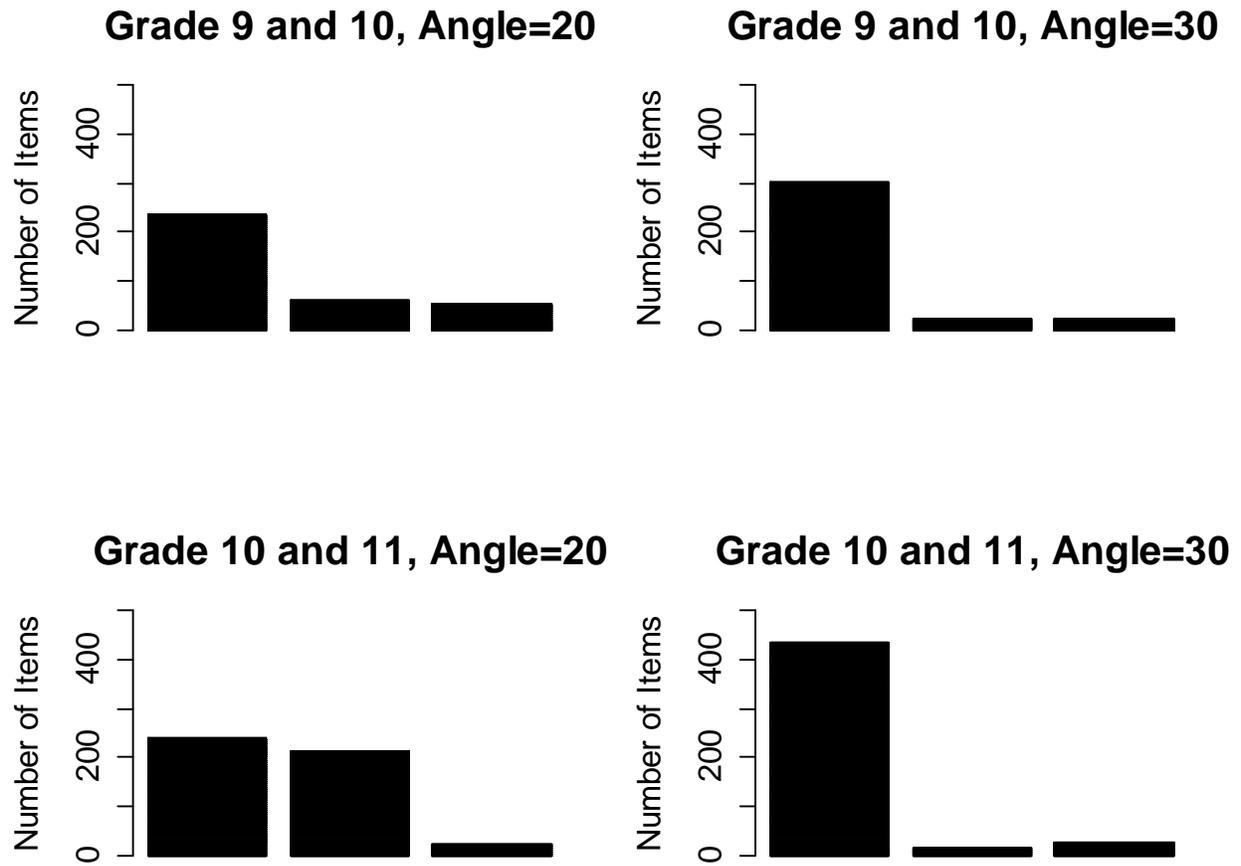


Figure 58. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (across grades)

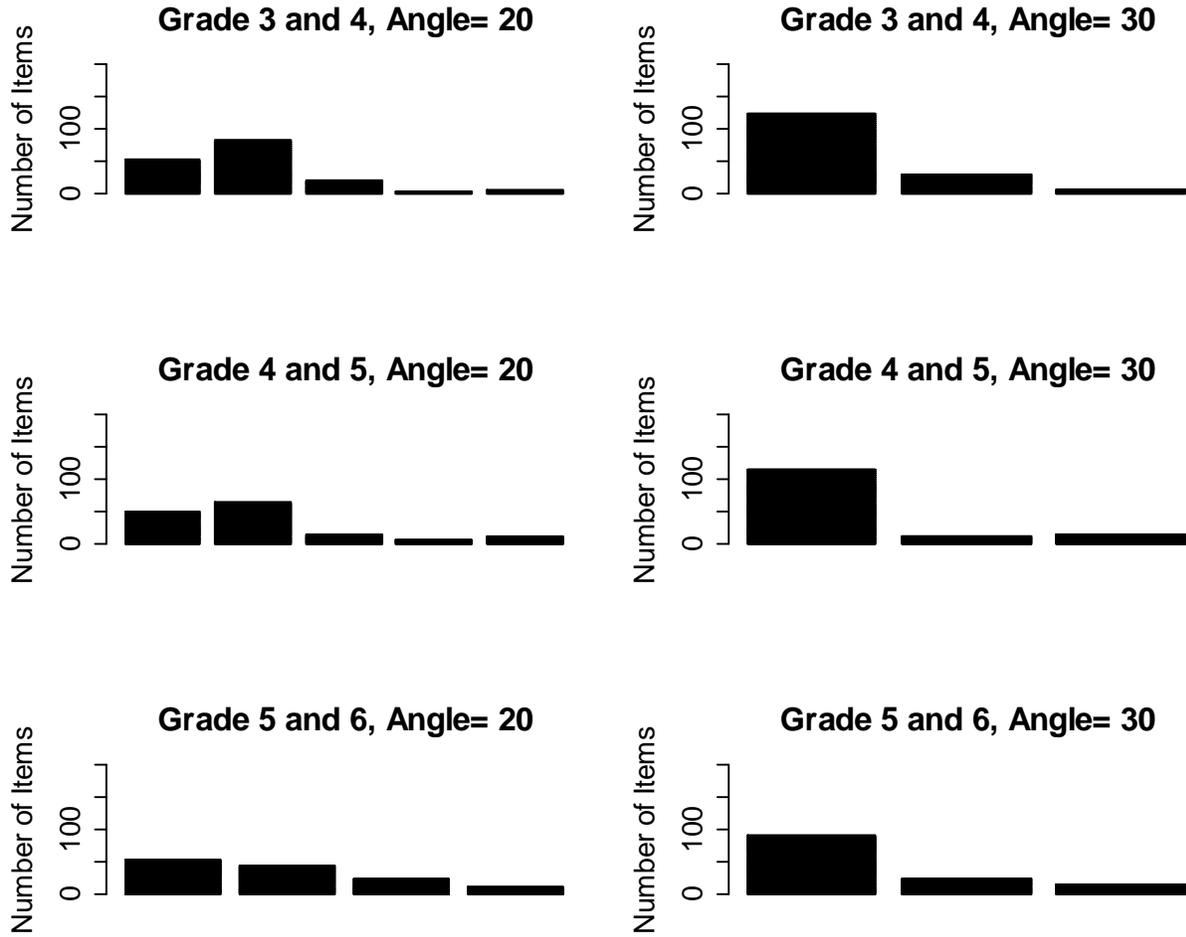


Figure 59. Clustering of Item Angle Measures for Grades 3 to 6, ELA/literacy (vertical linking)

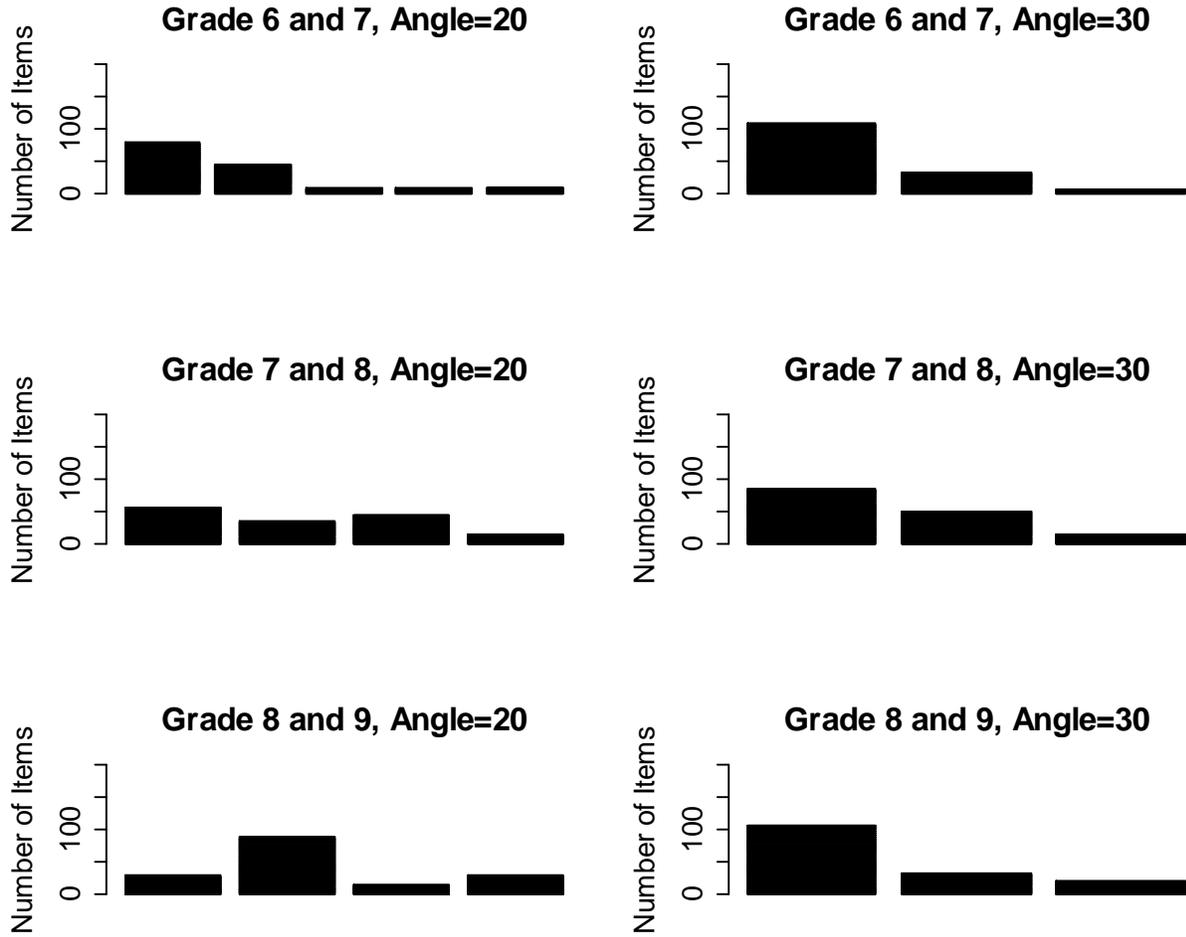


Figure 60. Clustering of Item Angle Measures for Grades 6 to 9, ELA/literacy (vertical linking)

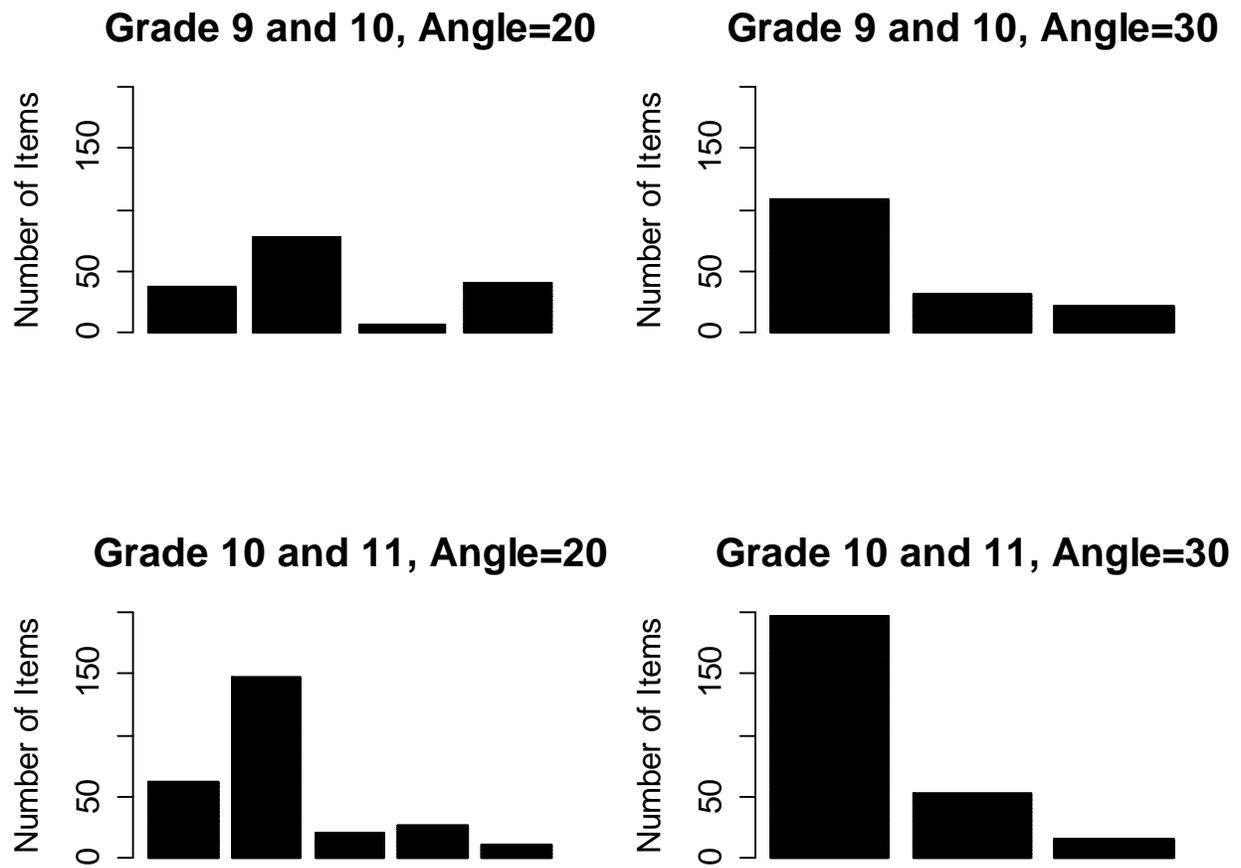


Figure 61. Clustering of Item Angle Measures for Grades 9 to 11, ELA/literacy (vertical linking)

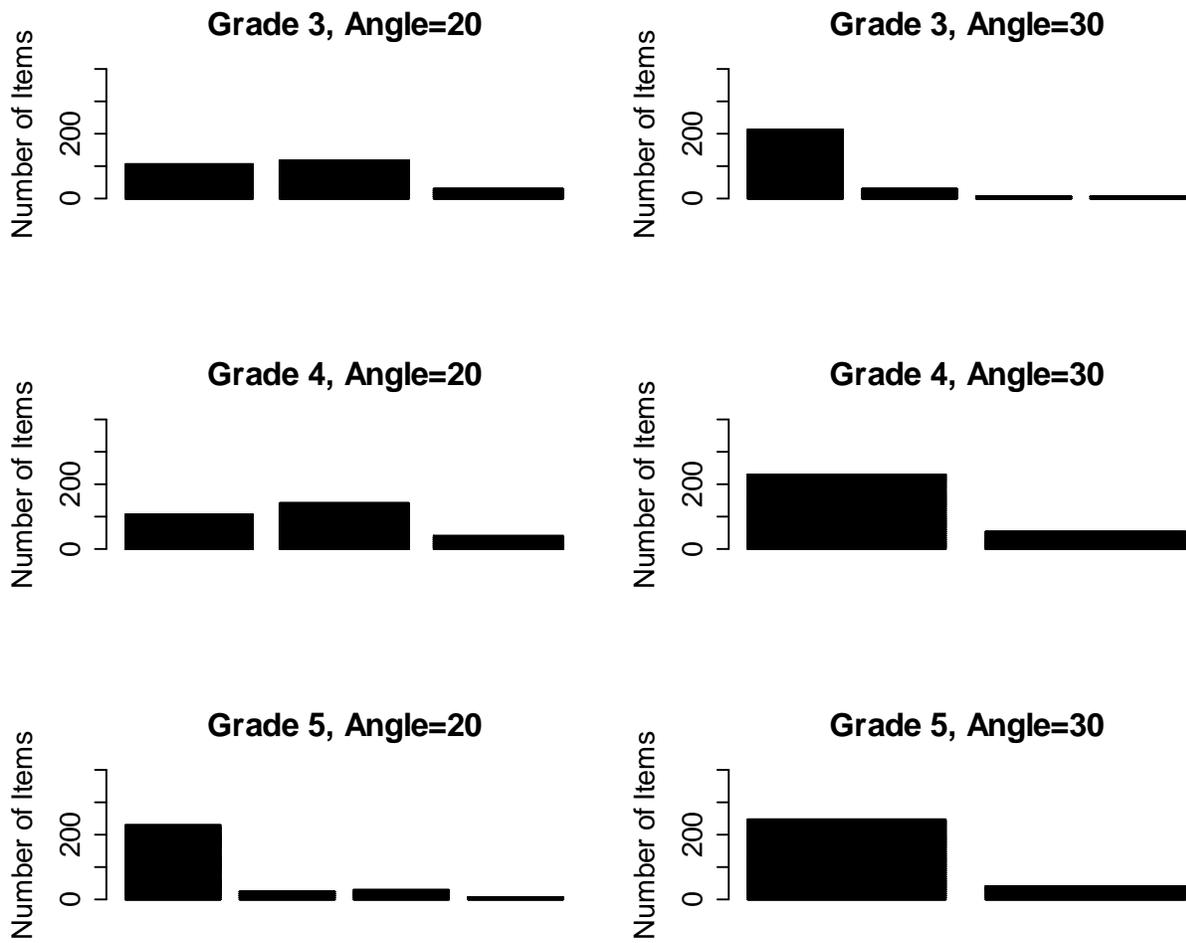


Figure 62. Clustering of Item Angle Measures for Grades 3 to 5, Mathematics (within grade)

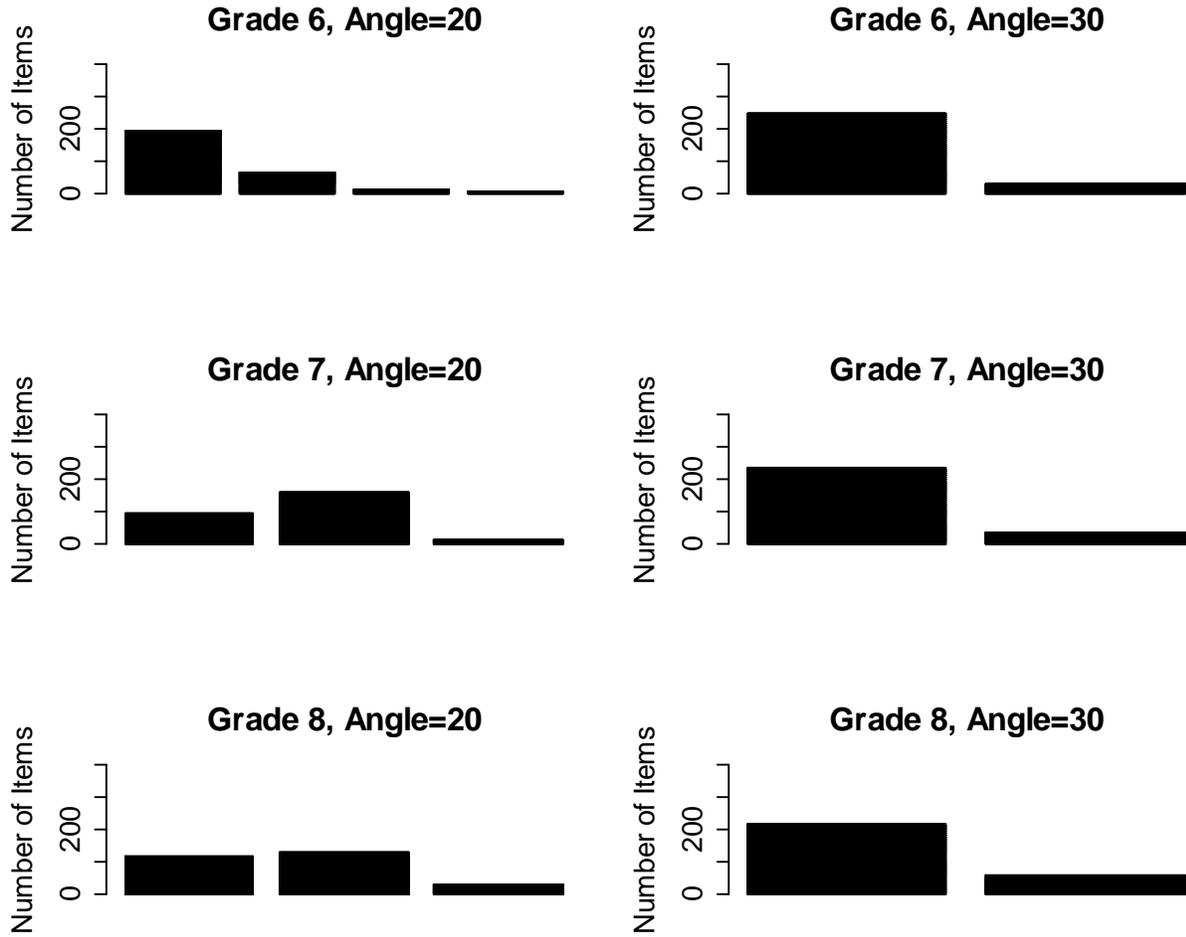


Figure 63. Clustering of Item Angle Measures for Grades 6 to 8, Mathematics (within grade)

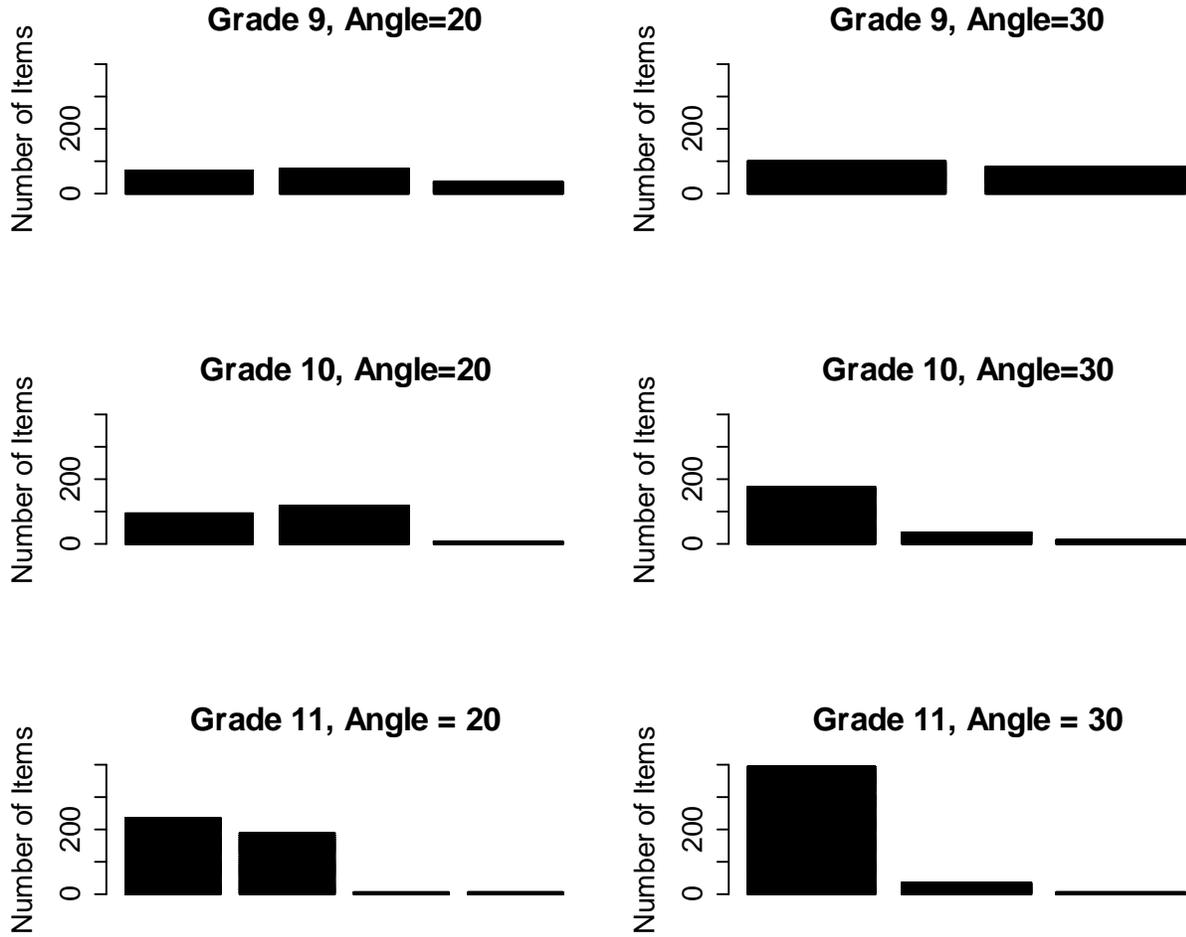


Figure 64. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (within grade)

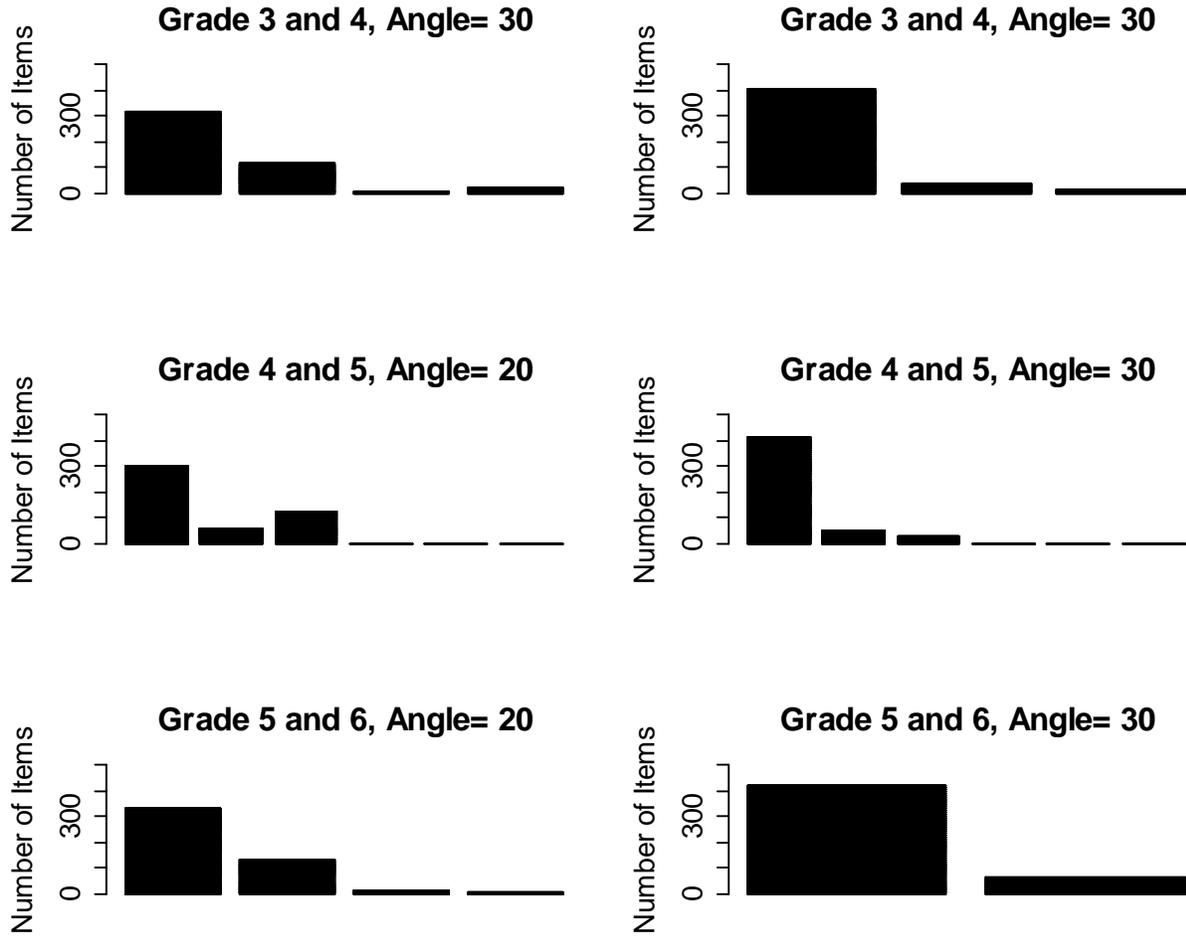


Figure 65. Clustering of Item Angle Measures for Grades 3 to 6, Mathematics (across grades)

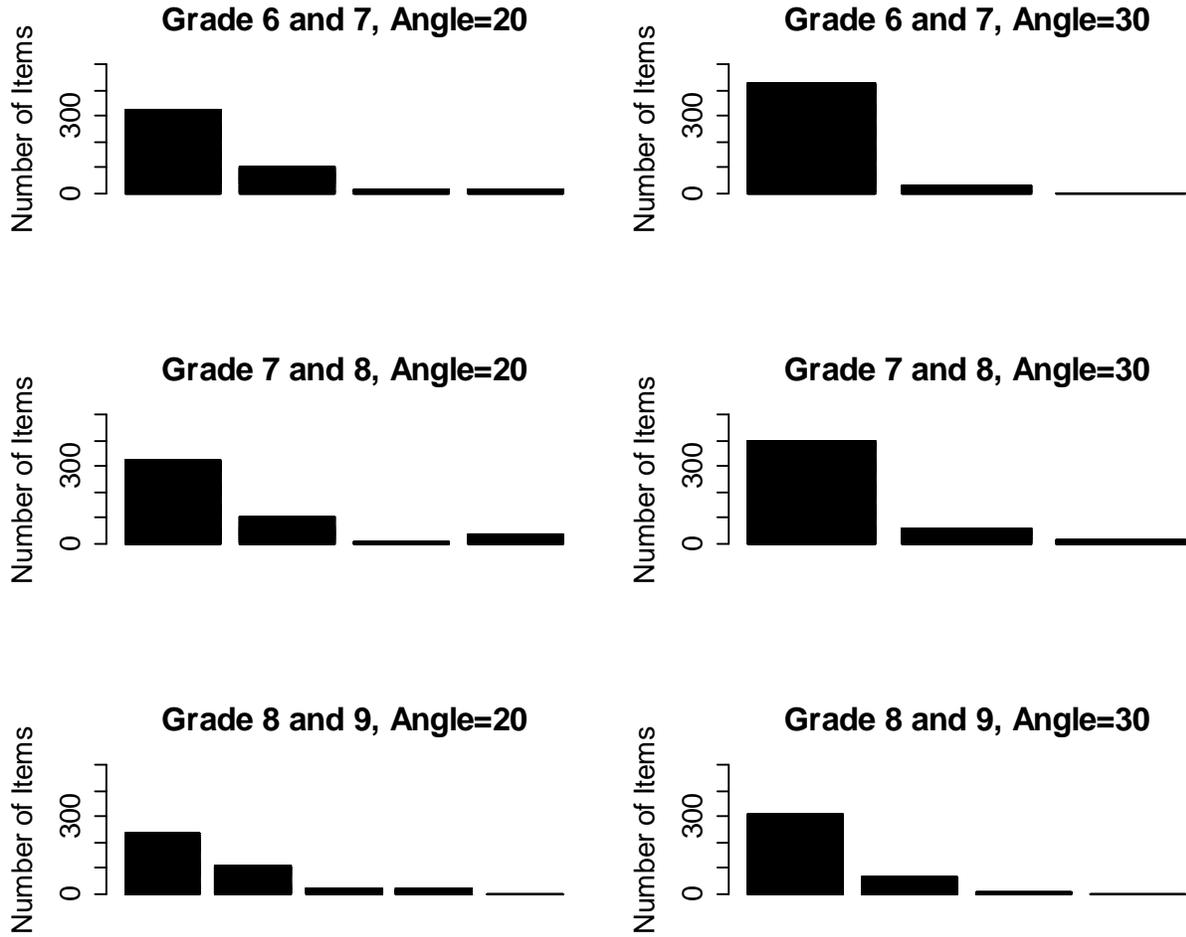


Figure 66. Clustering of Item Angle Measures for Grades 6 to 9, Mathematics (across grades)

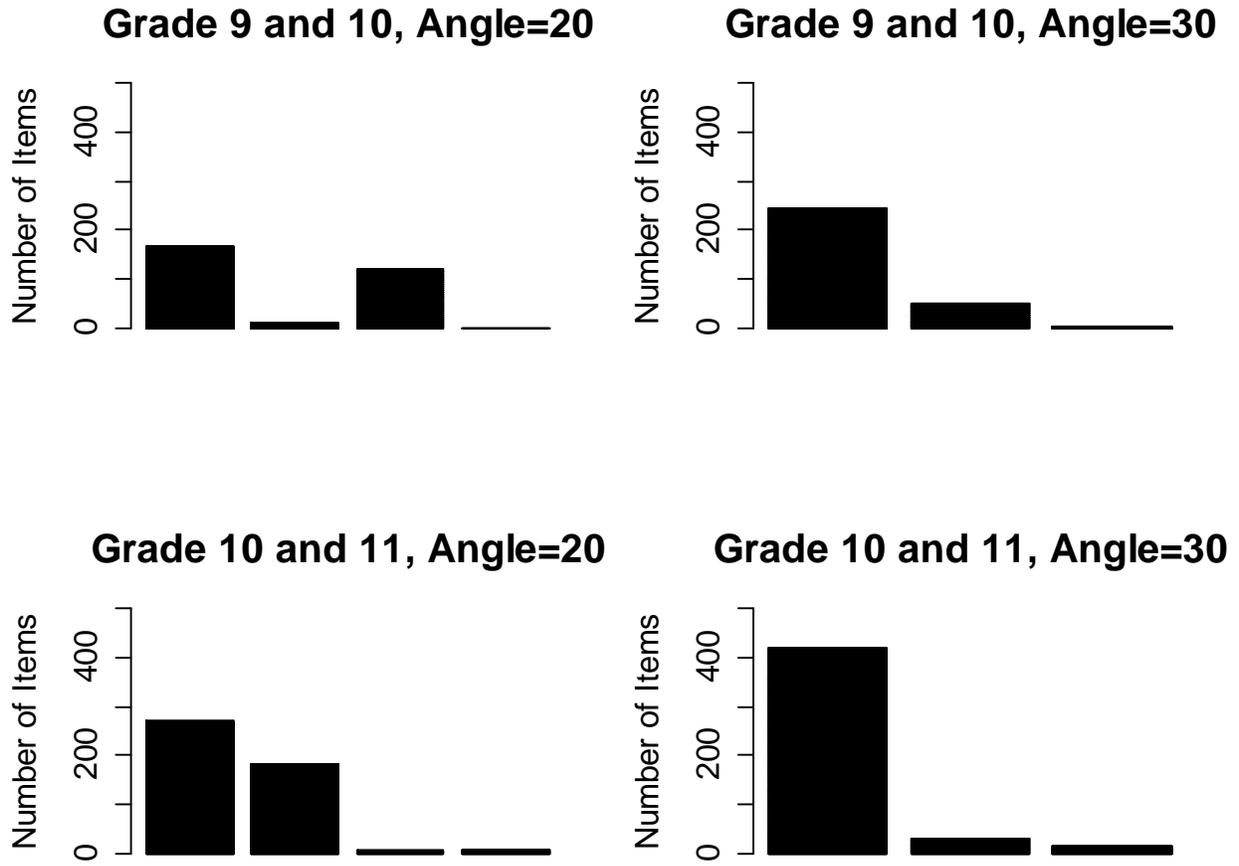


Figure 67. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (across grades)

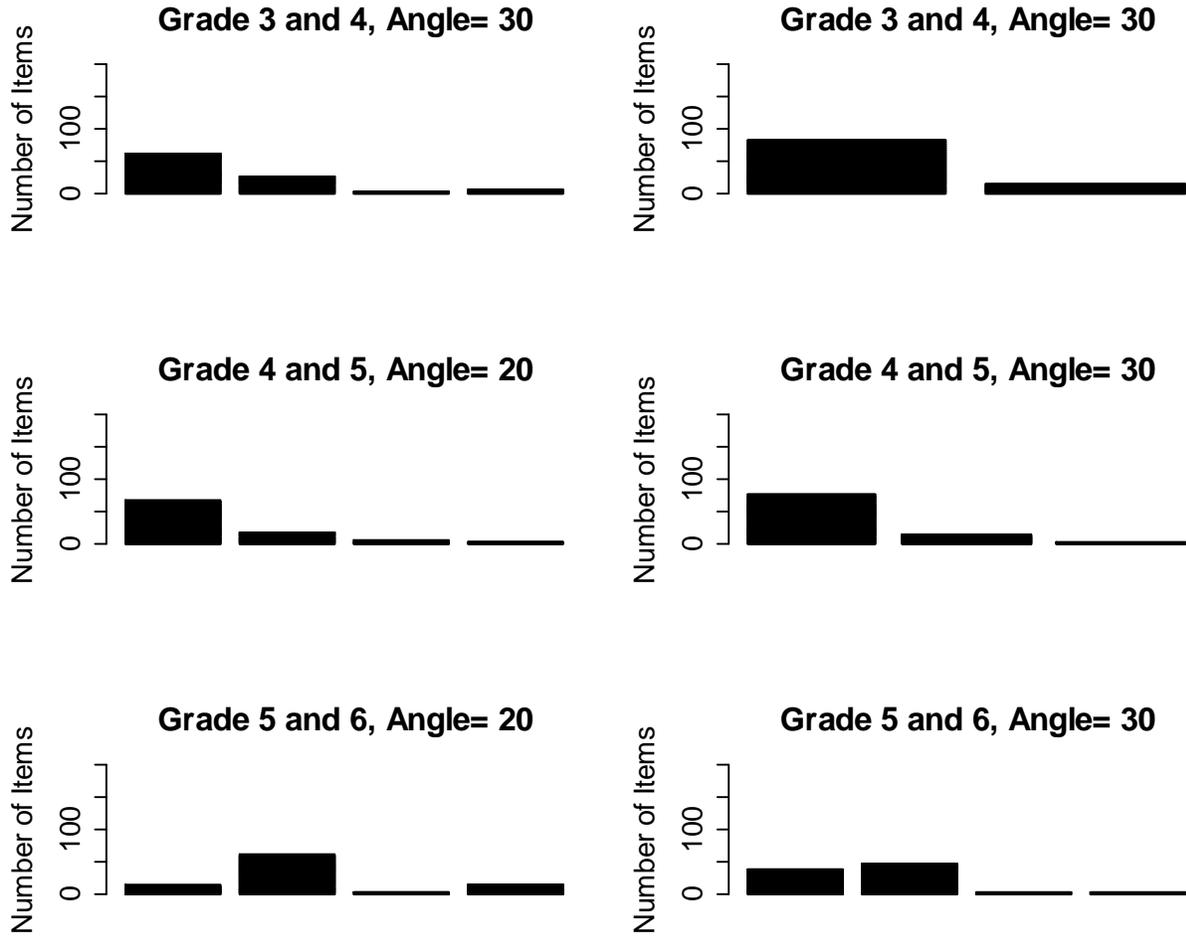


Figure 68. Clustering of Item Angle Measures for Grades 3 to 6, Mathematics (vertical linking)

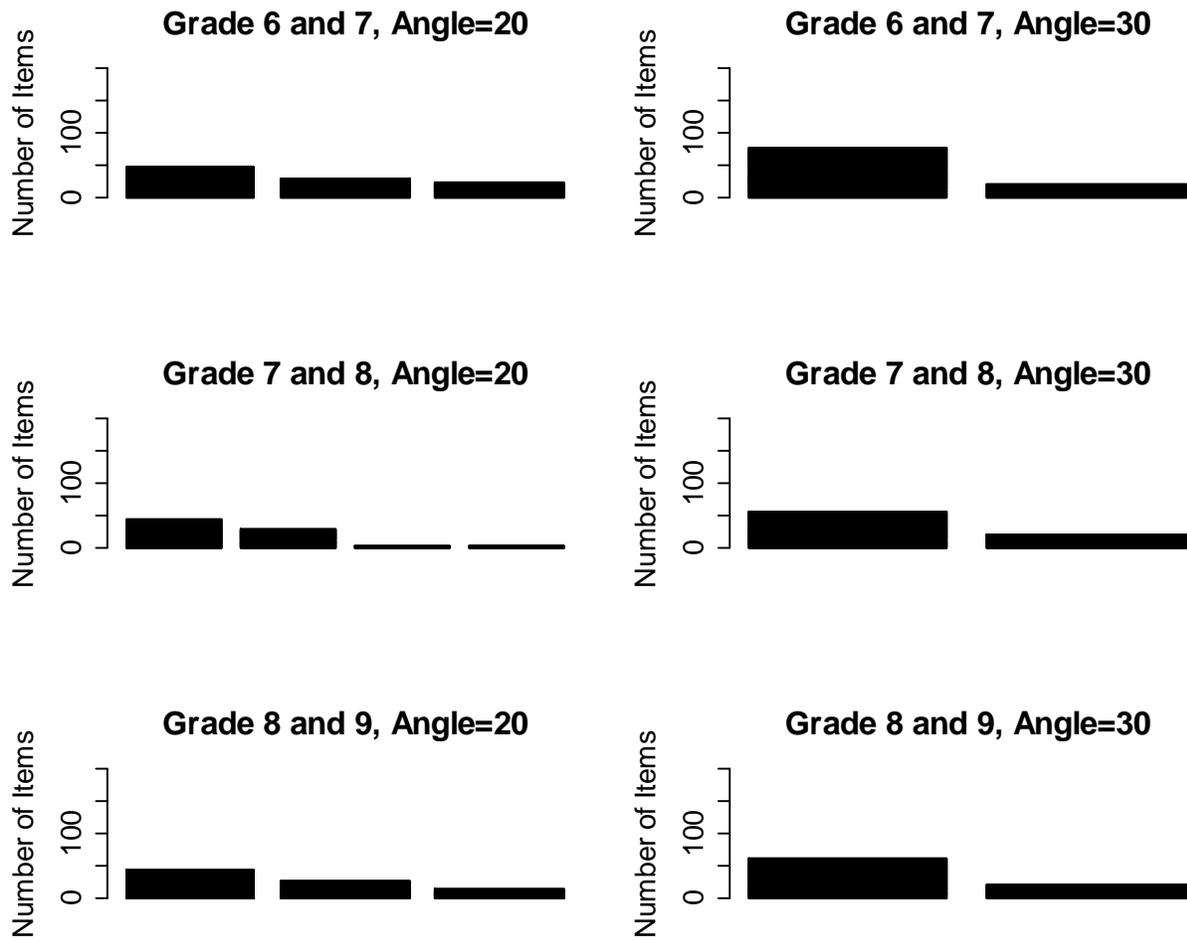


Figure 69. Clustering of Item Angle Measures for Grades 6 to 9, Mathematics (vertical linking)

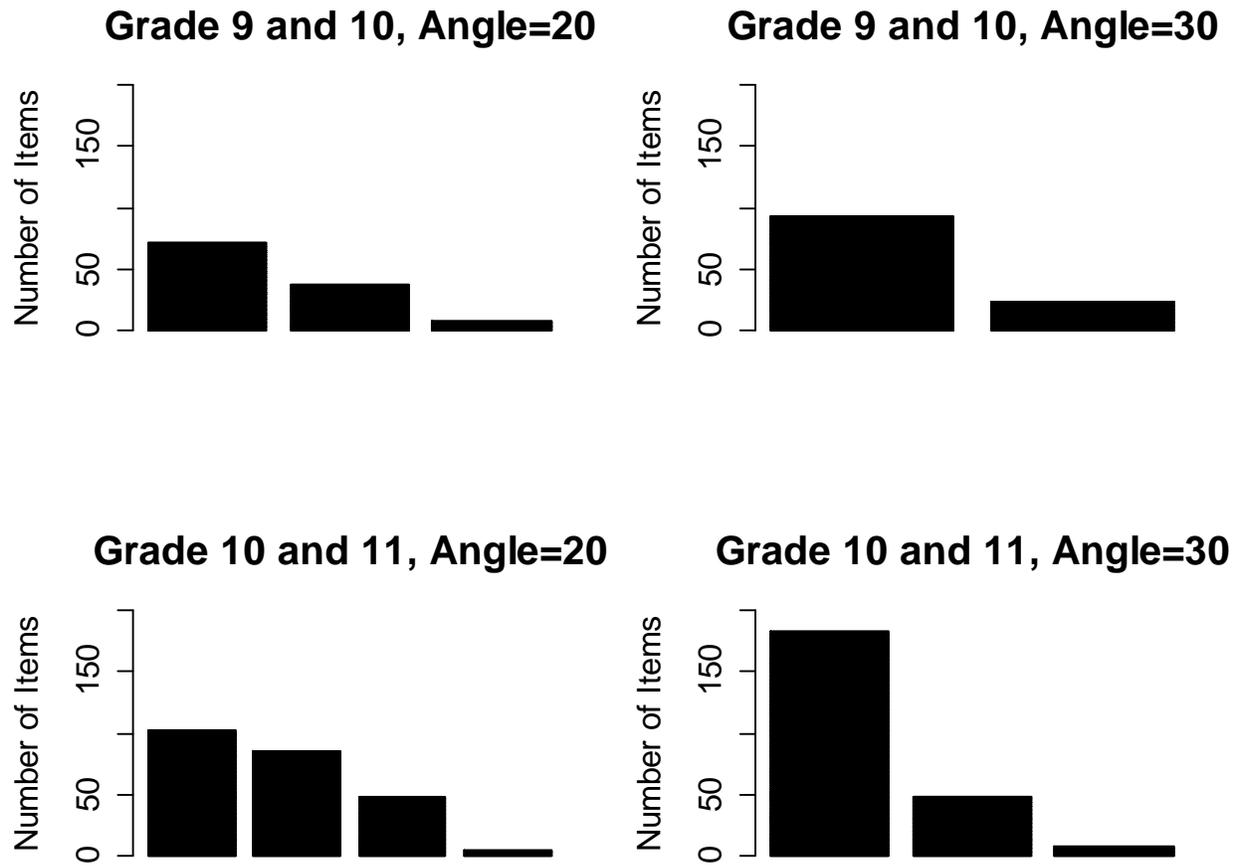


Figure 70. Clustering of Item Angle Measures for Grades 9 to 11, Mathematics (vertical linking)

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

$$P_i(\theta_j) = c_i + (1 - c_i) / \{1 + \exp[-Da_i(\theta_j - b_i)]\}$$

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 , 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 χ^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	χ^2 Mean	χ^2 SD	No. of Items	χ^2 Mean	χ^2 SD	No. of Items	χ^2 Mean	χ^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 χ^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	χ^2 Mean	χ^2 SD	No. of Items	χ^2 Mean	χ^2 SD	No. of Items	χ^2 Mean	χ^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	σ Estimate Summary				σ 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	<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	94	0.63	0.18	0.20	1.15	1.00	1.06	-1.06	3.38	-0.16	
	Score Categories	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	-0.12
		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
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			
10		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	
	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.¹ 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

¹The 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	<i>SD</i>	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

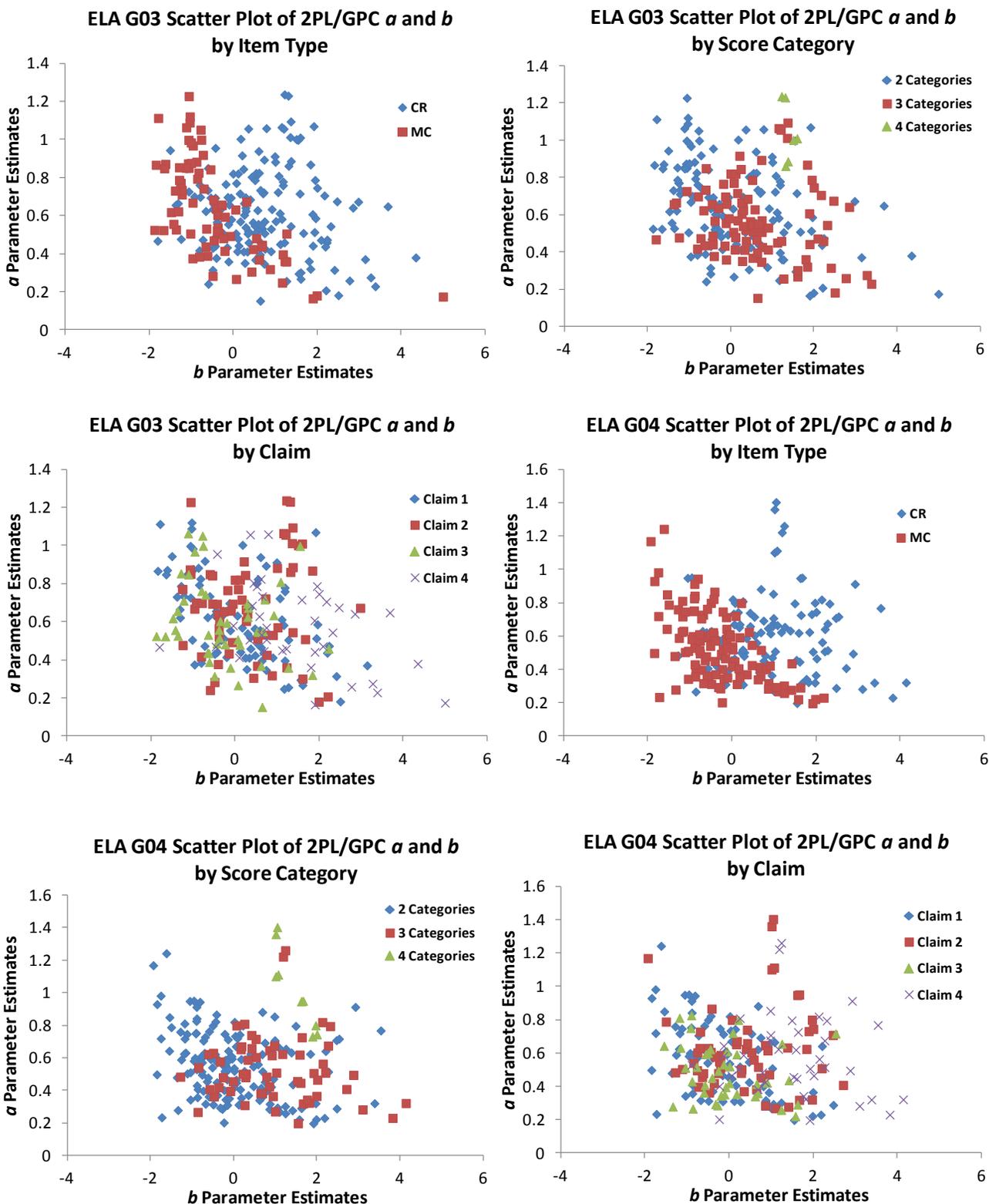
Based on the model comparison analysis results for the Pilot Test, the 2PL/GPC model could be adopted as the IRT model combination for calibrating Smarter Balanced items and establishing a vertical scale. The 2PL/GPC model provides flexibility for estimating a range of item discriminations without the 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:

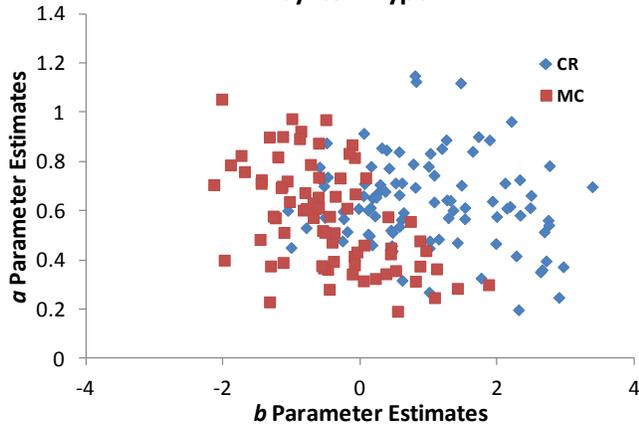
- There was a practical constraint that the scale will need to be established in the Field Test under very short timelines to support other activities such as standard setting. As a result, there will not be sufficient time to analyze the Field Test data under different IRT models. Therefore, it was 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.

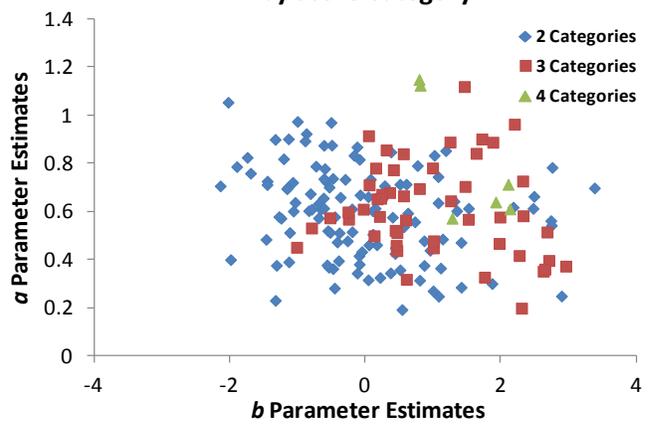
Figure 71. Scatter Plot of ELA/literacy 2PL/GPC Slope and Difficulty Estimates by Item Type, Score Category and Claim



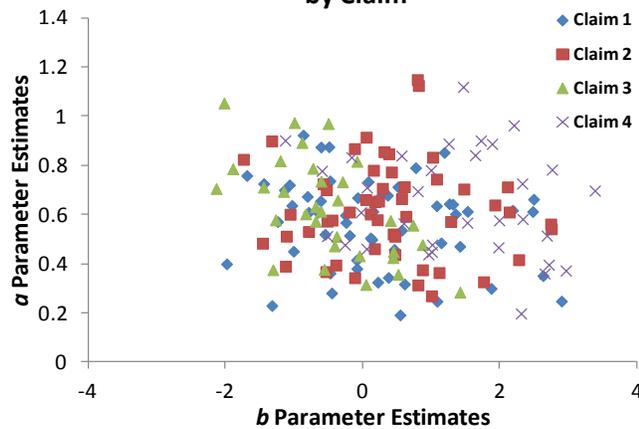
ELA G05 Scatter Plot of 2PL/GPC a and b by Item Type



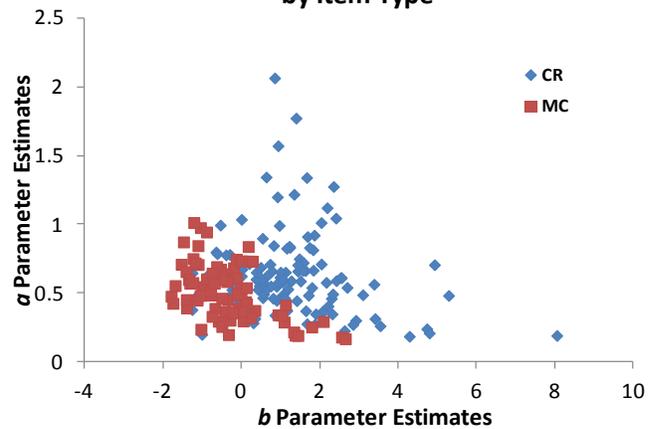
ELA G05 Scatter Plot of 2PL/GPC a and b by Score Category



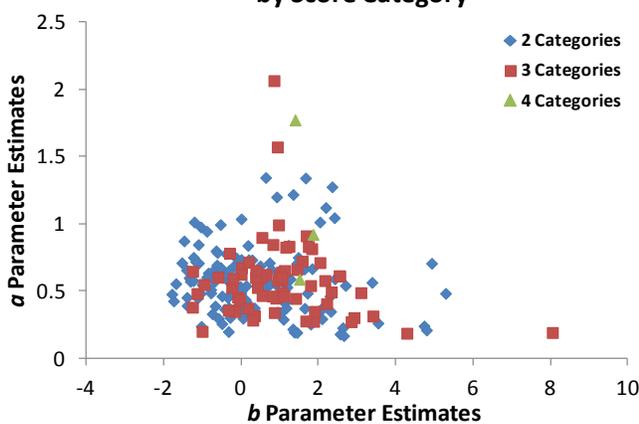
ELA G05 Scatter Plot of 2PL/GPC a and b by Claim



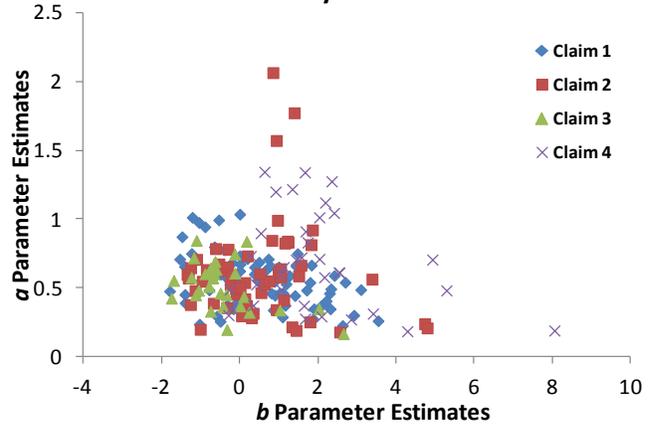
ELA G06 Scatter Plot of 2PL/GPC a and b by Item Type



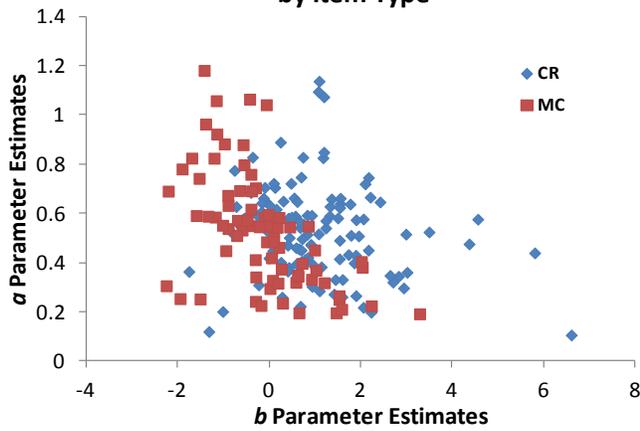
ELA G06 Scatter Plot of 2PL/GPC a and b by Score Category



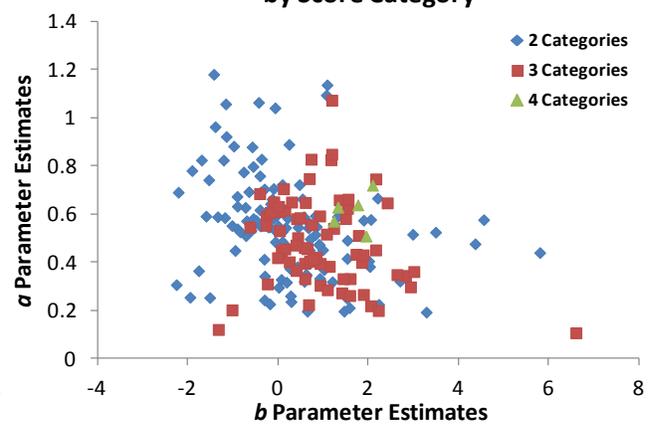
ELA G06 Scatter Plot of 2PL/GPC a and b by Claim



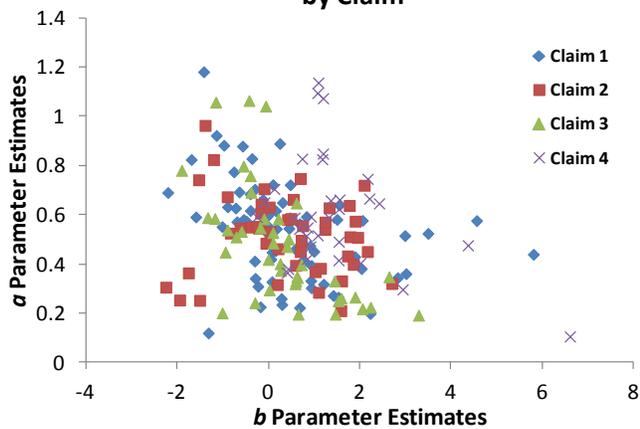
ELA G07 Scatter Plot of 2PL/GPC a and b by Item Type



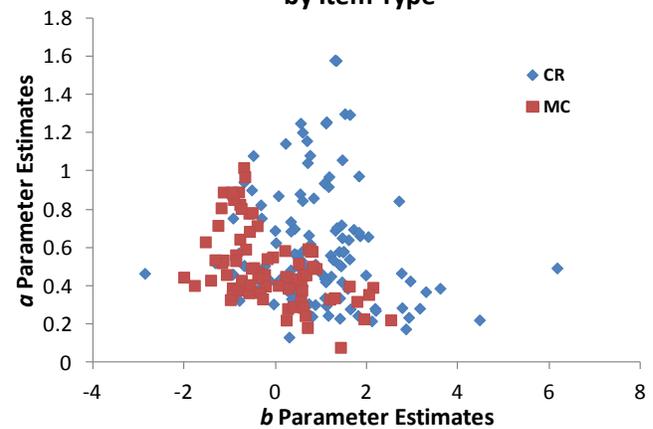
ELA G07 Scatter Plot of 2PL/GPC a and b by Score Category



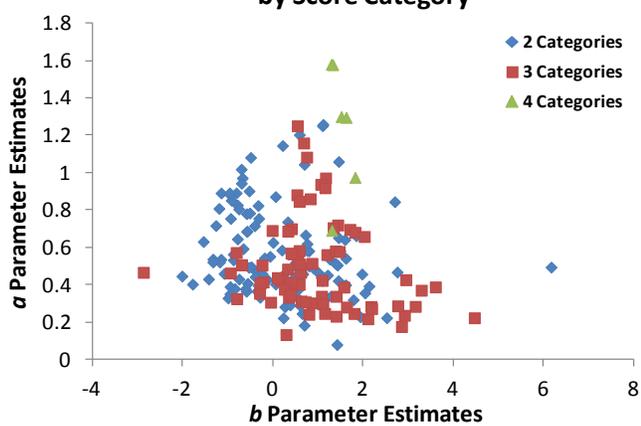
ELA G07 Scatter Plot of 2PL/GPC a and b by Claim



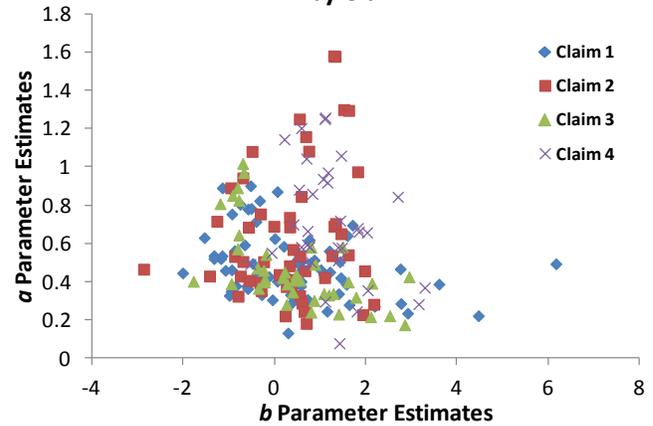
ELA G08 Scatter Plot of 2PL/GPC a and b by Item Type



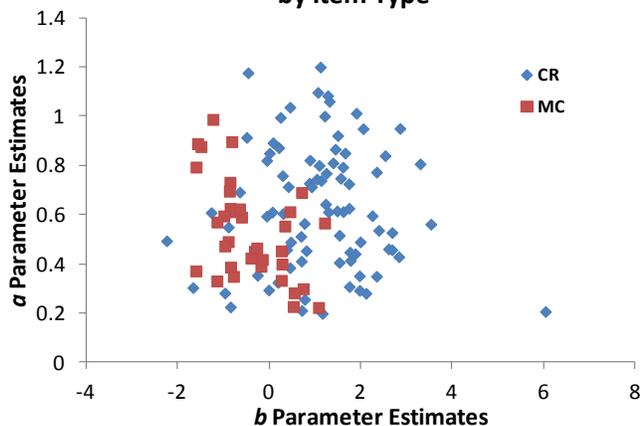
ELA G08 Scatter Plot of 2PL/GPC a and b by Score Category



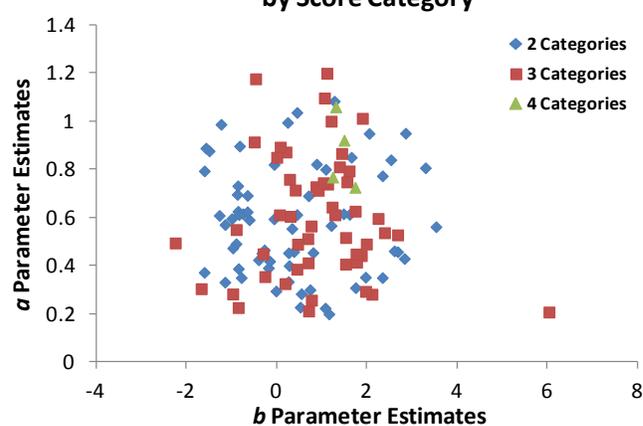
ELA G08 Scatter Plot of 2PL/GPC a and b by Claim



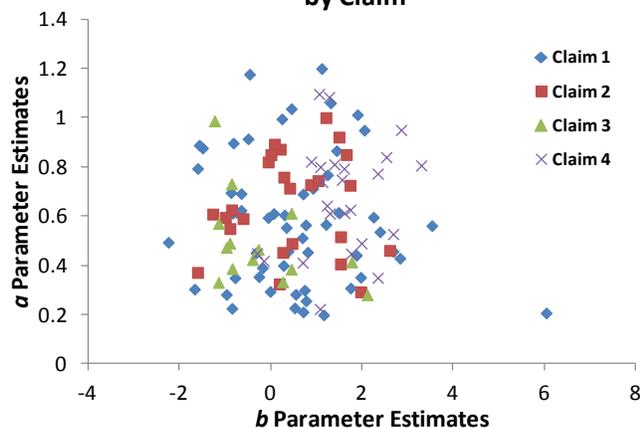
ELA G09 Scatter Plot of 2PL/GPC a and b by Item Type



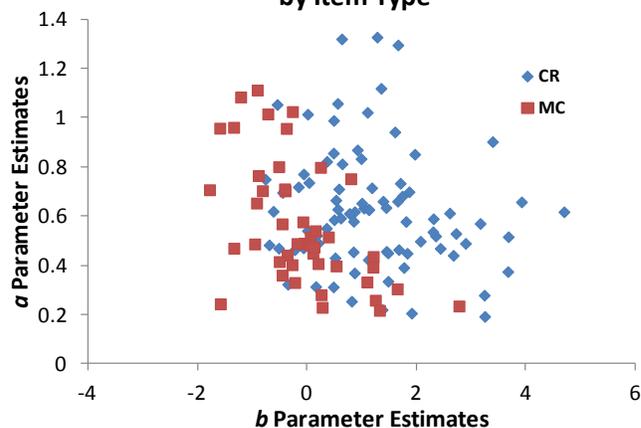
ELA G09 Scatter Plot of 2PL/GPC a and b by Score Category



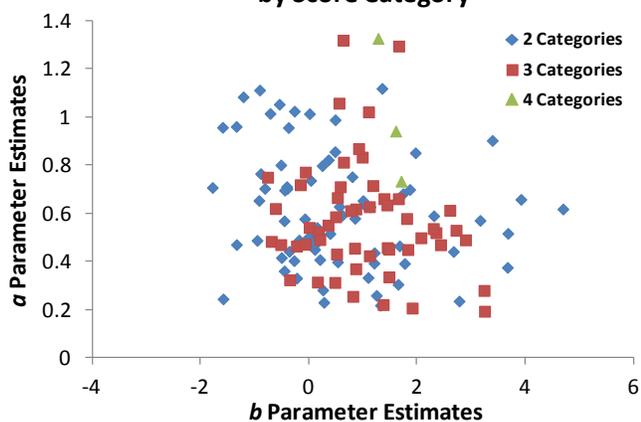
ELA G09 Scatter Plot of 2PL/GPC a and b by Claim



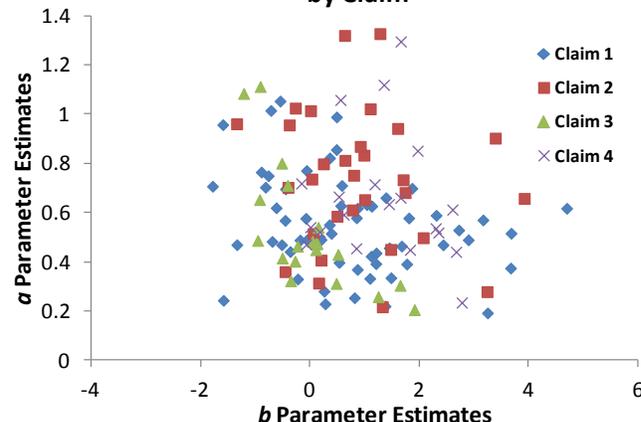
ELA G10 Scatter Plot of 2PL/GPC a and b by Item Type



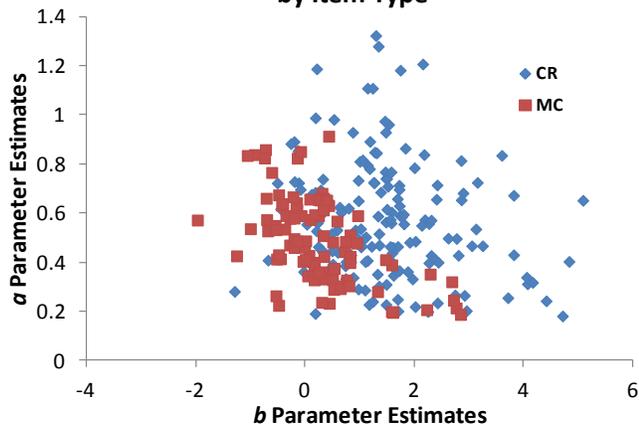
ELA G10 Scatter Plot of 2PL/GPC a and b by Score Category



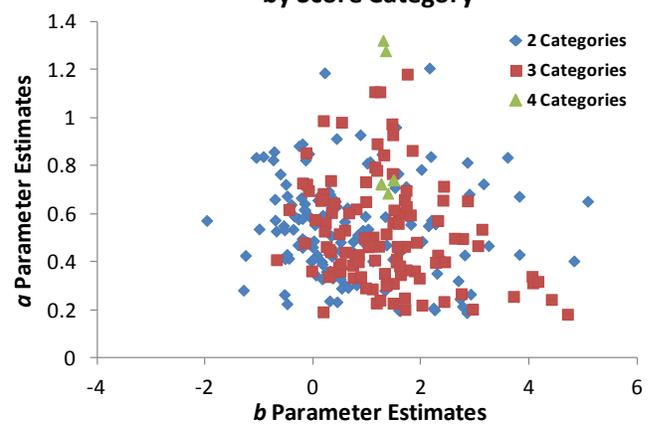
ELA G10 Scatter Plot of 2PL/GPC a and b by Claim



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

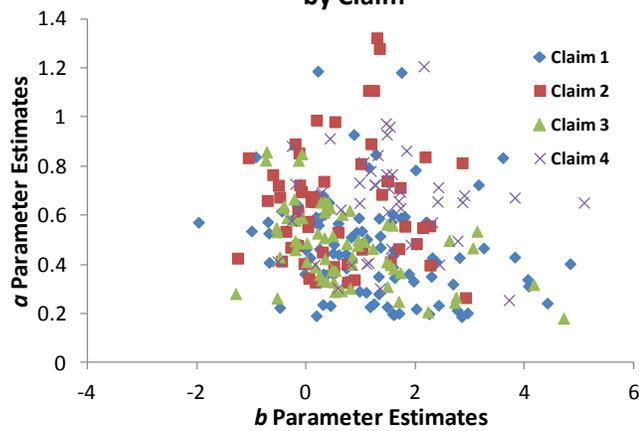
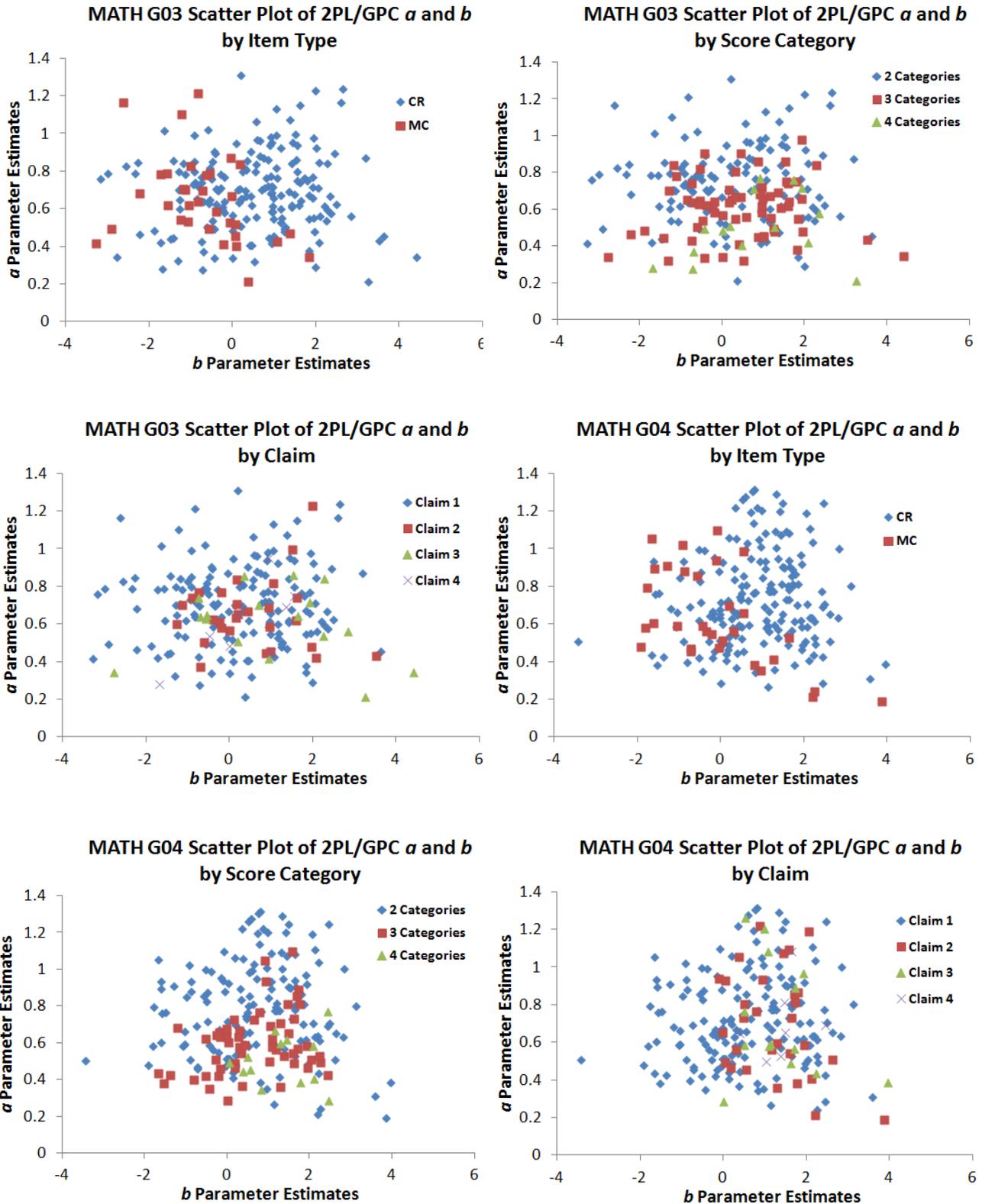
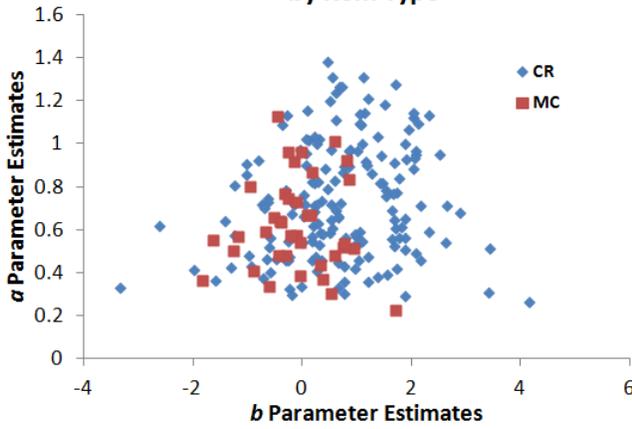


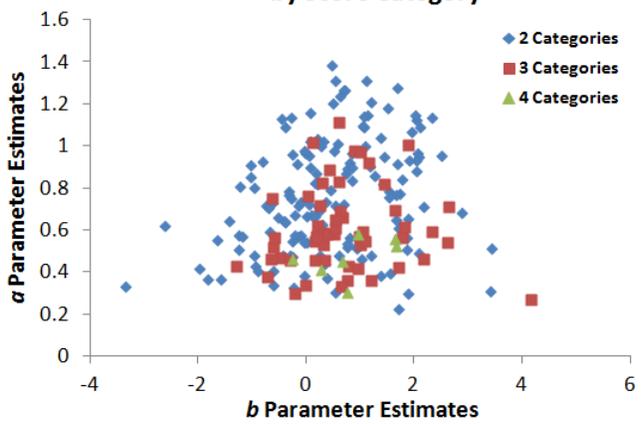
Figure 72. Scatter Plot of Mathematics 2PL/GPC Slope and Difficulty Estimates by Item Type, Score Category, and Claim



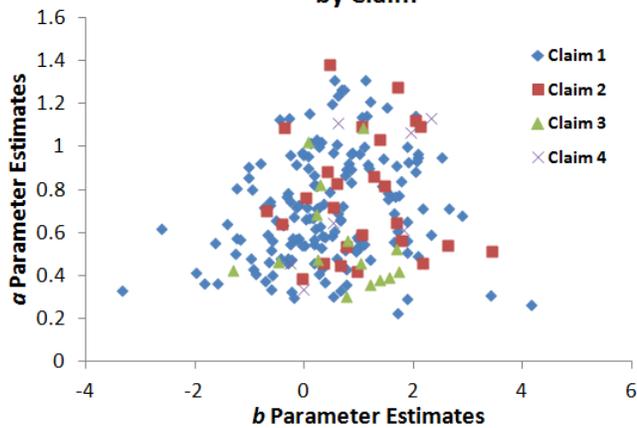
MATH G05 Scatter Plot of 2PL/GPC a and b by Item Type



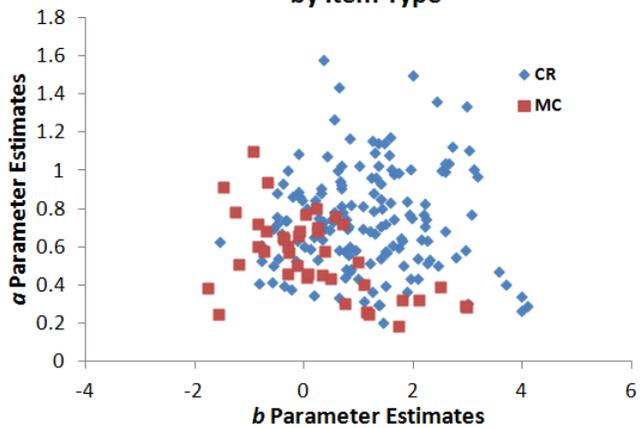
MATH G05 Scatter Plot of 2PL/GPC a and b by Score Category



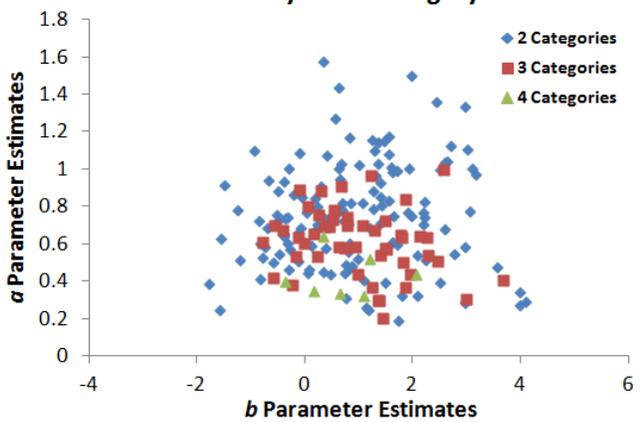
MATH G05 Scatter Plot of 2PL/GPC a and b by Claim



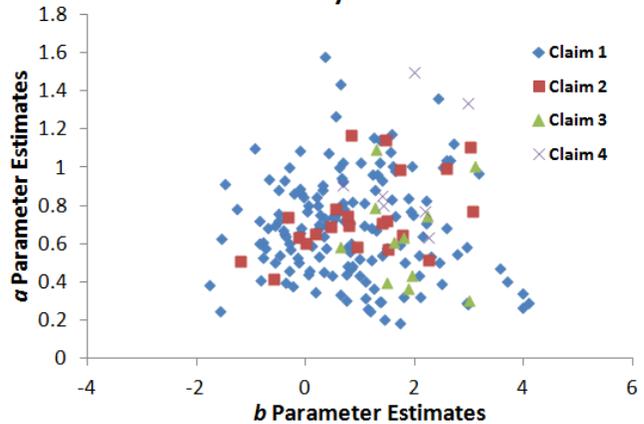
MATH G06 Scatter Plot of 2PL/GPC a and b by Item Type



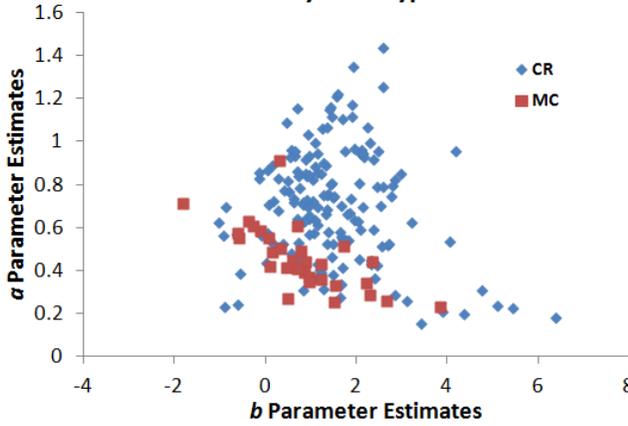
MATH G06 Scatter Plot of 2PL/GPC a and b by Score Category



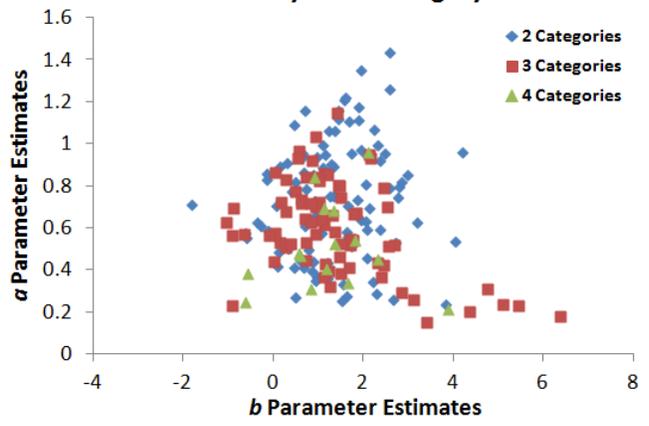
MATH G06 Scatter Plot of 2PL/GPC a and b by Claim



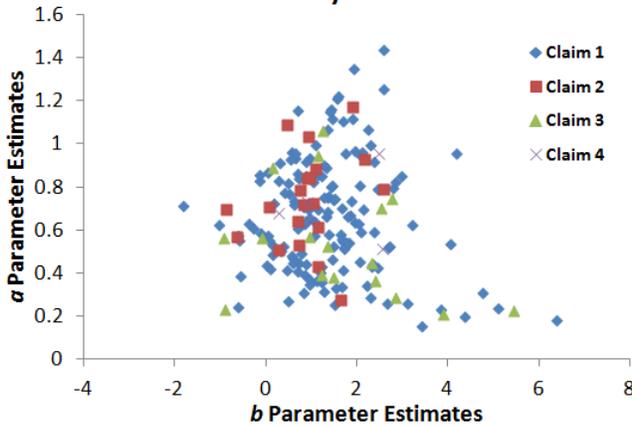
MATH G07 Scatter Plot of 2PL/GPC a and b by Item Type



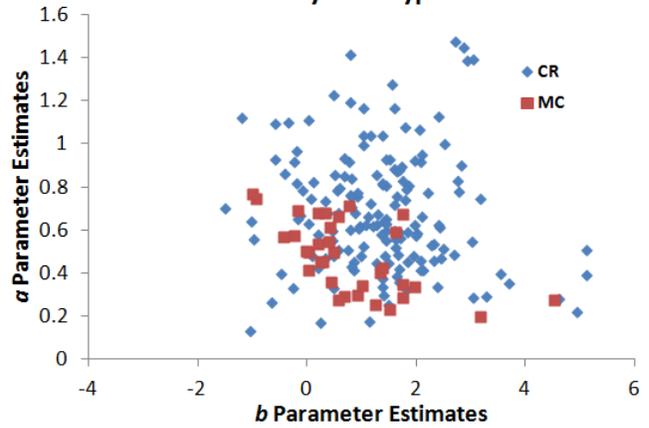
MATH G07 Scatter Plot of 2PL/GPC a and b by Score Category



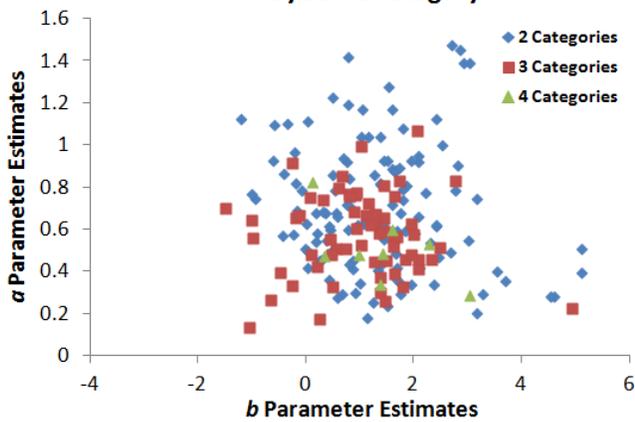
MATH G07 Scatter Plot of 2PL/GPC a and b by Claim



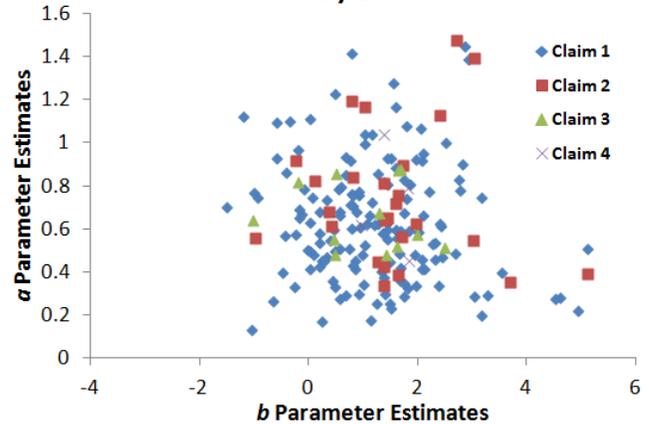
MATH G08 Scatter Plot of 2PL/GPC a and b by Item Type



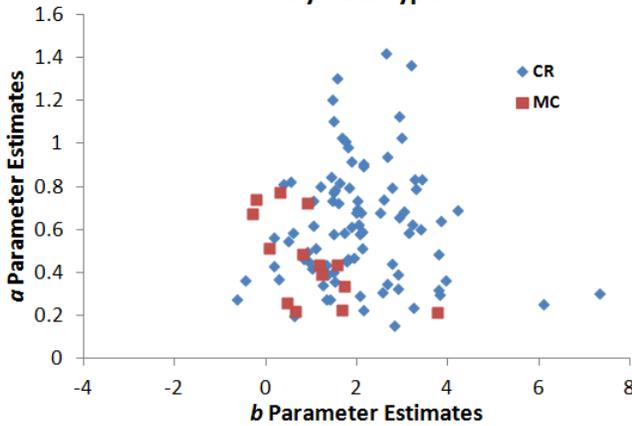
MATH G08 Scatter Plot of 2PL/GPC a and b by Score Category



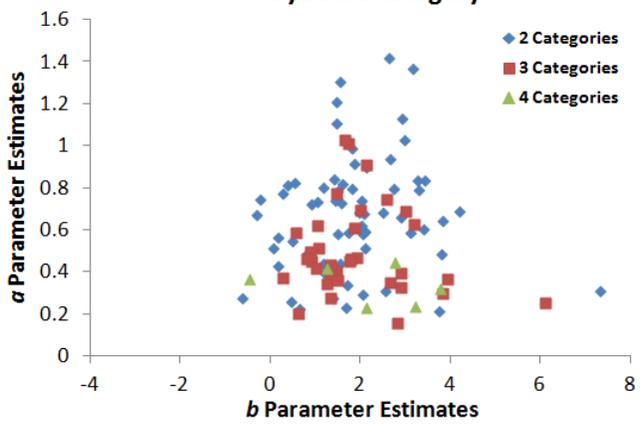
MATH G08 Scatter Plot of 2PL/GPC a and b by Claim



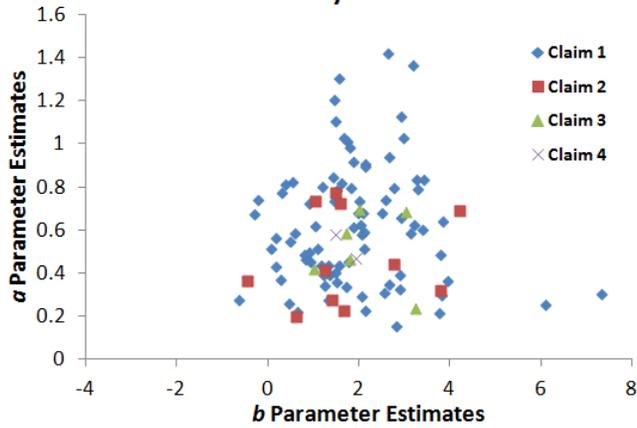
MATH G09 Scatter Plot of 2PL/GPC a and b by Item Type



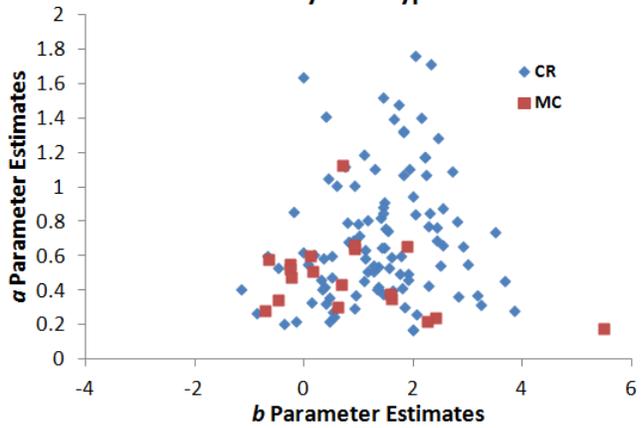
MATH G09 Scatter Plot of 2PL/GPC a and b by Score Category



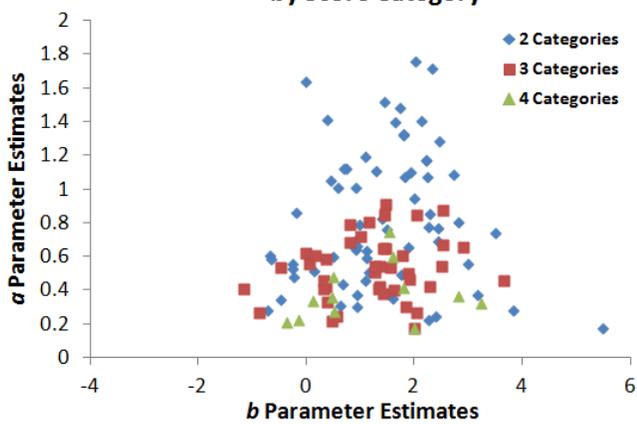
MATH G09 Scatter Plot of 2PL/GPC a and b by Claim



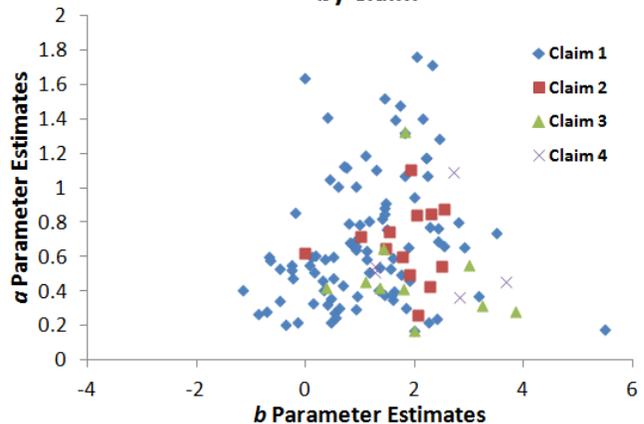
MATH G10 Scatter Plot of 2PL/GPC a and b by Item Type



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



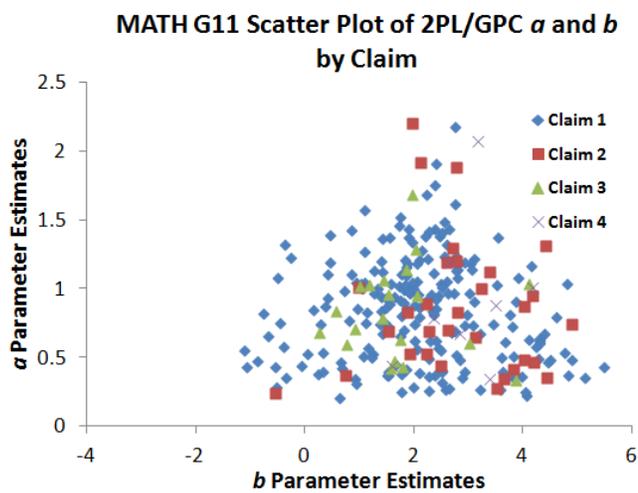
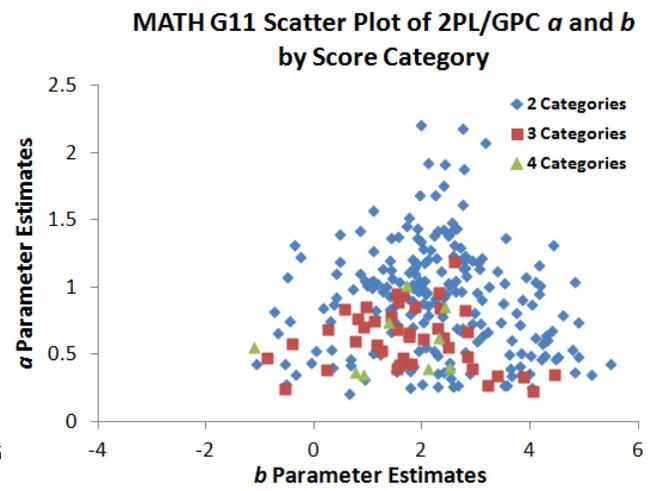
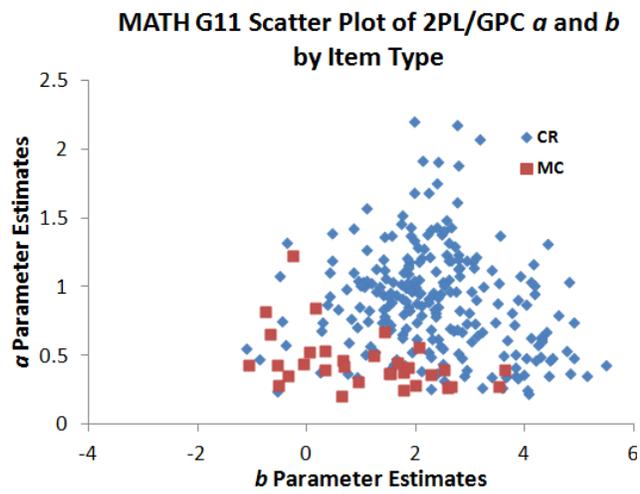
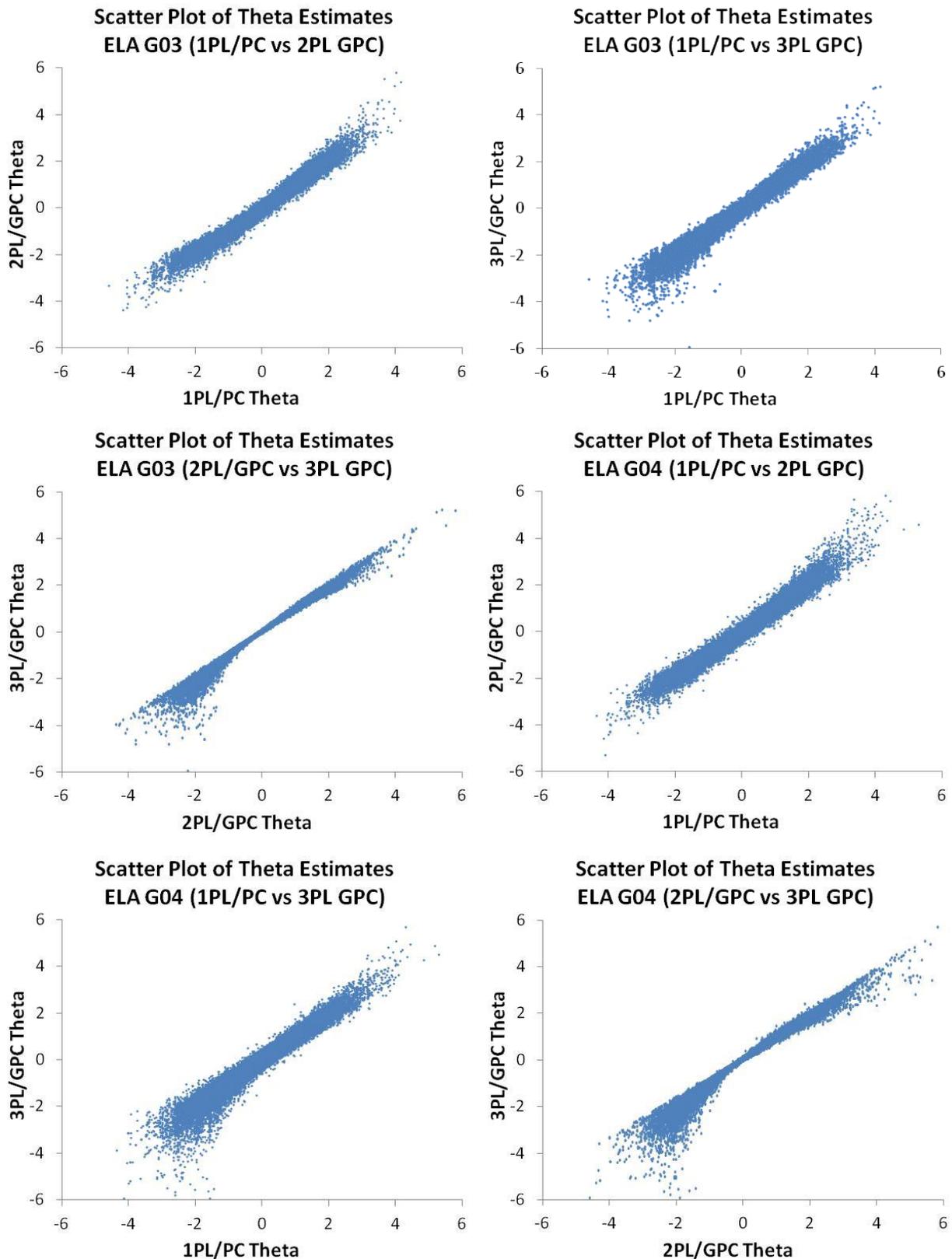
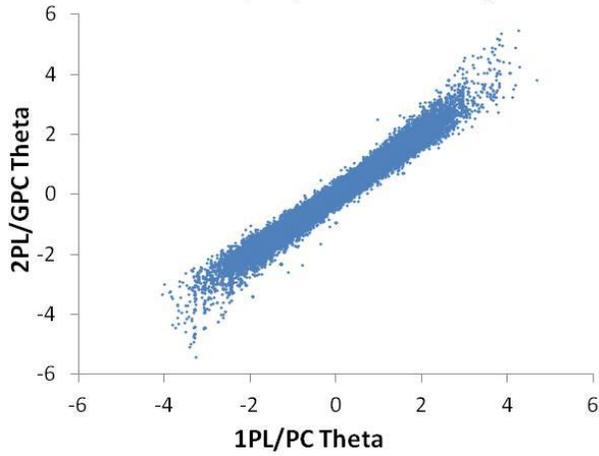


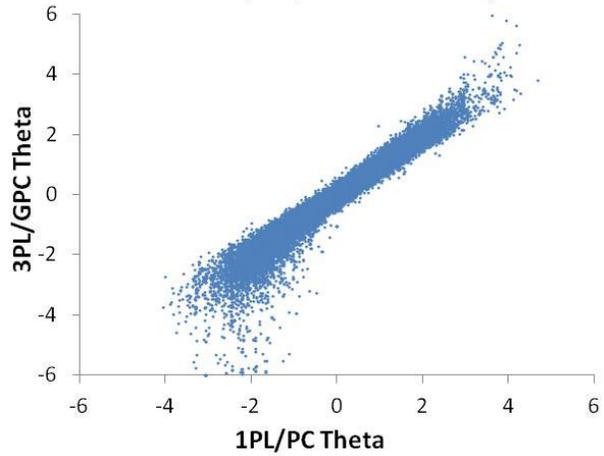
Figure 73. ELA/literacy Scatter Plots of Theta Estimates across Different Model Combinations



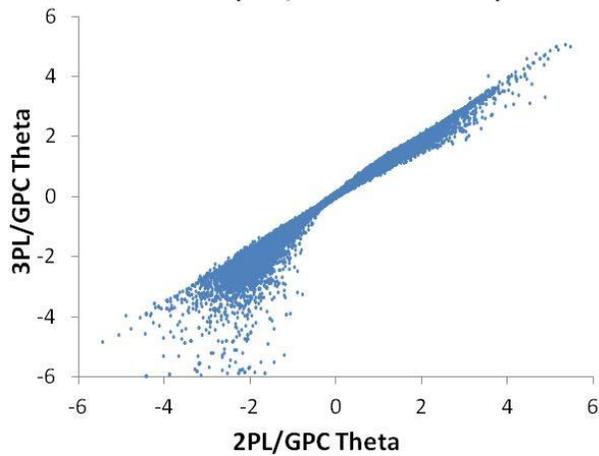
**Scatter Plot of Theta Estimates
ELA G05 (1PL/PC vs 2PL GPC)**



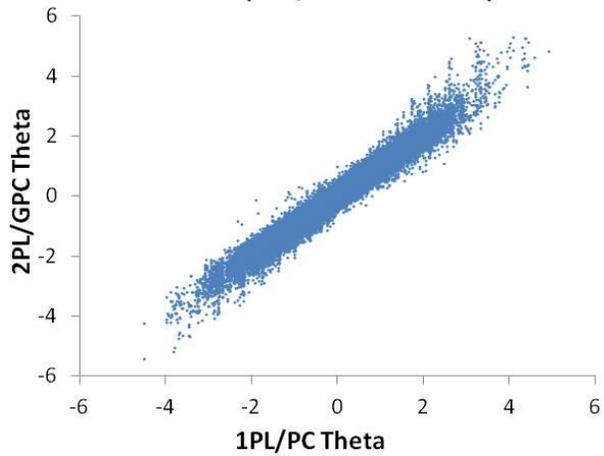
**Scatter Plot of Theta Estimates
ELA G05 (1PL/PC vs 3PL GPC)**



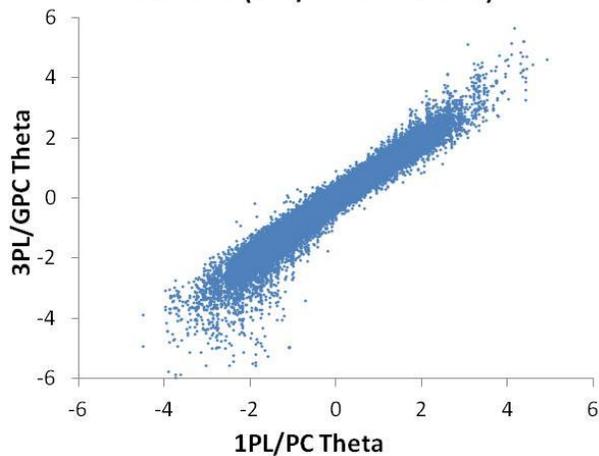
**Scatter Plot of Theta Estimates
ELA G05 (2PL/GPC vs 3PL GPC)**



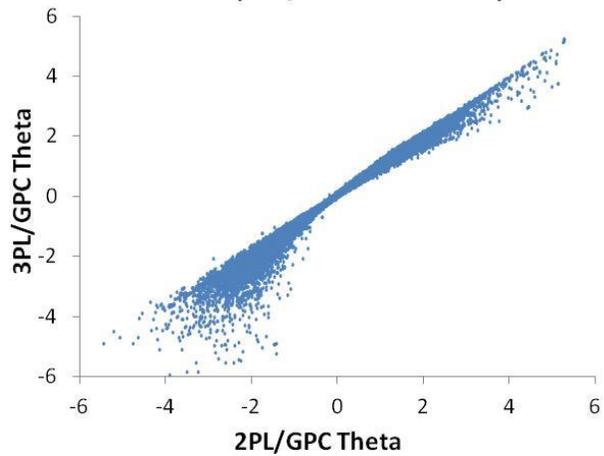
**Scatter Plot of Theta Estimates
ELA G06 (1PL/PC vs 2PL GPC)**



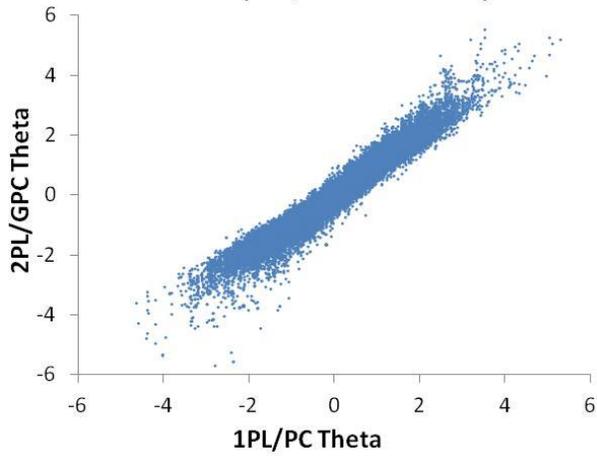
**Scatter Plot of Theta Estimates
ELA G06 (1PL/PC vs 3PL GPC)**



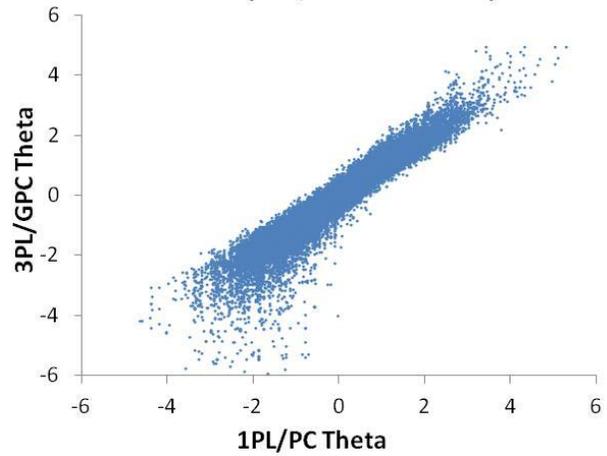
**Scatter Plot of Theta Estimates
ELA G06 (2PL/GPC vs 3PL GPC)**



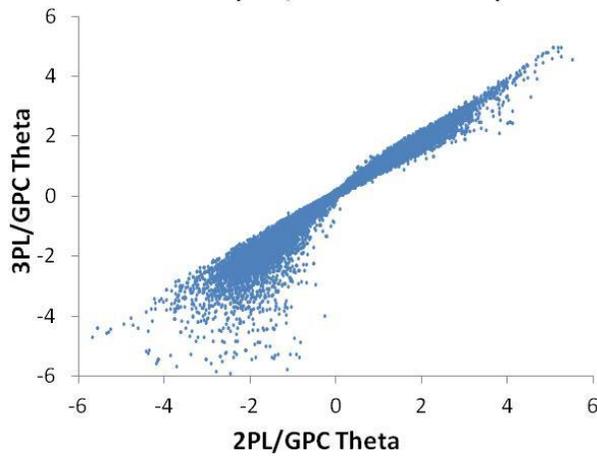
**Scatter Plot of Theta Estimates
ELA G07 (1PL/PC vs 2PL GPC)**



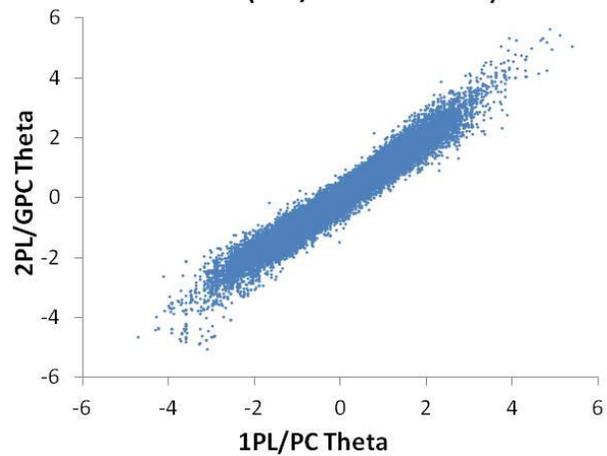
**Scatter Plot of Theta Estimates
ELA G07 (1PL/PC vs 3PL GPC)**



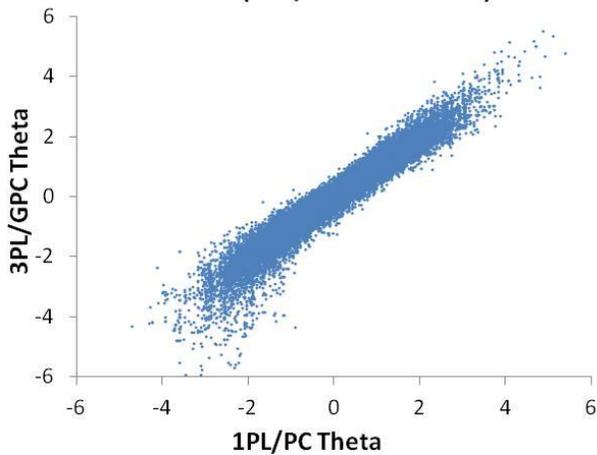
**Scatter Plot of Theta Estimates
ELA G07 (2PL/GPC vs 3PL GPC)**



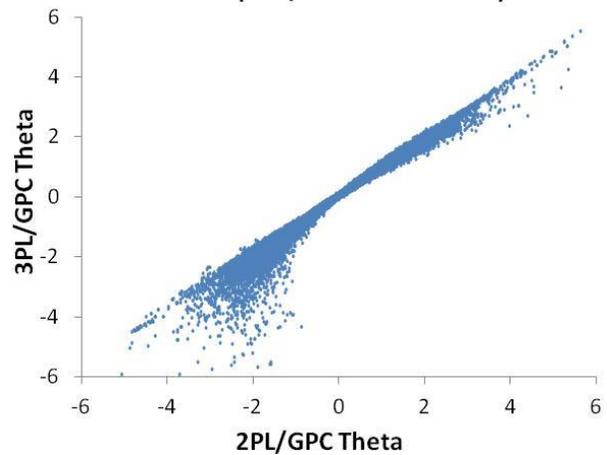
**Scatter Plot of Theta Estimates
ELA G08 (1PL/PC vs 2PL GPC)**



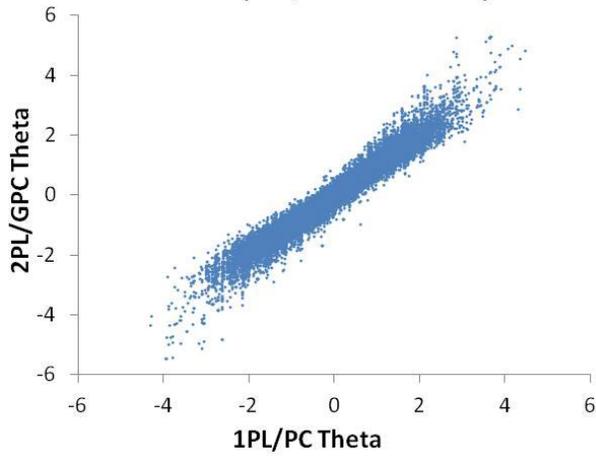
**Scatter Plot of Theta Estimates
ELA G08 (1PL/PC vs 3PL GPC)**



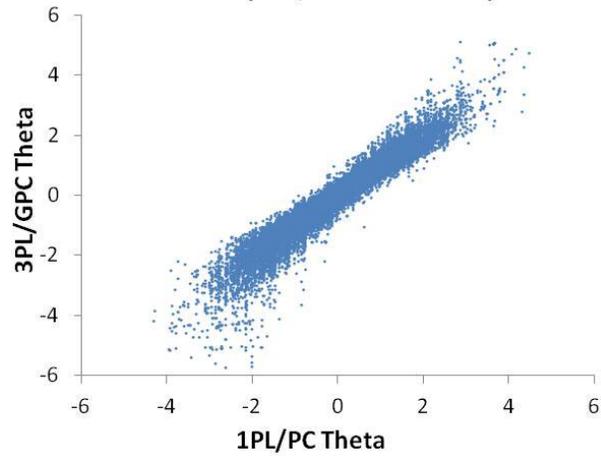
**Scatter Plot of Theta Estimates
ELA G08 (2PL/GPC vs 3PL GPC)**



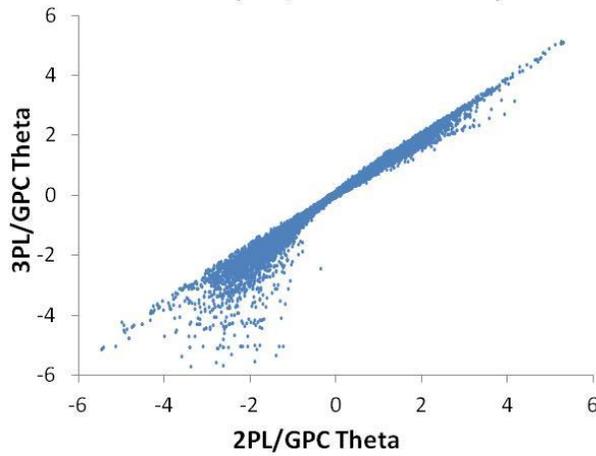
**Scatter Plot of Theta Estimates
ELA G09 (1PL/PC vs 2PL GPC)**



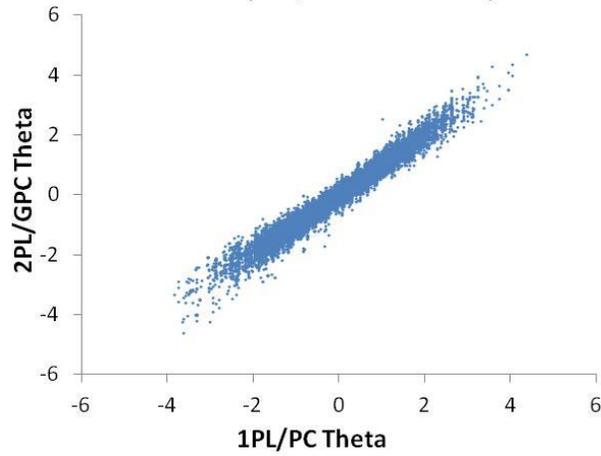
**Scatter Plot of Theta Estimates
ELA G09 (1PL/PC vs 3PL GPC)**



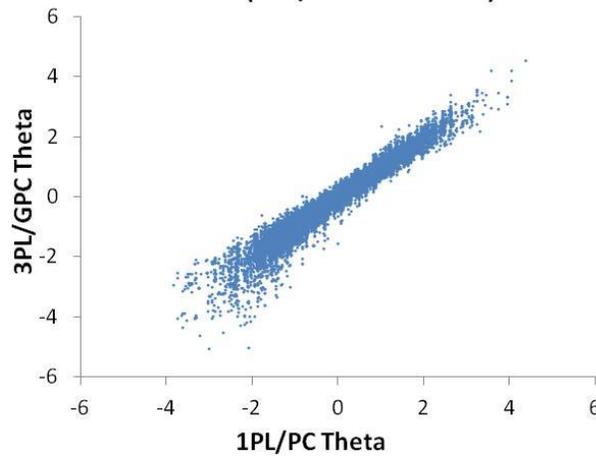
**Scatter Plot of Theta Estimates
ELA G09 (2PL/GPC vs 3PL GPC)**



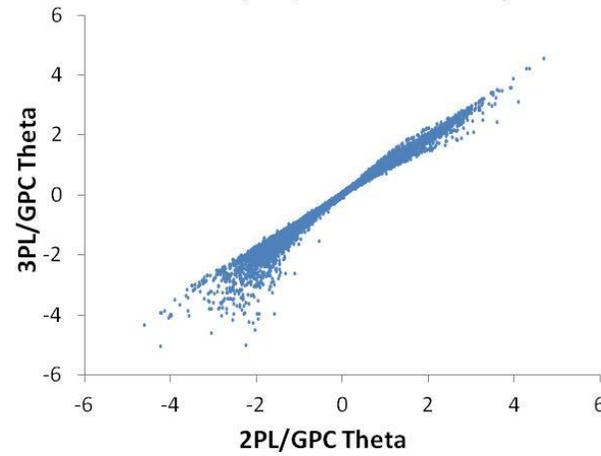
**Scatter Plot of Theta Estimates
ELA G10 (1PL/PC vs 2PL GPC)**



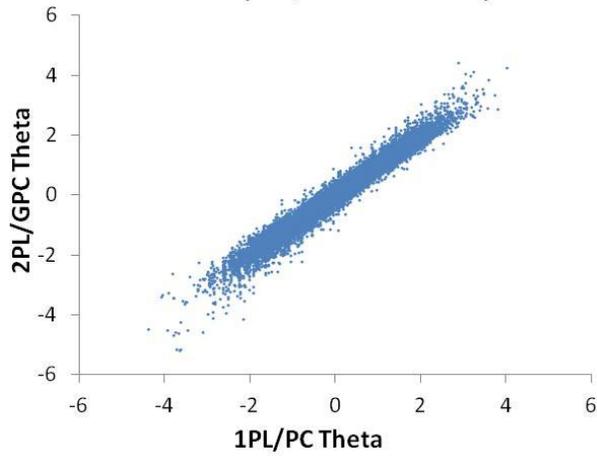
**Scatter Plot of Theta Estimates
ELA G10 (1PL/PC vs 3PL GPC)**



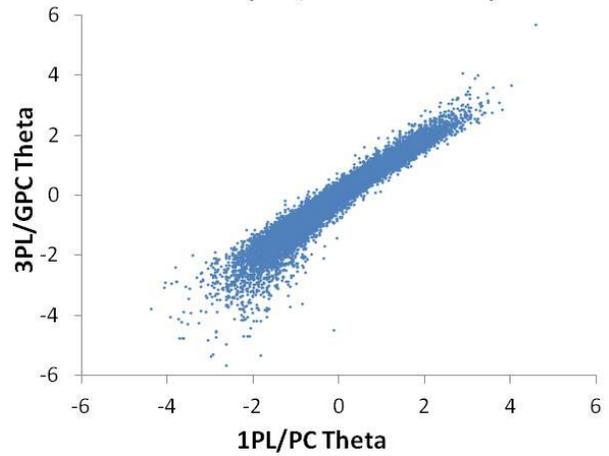
**Scatter Plot of Theta Estimates
ELA G10 (2PL/GPC vs 3PL GPC)**



**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)**

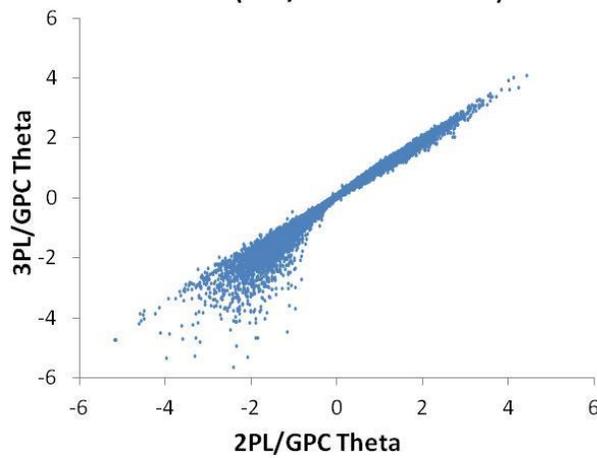
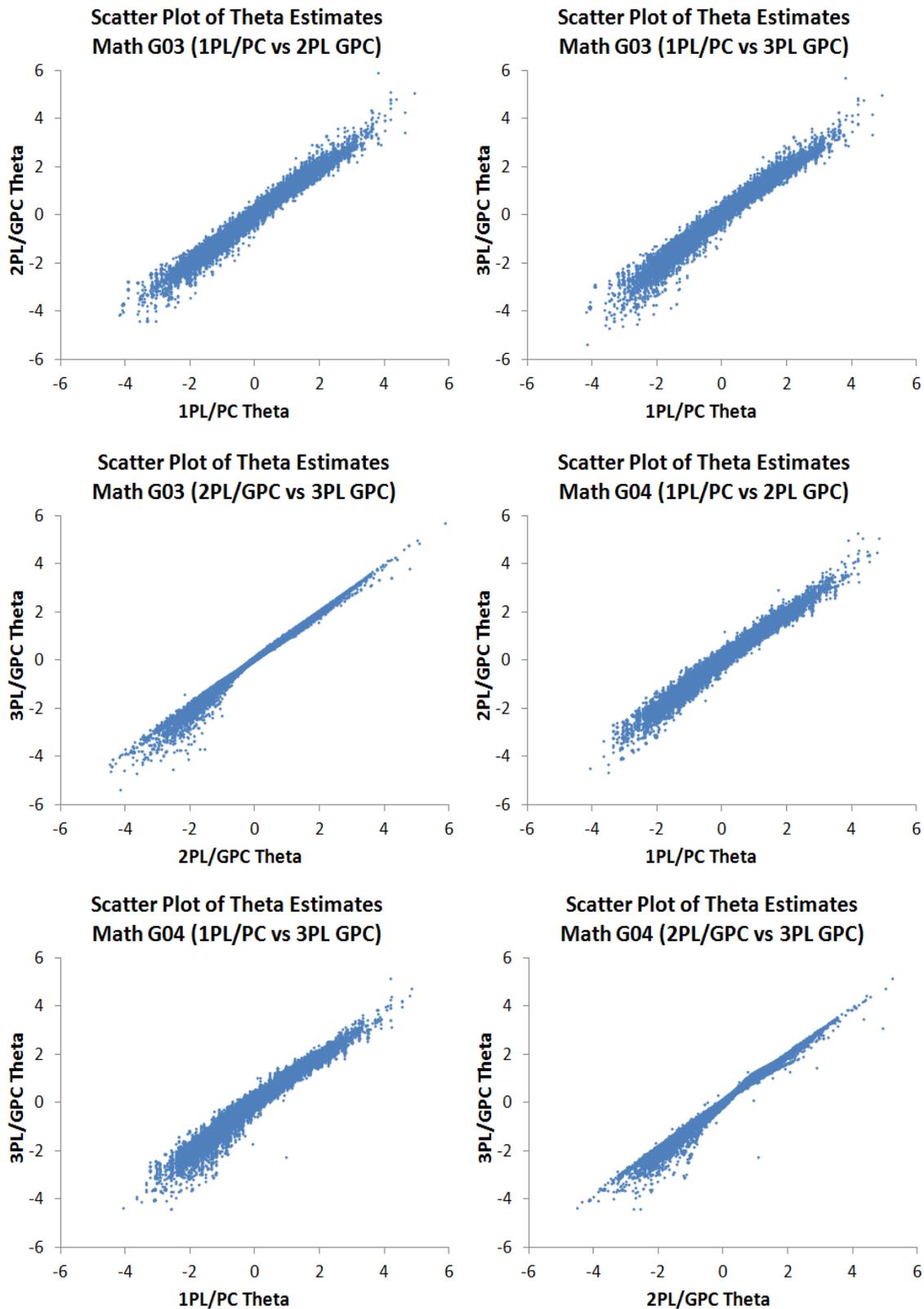
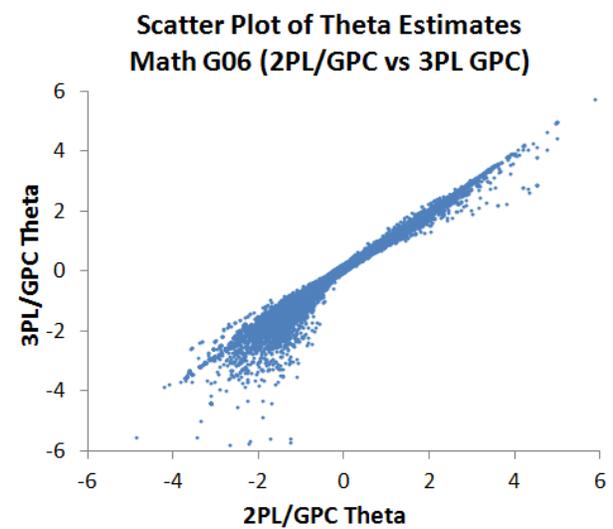
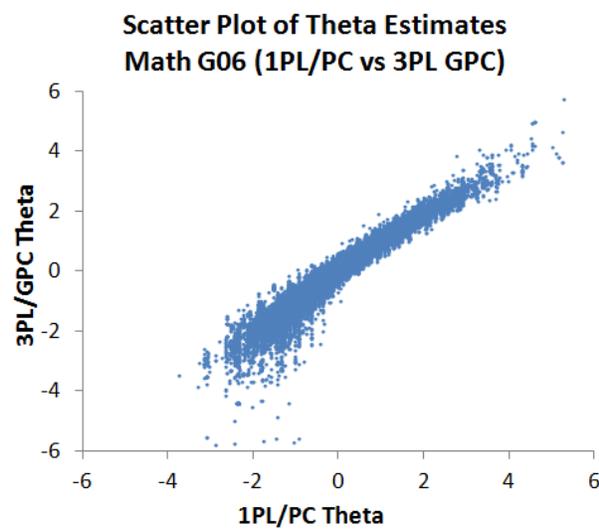
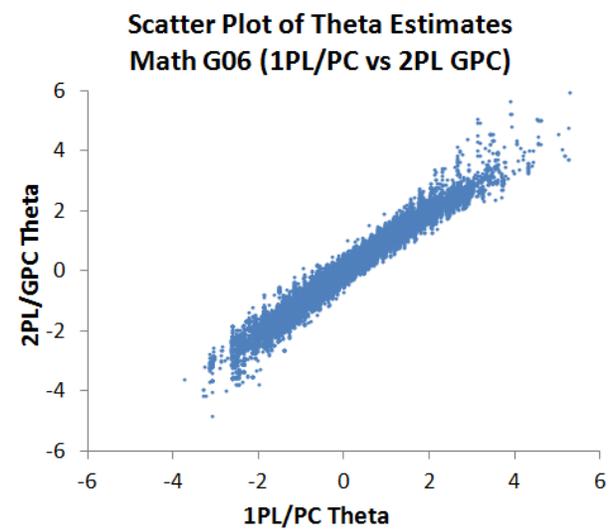
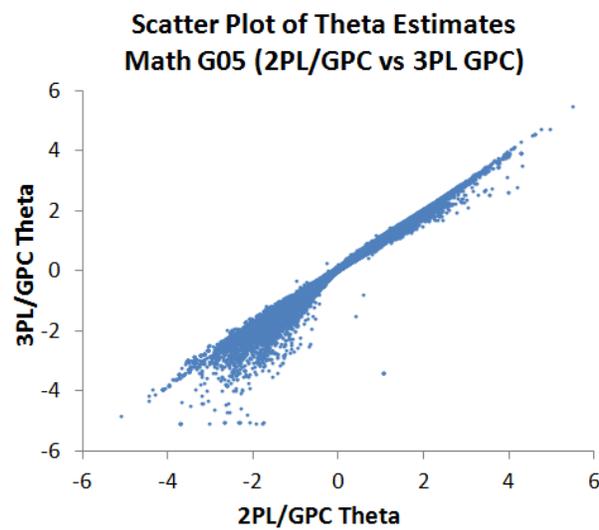
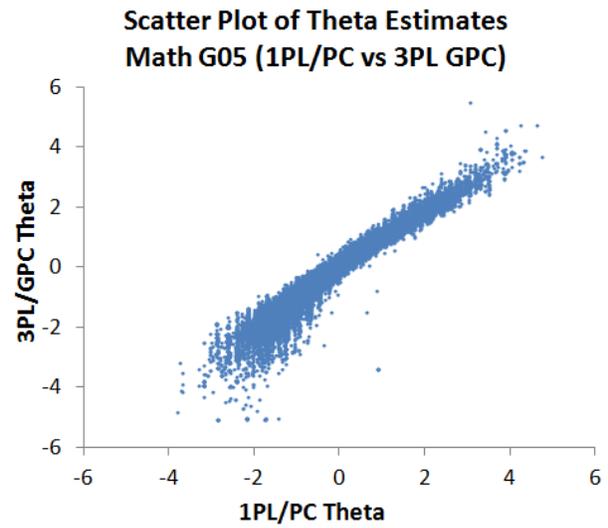
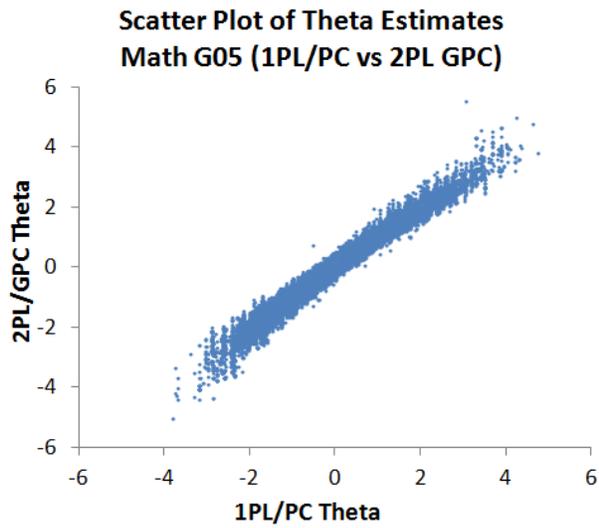
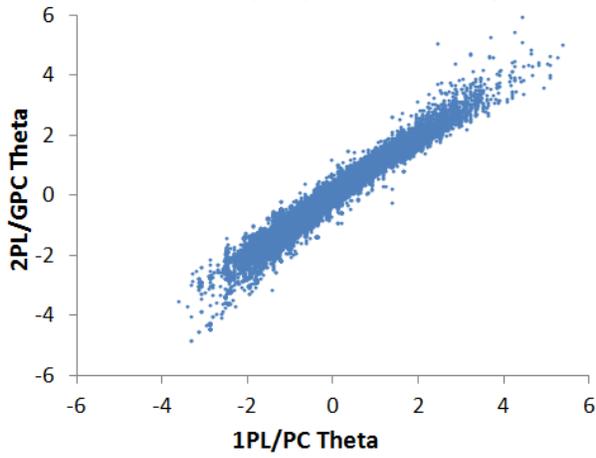


Figure 74. Mathematics Scatter Plots of Theta Estimates Across Different Model Combinations

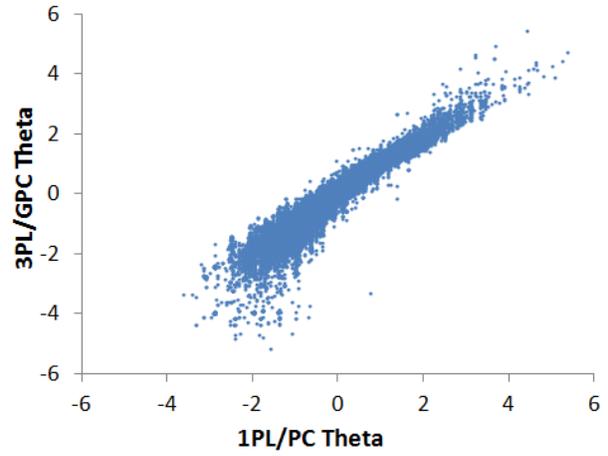




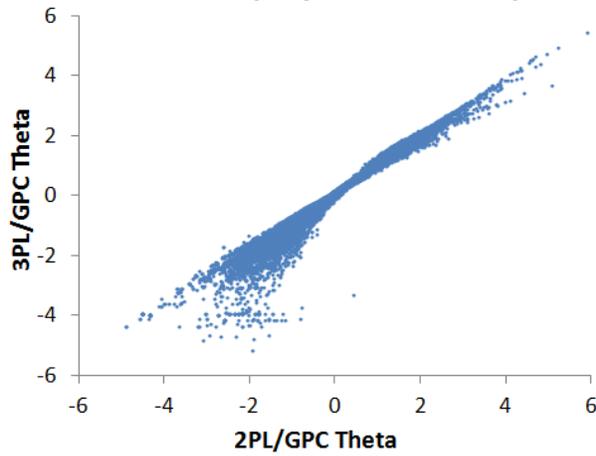
**Scatter Plot of Theta Estimates
Math G07 (1PL/PC vs 2PL GPC)**



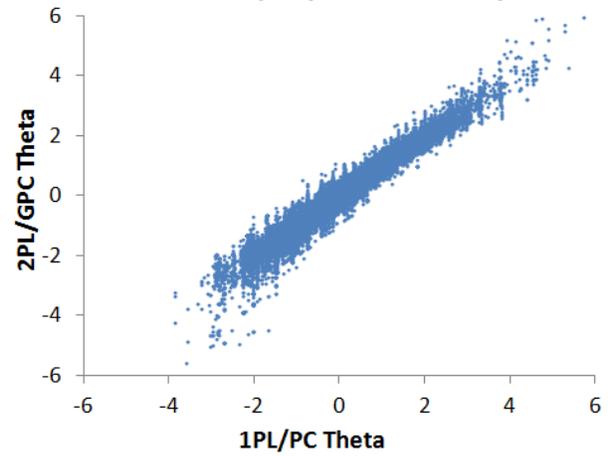
**Scatter Plot of Theta Estimates
Math G07 (1PL/PC vs 3PL GPC)**



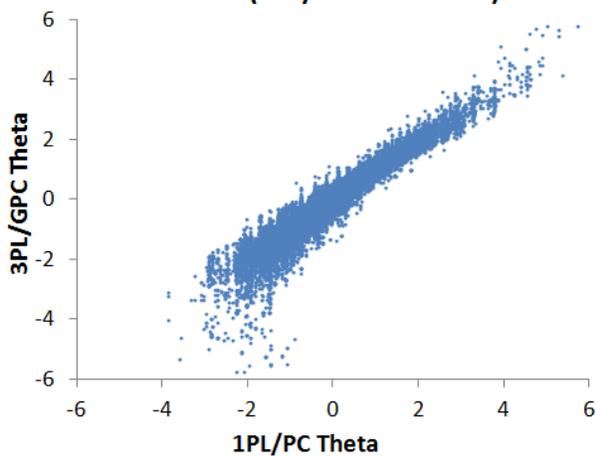
**Scatter Plot of Theta Estimates
Math G07 (2PL/GPC vs 3PL GPC)**



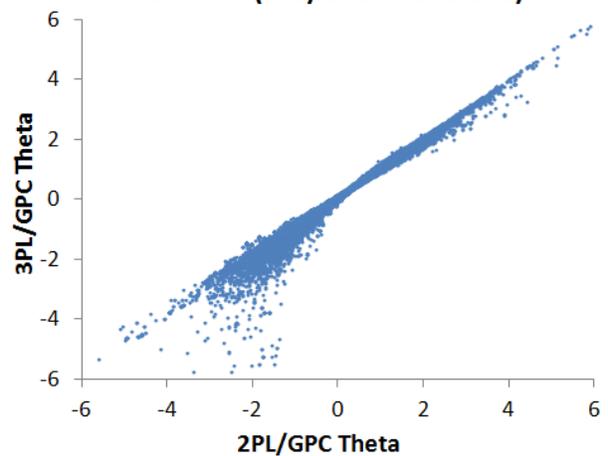
**Scatter Plot of Theta Estimates
Math G08 (1PL/PC vs 2PL GPC)**



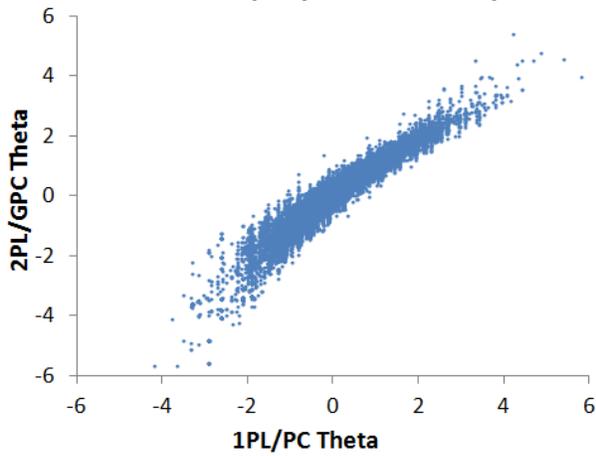
**Scatter Plot of Theta Estimates
Math G08 (1PL/PC vs 3PL GPC)**



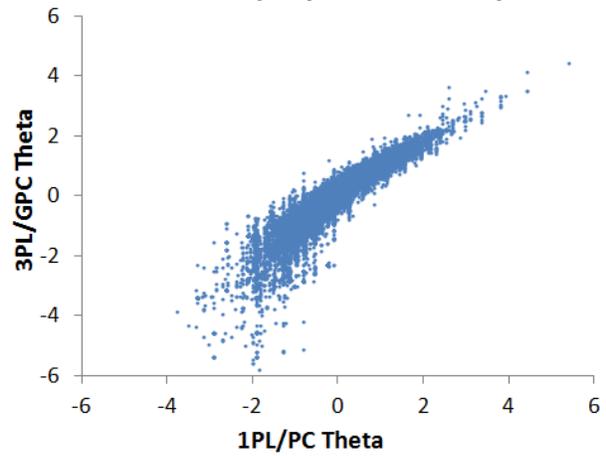
**Scatter Plot of Theta Estimates
Math G08 (2PL/GPC vs 3PL GPC)**



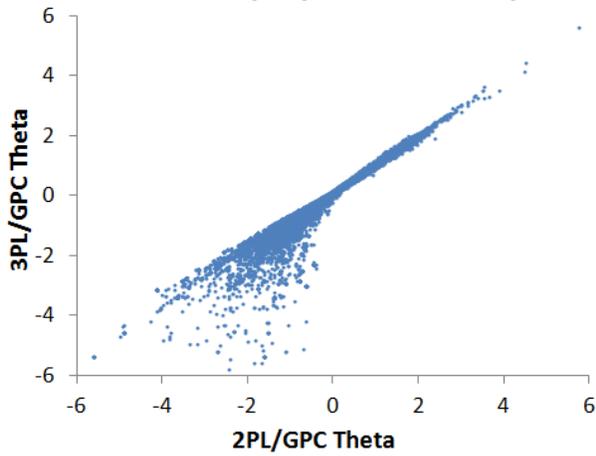
**Scatter Plot of Theta Estimates
Math G09 (1PL/PC vs 2PL GPC)**



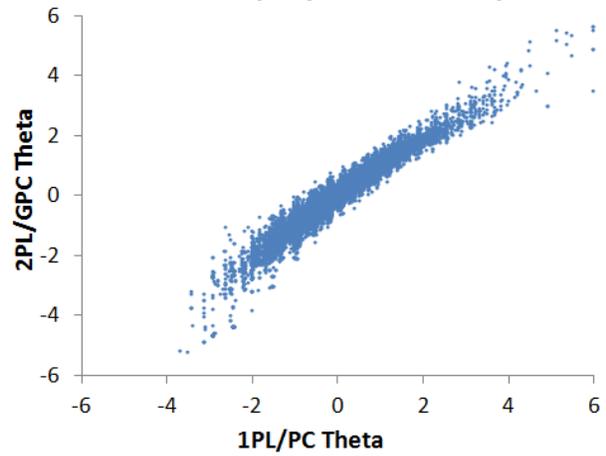
**Scatter Plot of Theta Estimates
Math G09 (1PL/PC vs 3PL GPC)**



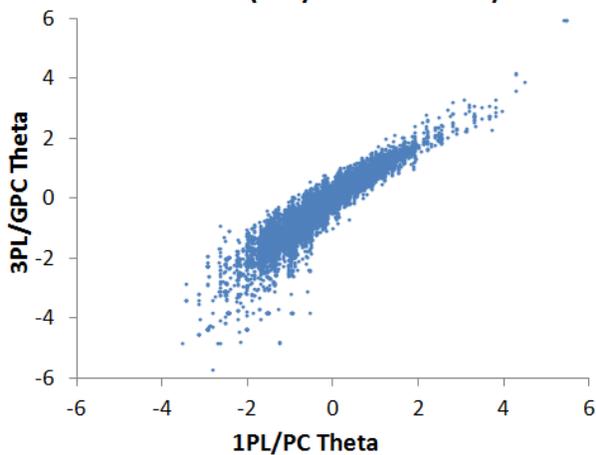
**Scatter Plot of Theta Estimates
Math G09 (2PL/GPC vs 3PL GPC)**



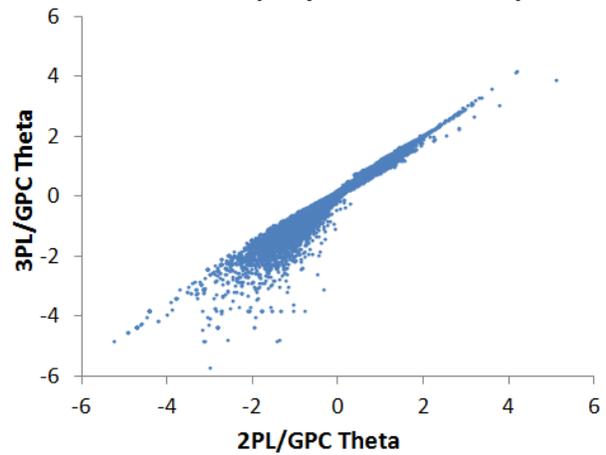
**Scatter Plot of Theta Estimates
Math G10 (1PL/PC vs 2PL GPC)**

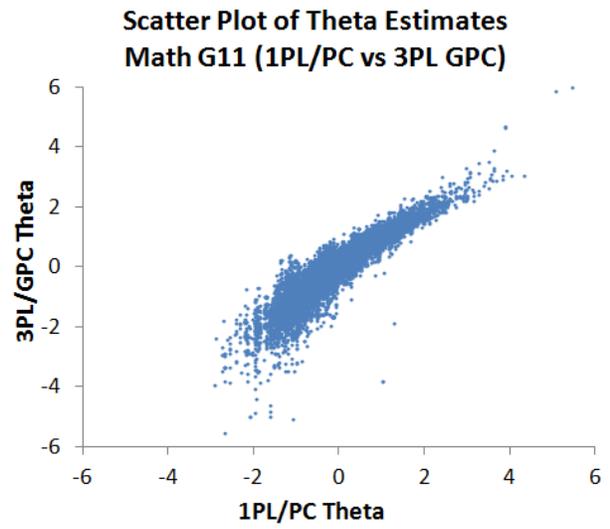
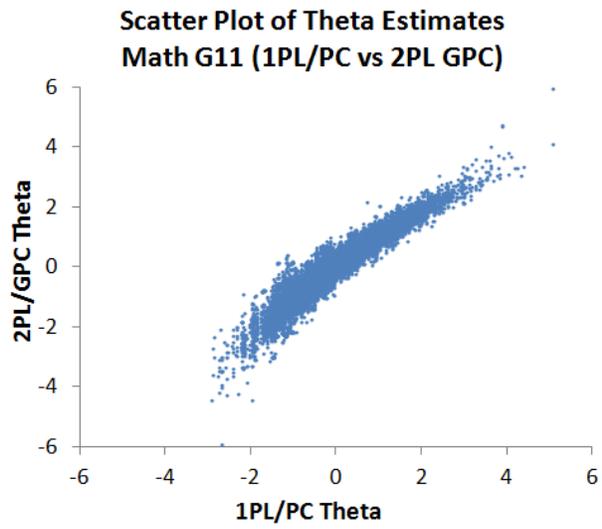


**Scatter Plot of Theta Estimates
Math G10 (1PL/PC vs 3PL GPC)**



**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). *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.
- 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). *BLOG-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.