

Chapter 8 Field Test Data Steps and Classical Test Analyses 3

 Introduction 3

 Data Inclusion/Exclusion Rules for Items and Students..... 4

 Table 1. Summary of Students Excluded and Resulting Sample Size. 5

 Table 2. Summary of ELA Item Exclusions (Item Pool Calibration) by Type..... 7

 Table 3. Summary of Mathematics Item Exclusions (Item Pool Calibration) by Type..... 7

 Item Pool Composition (Vertical Scaling and Item Pool Calibration Steps) 7

 Table 4. Summary of ELA Vertical Scaling Items by Purpose and Type..... 8

 Table 5. Summary of Vertical Scale Mathematics Items by Purpose and Type. 9

 Table 6. Number of On-grade Vertical Scaling Items by Content Area and Characteristics..... 10

 Table 7. Number of Off-grade Vertical Linking Items by Content Area and Characteristics. 11

 Table 8. Number of On-grade Calibration Items by Content Area and Characteristics. 12

 Classical Item and Test Analysis 12

 Item Difficulty..... 13

 Item Discrimination..... 14

 Distractor Analysis. 15

 Reliability Analyses. 16

 Item Flagging Criteria for Content Data Review. 16

 Table 9. Item Flagging Based on Classical Statistics and Judgmental Review. 17

 Table 10. Summary of Vertical Scaling Items with Flags (ELA)..... 18

 Table 11. Summary of Vertical Scaling Items with Flags (Mathematics). 20

 Table 12. Summary of Item Flags for the Item Pool Calibration (ELA). 22

 Table 13. Summary of Items with Flags for the Item Pool Calibration (Mathematics). 23

 Field Test Classical Results 24

 Vertical Scaling Results: Classical Item and Test Statistics..... 24

 Table 14. Number of Items, Average Item Difficulty, and Discrimination for ELA Vertical Scaling Items. 25

 Table 15. Number of Items, Average Item Difficulty, and Discrimination for Mathematics Vertical Scaling Items. 26

 Figure 1. P-value Plots for Vertical Linking Items (ELA) (AIS is used here as association between p-values)..... 27

 Figure 2. P-value Plots for Vertical Linking Items (Mathematics) (AIS is used here as association between p-values) 28

 Table 16. Pearson Correlation between CAT and Performance Tasks for the Vertical Scaling. . 29

Table 17. Number of Items, Average Item Difficulty, and Discrimination for the for Item Pool Calibration.....	30
Table 18. Pearson Correlations between CAT and Performance Tasks for the Item Pool Calibration.....	31
Table 19. Test Reliability and SEM of Performance Tasks for the Item Pool Calibration.	32
Figure 3. Distributions of Total Raw Scores as a Percentage of the Corresponding Maximum Possible Score for the Item Pool Calibration Sample (ELA)	34
Figure 4. Distributions of Total Raw Scores as a Percentage of the Corresponding Maximum Possible Score for the Item Pool Calibration Sample (Mathematics)	35
Subgroup Analysis of Test Difficulty for the Item Pool Calibration.....	36
Table 20. Summary of Average Test Difficulty by Subgroup for ELA.	36
Table 21. Summary of Average Test Difficulty by Subgroup for Mathematics	40
Differential Item Functioning (DIF) Analyses for the Calibration Item Pool	43
Table 22. Definition of Focal and Reference Groups.	44
Table 23. DIF Flagging Logic for Selected-Response Items.	44
Table 24. DIF Flagging Logic for Constructed-Response Items	45
Table 25. Number of DIF Items Flagged by Category (ELA, Grades 3 to High School).	47
Table 26. Number of DIF Items Flagged by Category (Mathematics, Grades 3 to High School).	48
Prospective Evidence in Support of Rater Agreement	49
Monitoring Scoring Processes.....	49
Monitoring Raters and Associated Statistics.....	49
Monitoring Automated Scoring and Associated Statistics.	50
External Assessments: NAEP and PISA.....	50
Table 27. Comparison of Features across the Smarter Balanced, NAEP, and PISA Assessment Programs.....	51
PISA Overview.	52
NAEP Overview.....	52
Results	53
Table 28. Number of Items, Average Item Difficulty, and Discrimination for NAEP and PISA Items.	54
Figure 5. Comparison of NAEP Item Difficulty and Values Obtained from Smarter Balanced Samples	55
References	56

Chapter 8 Field Test Data Steps and Classical Test Analyses

Introduction

There were two steps or phases for the Smarter Balanced Field Test. The purpose of the first step of the Field Test was a) to include a robust set of items to establish the horizontal and vertical scales using a representative student sample, and b) to perform the achievement level setting conducted in the fall of 2014. In the vertical scaling sample, selected computer adaptive test (CAT) items and performance tasks (PTs) were administered to students at adjacent lower grades to permit vertical linking (i.e., the off-grade administration of items/tasks). A more manageable number of items and students was needed to conduct the vertical scaling (i.e., standard setting) within the short time confines of the Field Test in order to permit the 2014 achievement level setting. The purpose of the second step of the Field Test was to calibrate a large, robust pool of items onto the scale established in the vertical scaling analysis. All the items in the vertical scaling were also administered on-grade to students in the calibration study for linking purposes. In the calibration step, students were administered items/tasks that were targeted for that grade (i.e., on-grade administration only). This horizontal calibration (item-pool calibration) and linking resulted in the final entire parameter estimates for the Smarter Balanced item pool. The test windows (i.e., administration) of the vertical scaling and item pool calibration steps overlapped in the spring of 2014. The primary focus here is to demonstrate the properties of the Smarter Balanced items pool at the conclusion of the item-pool calibration step that reflected the items available for operational administrations. The results presented are for the item pool calibrations step unless they are also explicitly labeled as vertical scaling outcomes.

In accordance with the Field Test design, students were administered a CAT component, which was intended to conform to the Smarter Balanced test blueprint, and a selected performance task. The CAT and performance task components work in conjunction to fulfill the content requirements for the test blueprint. They consist of a variety of different item types, some of which were machine scored. All single-selection selected-response (SR) items had three to five answer choices. Multiple-selection selected-response (MSR) items provided five to eight answer choices. The performance task items had scores ranging from zero to a maximum of four score points. In the case of ELA/literacy (ELA), the extended student-writing sample was scored for three dimensions of writing (purpose/focus/organization, evidence/elaboration, and conventions). A performance task was expected to have approximately 4 to 6 scorable units (i.e., items) yielding approximately 12 to 15 score points in total. For example in grade 6 mathematics, a task could have six items with maximum score levels corresponding to 1, 2, 2, 2, 2, 3 that pertained to three short answer items and three equation items. In grade 11 ELA, there were 11 maximum raw score points associated with a task; six points for the extended writing response, two short answer items and a matching item.

The CAT items were administered in the context of linear-on-the-fly testing (LOFT), in which the test content that was sampled conformed to the Smarter Balanced test blueprint. Unlike a fixed or linear test form (e.g., a paper-and-pencil administration), there are no intact test forms common across students which are necessary to compute classical test reliability or the overall “number correct” common across substantial numbers of students. The primary advantage of the LOFT administration is its efficiency in test delivery since test forms can be constructed dynamically for each student that conform to the CAT test blueprint. Given the size of the item pool in a grade and content area, it would have been difficult to construct the necessary number of linear, blueprint conforming test forms without a LOFT administration.

Decisions concerning the data steps are included here since they had important implications for the resulting item quality and composition of the resulting item pools in ELA/literacy and mathematics. Further information can be found on vertical and horizontal scaling in Chapter 9 and on test design

in Chapter 4. An explanation of the differential item functioning (DIF) methods used is given in Chapter 6 on the Pilot Analysis.

Data Inclusion/Exclusion Rules for Items and Students

The first step was to create a sparse data matrix for analysis reflecting item scores as well as missing information by design. Each row of the matrix was a student response vector where the columns were the available items. For a given grade, the dimension of this sparse matrix is the total number of students times the total number of unique items (i.e., scorable units). Many of the cells of this matrix represent items not administered to students by design. This “missing” information was indicated in the sparse matrix as “not presented” items, which is the typical practice when multiple test forms exist. Smarter Balanced defined condition codes for various sorts of invalid responses to polytomous items that consisted of the following designations: B (Blank), U (Insufficient), F (Non-scorable language), T (Off topic), and M (Off purpose). These condition codes were ultimately resolved as scores of zero in the data matrix used in calibration. This data matrix for each grade and content area was then the focus of the subsequent analyses.

The inclusion/exclusion rules were applied prior to the classical test analysis and IRT calibrations in order to ensure that the best possible statistical outcomes resulted. Inclusion and exclusion logic for both items and students were also implemented using IRT statistics. This IRT item exclusion might include issues like non-convergence during parameter estimation or very low IRT discrimination values. They are included in this data step to help avoid confusion concerning the final number of Field Test students and items. Since there were a limited number of performance tasks, extra effort was made to preserve the associated items within a given task. First, the student exclusion rules are presented. These are followed by rules applied to both the CAT selected-response (SR), constructed-response (CR) items, and the performance tasks (PT).

Student Exclusions. The following rules were implemented in the vertical scaling step. For the item calibration step, short tests (less than 9 items) were included in the analysis.

1. A record was excluded if a student was deemed to not have made a reasonable attempt on the Smarter Balanced Field Test. Students were eliminated if their response time (i.e., test duration) was very short, which likely indicated that a reasonable effort was not made or some other anomaly occurred. Test duration was defined as the time when the student entered the test administration until the test was completed using the “submit” button.
 - a. A student record was excluded if the full-length CAT test event was completed in less than 15 minutes.
 - b. Note that in California, students took shortened ELA/literacy and mathematics CAT components expected to be approximately 25 items in each content area as opposed to 50 items in a single content area. Half-length CAT events administered to California students were eliminated if the test duration was less than eight minutes.
 - c. If the ELA/literacy performance task was completed in less than 15 minutes or the mathematics performance task was completed in less than ten minutes, the student’s score was excluded.
2. All student records with a zero on all item scores were excluded.
3. Students with scored responses to less than nine items were excluded.

The impact of applying these student exclusion rules is shown in Table 1. The number of students excluded was relatively negligible except in high school ELA/literacy and mathematics. Table 1 reflects the number of students that have valid performance task scores (all students have CAT scores), which may be lower than the student counts in other tables.

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

Vertical Scaling			Item Pool Calibration			
Grade	No. Students	No. Excluded	No. Valid	No. Students	No. Excluded	No. Valid
ELA						
3	23,788	585	23,203	85,889	1,830	84,059
4	36,271	572	35,699	94,915	1,393	93,522
5	32,220	614	31,606	88,293	1,503	86,790
6	32,229	681	31,548	93,536	1,790	91,746
7	32,005	1,126	30,879	93,431	2,895	90,536
8	37,129	1,213	35,916	98,433	3,163	95,270
HS	57,608	7,073	50,535	261,405	27,462	233,943
Mathematics						
3	25,671	848	24,823	95,143	2,604	92,539
4	39,522	595	38,927	109,441	1,645	107,796
5	42,818	433	42,385	108,412	1,186	107,226
6	34,014	775	33,239	117,691	2,172	115,519
7	32,176	1,885	30,291	117,049	5,342	111,707
8	38,856	2,135	36,721	116,459	5,656	110,803
HS	56,658	7,375	49,283	262,111	37,425	224,686

Item Exclusions and Data Steps. Item quality was inspected using frequency distributions and classical item statistics prior to conducting the IRT calibration. After consultation with Smarter Balanced, poor-quality items were excluded by using either statistical or judgmental rules. Items were excluded based on the following rules and guidelines:

1. all selected-response items with rounded item difficulty at or below 0.10;
2. CAT (polytomous, non-selected-response items) with a rounded item difficulty at or below 0.02,

3. performance tasks (performance task non-selected-response items) with a rounded average item difficulty at or below 0.01;
4. CAT polytomous items with any score categories having 10 or fewer observations;
5. all dichotomous items that have 30 or fewer observations obtaining a score point of 1;
6. items having fewer than 500 observations; and
7. selected-response items incorporating the combined psychometric staff evaluation of the item-total correlation and empirical item plots.

For constructed-response items, score categories with fewer than 10 students at on-grade level were collapsed with neighboring categories in both on-grade and off-grade data sets. If the category that needed to be collapsed was a middle category, it was collapsed with the category with smaller number of observations.

Using IRT-derived rules, additional item exclusions were performed to ensure the most reasonable item and student estimates would result. Items were excluded based on the following IRT-derived rules.

- a. Non-convergence during Marginal Maximum Likelihood (MML) estimation
- b. Discrimination parameter estimates below 0.10
- c. The quality of additional items were evaluated based on
 - i. Selecting outliers by rank ordering the IRT discrimination parameters and classical item-total correlations
 - ii. Selecting outliers by rank ordering the IRT difficulty parameter and observed p-value
 - iii. Identifying unreasonably high chi-square by rank ordering sample size and chi-square
 - iv. Identifying large standard errors for IRT discrimination and/or difficulty parameters
 - v. Item characteristic curves with poor fit between observed and expected performance.

Tables 2 and 3 show a summary of the total item inventory—the items lost strictly to content and scoring decisions, items analysis, and IRT exclusions. They show the number of items that survived (Final Pool) after all the exclusion rules were applied in ELA and mathematics for the item pool. The subsequent IRT exclusions were included here for completeness. These tables list the original inventory of all items developed and the number of items not used or otherwise scored for content reasons. The “sample size” column shows the number of items eliminated for small sample size or fewer than 10 observations in a score category and applying the various exclusion rules. No items were dropped from the calibration analysis because of DIF. A large number of items were precluded from IRT analysis due to small sample size in high school ELA and mathematics based on classical test analysis. The final set of items was used to derive the classical item and test statistics of record and those entering into the IRT scaling labeled under the “Resulting Pool” column. A significant number of items were not calibrated due to an insufficient sample size in high school. These items can be piloted and scaled in subsequent operational administrations.

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

Grade	Initial	Content	Small Sample Size		Poor Item Statistics		Final
	Pool	Issues	(50,300)	(300,500)	Classical	IRT	Pool
3	1,045	30	13	18	69	19	896
4	965	17	13	19	38	22	856
5	975	23	31	14	65	19	823
6	984	23	19	11	60	22	849
7	1,033	27	20	11	77	23	875
8	1,010	20	17	23	95	19	836
HS	3,371	61	272	386	248	33	2,371

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

Grade	Initial	Content	Small Sample Size		Poor Item Statistics		Final
	Pool	Issues	(50,300)	(300,500)	Classical	IRT	Pool
3	1,163	1	-	-	44	4	1,114
4	1,207	9	-	-	56	12	1,130
5	1,108	2	-	-	45	18	1,043
6	1,115	8	-	-	80	9	1,018
7	1,037	5	-	-	76	14	942
8	1,036	9	-	-	103	30	894
HS	3,386	75	25	772	433	55	2,026

Item Pool Composition (Vertical Scaling and Item Pool Calibration Steps)

Since the vertical scaling item sets were used to establish the Smarter Balanced scales, it is important to delineate the composition of the items types. Tables 4 and 5 classified items by purpose and type for the vertical scaling. Items were targeted for on-grade or off-grade administration for the vertical scaling. The mixture of items types included both selected- and constructed-response items. Constructed-response items could be dichotomously (right/wrong) or polytomously (with provision for partial credit) scored. The item counts reflect both CAT and

performance task items. NAEP and PISA items were also given in selected grades. These tables also show the number of items that remained for vertical scaling after all the item exclusions were applied. Table 6 shows the distributions of the on-grade items by claim and item type. The claims in ELA pertain to reading, writing, listening/speaking, and research, respectively. In mathematics, the claims pertain to concepts/processes, problem solving, communicating reasoning, and modeling/data analysis. Table 7 shows the same types of information for the vertical linking items. Table 8 shows the distribution for all items by claim and type from the calibration item pool for ELA and mathematics. All items contained in the calibration step consisted of all available items in the pool targeted in the Field Test for on-grade administration; this was inclusive of the vertical scaling items. The readministration of “on-grade vertical scaling” items was necessary to link the item pool calibration items onto the scale.

Table 4. Summary of ELA Vertical Scaling Items by Purpose and Type.

Item Purpose	Response Type	Score Type	Grade						
			3	4	5	6	7	8	HS
On-Grade	Selected-response		115	92	95	81	77	91	142
	Other	Dichotomous	110	119	122	114	122	113	216
		Polytomous	36	31	39	37	39	39	52
Off-Grade Vertical Linking	Selected-response			57	53	46	27	38	39
	Other	Dichotomous		45	62	63	58	62	58
		Polytomous		18	18	22	22	23	10
NAEP	Selected-response			22				20	12
	Other	Dichotomous		2				2	4
		Polytomous		4				8	11
PISA	Selected-response								17
	Other	Dichotomous							12
		Polytomous							4

Table 5. Summary of Vertical Scale Mathematics Items by Purpose and Type.

Item Purpose	Response Type	Score Type	Grade						
			3	4	5	6	7	8	HS
On-Grade	Selected-response		48	65	78	21	39	41	66
	Other	Dichotomous	221	212	175	174	185	159	203
		Polytomous	35	29	53	27	15	30	50
Off-Grade Vertical Linking	Selected-response			11	12	31	9	7	18
	Other	Dichotomous		76	71	56	55	60	56
		Polytomous		17	12	15	7	6	7
NAEP	Selected-response			20				19	18
	Other	Dichotomous		2				6	4
		Polytomous		8				8	6
PISA	Selected-response								19
	Other	Dichotomous							44
		Polytomous							11

Table 6. Number of On-grade Vertical Scaling Items by Content Area and Characteristics.

Item Type	Grade						
	3	4	5	6	7	8	HS
ELA							
Total	261	242	256	232	238	243	410
Selected-response	115	92	95	81	77	91	142
Dichotomous	110	119	122	114	122	113	216
Polytomous	36	31	39	37	39	39	52
Claim 1	94	72	91	71	75	83	181
Claim 2	70	67	67	67	70	66	126
Claim 3	50	51	46	45	46	49	39
Claim 4	47	52	52	49	47	45	64
Mathematics							
Total	304	306	306	222	239	230	319
Selected-response	48	65	78	21	39	41	66
Dichotomous	221	212	175	174	185	159	203
Polytomous	35	29	53	27	15	30	50
Claim 1	184	182	182	107	134	130	191
Claim 2	17	17	17	20	10	15	22
Claim 3	47	51	49	40	38	35	44
Claim 4	19	23	20	19	21	17	33
Unclassified	37	33	38	36	36	33	29

Table 7. Number of Off-grade Vertical Linking Items by Content Area and Characteristics.

Item Type	Grade					
	4	5	6	7	8	HS
ELA						
Total	120	133	131	107	123	107
Selected-response	57	53	46	27	38	39
Dichotomous	45	62	63	58	62	58
Polytomous	18	18	22	22	23	10
Claim 1	40	54	53	41	49	48
Claim 2	34	32	31	29	34	25
Claim 3	26	25	25	18	24	21
Claim 4	20	22	22	19	16	13
Mathematics						
Total	104	95	102	71	73	81
Selected-response	11	12	31	9	7	18
Dichotomous	76	71	56	55	60	56
Polytomous	17	12	15	7	6	7
Claim 1	58	55	60	28	36	46
Claim 2	4	3	4	4	3	5
Claim 3	16	18	13	15	9	12
Claim 4	7	6	7	6	7	6
Unclassified	19	13	18	18	18	12

Table 8. Number of On-grade Calibration Items by Content Area and Characteristics.

Item Type	Grade						
	3	4	5	6	7	8	HS
ELA							
Total	896	856	823	849	875	836	2371
Selected-response	386	336	286	300	280	301	875
Dichotomous	381	376	379	413	437	372	1216
Polytomous	129	144	158	136	158	163	280
Claim 1	317	259	265	274	299	258	867
Claim 2	243	248	241	257	262	241	729
Claim 3	163	157	142	147	152	174	383
Claim 4	173	192	175	171	162	163	392
Mathematics							
Total	1114	1130	1043	1018	942	894	2026
Selected-response	166	189	239	101	109	161	530
Dichotomous	815	789	633	792	743	610	1247
Polytomous	133	152	171	125	90	123	249
Claim 1	672	677	613	576	519	493	1123
Claim 2	55	68	55	77	71	59	147
Claim 3	166	145	168	132	120	134	433
Claim 4	68	77	84	68	67	64	185
Unclassified	153	163	123	165	165	144	138

Before presenting the classical results, a discussion of processing of ELA essay scores for performance tasks is necessary. For performance tasks in ELA/literacy, students were administered a writing task (i.e., extended writing response) that is scored on three dimensions of writing that correspond to organization (0-4 points), elaboration (0-4 points), and conventions (0-2 points). That is, three separate scores (i.e., scorable units) are obtained for a single student writing sample. The correlations between the dimensions of organization and elaboration exceeded 0.95 in many instances. This high degree of dependence precluded them from being calibrated as separate items due to very high local item dependence. As a result, the two writing dimensions for organization and elaboration were averaged and rounded up if necessary. This resulted in a single 0 to 4 point score for these two dimensions along with the original conventions score (0-2) for the long writing task. These three ELA/literacy raw scores from the long writing task were then used in the IRT scaling.

Classical Item and Test Analysis

Classical (traditional or observed) item and test statistics are presented here for both items and tests. Tests are defined as the collection of items administered to students for the CAT using Linear-on-the-Fly-Testing (LOFT) administration combined with a performance task. This test definition

pertains to the items remaining after all the item exclusions were applied. Both CAT and performance task components were needed to fulfill the test blueprint. Each statistic provides some key information about the quality of each item or test based on empirical data from the Smarter Balanced assessments. Classical measures include statistics such as item difficulty, item-test correlations, and test statistics (e.g., reliability). Other descriptive measures, such as the percentage of students at each response option or score level, were used to evaluate item functioning but were not reported here. Classical test analyses were conducted in part to gain information about the quality of items, such as the following:

- Based on item difficulty, is the item appropriate for testing at a given grade level?
- How effective is the item in distinguishing students with high and low ability? Did higher ability students perform better on the item than lower ability students?
- For selected-response (SR) items, is the key the only correct choice? Are all item distractors wrong? Are distractors constructed in a way that is more attractive to low ability students compared with high ability ones? Are high ability students more likely to choose the key than the distractors?
- For constructed-response (CR) items, do high ability students tend to score in upper score categories and less able students in lower ones?
- For an item that is administered in multiple grades for the purpose of vertical scaling, do students in a higher grade level tend to perform better on the item than students in a lower grade level?
- Does the item show DIF? In other words, does the item tend to be especially difficult for a specified group of students with comparable levels of ability?
- Are scores sufficiently reliable for the intended purposes?

To address these properties, the analyses include several components: item difficulty, item discrimination, item response distribution for CR items, differential item functioning and score reliability. In the context of the Field Test, these statistics also provided information that was used to exclude poorly functioning items prior to the IRT calibration step or to inform future, item writing activities on the part of content developers.

As mentioned at the outset, the presentation of statistics is more difficult to summarize in some respects since no fixed forms containing a set number of items exist in the Field Test. There were many potential combinations of CAT test forms presented to students due to the LOFT administration and the sheer number of items in the pool in a given grade and content area. Several types of classical analysis rely on the provision of a criterion variable for computing item-test correlations or differential item functioning that is typically defined as the total raw score. Since there are many variations in the CAT items presented to students due to the LOFT administration along with performance tasks, defining a common test criterion was not possible. To circumvent these problems and provide the best available criterion score, the student ability (i.e., theta estimate) was used. As a result, item-test correlations and DIF depended on incorporating IRT ability as the criterion score. Chapter 9 on the IRT analysis provides a description of the methods used to compute theta. This modified classical test analysis was conducted to obtain additional evidence concerning item and test properties, item pool characteristics, and eventually performing the data review of items by content developers for operational administrations.

Item Difficulty. The percent of maximum possible score is computed for each item as an indicator of item difficulty with a range of 0.0 to 1.0. A relatively higher value indicates an easier item. An item difficulty of 1.0 indicates that all students received a perfect score on the item. An average item

score of 0.0 for an item indicates that no students answered the item correctly or only received partial credit for the item in the case of polytomous or CR items.

For dichotomous items and SR items, the percent of maximum possible score is simply equivalent to the percentage of students who answered the item correctly. The formula for p -value for selected response is

$$p\text{-value}_{SR} = \frac{\sum X_{ic}}{N_i},$$

where X_{ic} is the number of students that answered item i correctly, and N_i is the total number of students observed for item i .

A polytomous item is an item that is scored with more than two ordered categories, such as the scores from the ELA/literacy performance task essay. For polytomous items (i.e., constructed-response), the p -value is defined as

$$p\text{-value}_{CR} = \frac{\sum X_{ij}}{N_i \times \text{Max}(k)},$$

where X_{ij} is the score assigned for a given constructed-response item and k are the score levels associated with the item. Another interpretation is that item difficulty for constructed-response items is the mean score for the item divided by the maximum score. For example, a polytomous item had scores ranging from a low score of zero to three as the maximum and the observed mean score was 2.1. The observed percent of maximum can also be calculated as $2.1/3 = 0.70$, or 70 percent, of the maximum score was achieved by students on this hypothetical constructed-response item. In the case of a selected-response item (i.e., multiple-choice), the maximum score is one by definition and defaults to the selected-response p -value.

A wide item difficulty range is needed to measure student ability that can vary greatly particularly for operational administrations of the adaptive test. Very easy or difficult items require additional review to ensure that the items are valid and are grade appropriate. Note that some items served as anchor items in vertical scaling. These items are administered across multiple grade levels and therefore can have several sets of grade-level-specific classical item statistics. For vertical scaling, item difficulty across different grade levels was assessed to evaluate if students in the upper grade level generally performed better in comparison with a lower grade level.

Item Discrimination. Item discrimination evaluates how well an item distinguishes between low and high ability students. In classical (non-IRT) item analysis it is the correlation between the item score and total test score. The expectation is that high ability students will outperform low ability students on an item. The item discrimination statistic is calculated as the correlation coefficient between the item score and criterion score (i.e., IRT ability estimate). A relatively high item-total correlation coefficient value is desired, as it indicates that students with higher scores on the overall test tended to perform better. In general, item-total correlation ranges from -1.0 (for a perfect negative relationship) to 1.0 (for a perfect positive relationship). However, a negative item-total correlation typically signifies a problem with the item, as the higher ability students generally are getting the item wrong or a low score and the lower ability students are getting the item right or are assigned a higher score level.

Some coefficients used in computing item-total correlations are the point-biserial and polyserial correlation coefficient. The point-biserial correlation is used for dichotomous items and polyserial correlation used for polytomous items. The point-biserial correlation coefficient is a special case of

the Pearson correlation coefficient used for dichotomous items. The point-biserial correlation is computed using

$$r_{ptbis} = \frac{(\mu_+ - \mu_-)}{\sigma_{tot}} \sqrt{pq}$$

where μ_+ is the mean criterion score of examinees answering the item correctly, μ_- is the mean criterion score of the examinees answering the item incorrectly, σ_{tot} is the standard deviation of the criterion score, p is the proportion of examinees answering the item correctly, and q equals $(1 - p)$.

The polyserial correlation measures the relationship between a polytomous item and the criterion score. Polyserial correlations are based on a polyserial regression model (Olsson, 1979; Drasgow, 1988), which assumes that performance on an item is determined by the examinee's position on an underlying latent variable that is normally distributed at a given criterion score level. Based on this approach, the polyserial correlation can be estimated as

$$r_{polyreg} = \frac{\beta \sigma_{tot}}{\sqrt{\beta^2 \sigma_{tot}^2 + 1}}$$

in which β is a series of parameters estimated using maximum likelihood and σ_{tot} is the standard deviation of the criterion score. The biserial correlation could have been chosen for dichotomous items but the point-biserial and its interpretation is more familiar to many users.

Distractor Analysis. For each selected-response item, distractor analyses were conducted. The quality of distractors is an important component of an item's overall quality. Distractors should be clearly incorrect, but at the same time plausible and attractive to lower ability students. The following distractor analyses are conducted to evaluate the quality of distractors.

- The percentage of students at each response option is calculated. For the key (i.e., the correct answer), this percentage is the item difficulty value. If the percentage of students who selected a distractor is greater than the percentage of students who selected a key, the item should be examined to determine if it has been incorrectly keyed or double-keyed.
- The point-biserial correlation is calculated for each response option. While the key should have a positive point-biserial correlation with the criterion score, the distractors should exhibit negative point-biserial correlations (i.e., lower ability students would likely choose the distractors, while the higher ability students would not).
- The average ability level (measured by criterion score) is calculated for students at each response option. Students choosing the key should be of higher ability levels than students choosing distractors.
- The percentage of high ability students at each response option is calculated. High ability students were defined as the top 20 percent of students in the ability distribution (grade and content area). If the percentage of high ability students who selected a distractor is greater than the percentage of high ability students who selected a key, the item should be examined further.

For each constructed-response item, the following statistics are evaluated.

- The percentage of students at each score level is calculated. If there were very few students at certain score levels, this might suggest that some score categories need to be collapsed or that the scoring rubric needs adjustment.

- The average ability level is calculated for students at each score level. Students at a higher score level on this item should be of higher ability levels (i.e., having higher average ability estimates) than students at a lower score level on this item.
- The item-test correlation is computed using the polyserial correlation.

Reliability Analyses. The variance in the distributions of test scores, essentially the differences among individuals, is partly due to real differences in the knowledge, skills, or ability being tested (true variance) and partly due to random errors in the measurement process (error variance). Score reliability is an estimate of the proportion of the total variance that is true variance. The estimates of reliability used here are internal-consistency measures. The formula for the internal consistency reliability, as measured by Cronbach’s coefficient alpha (Cronbach, 1951), is

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_x^2} \right],$$

where n is the number of items, σ_i^2 is the variance of scores on the i^{th} item, and σ_x^2 is the variance of the total score (sum of scores on the individual items).

The standard error of measurement (SEM) provides a measure of score instability in the score metric. The formula for computing the SEM is

$$\sigma_e = \sigma_x \sqrt{1 - \alpha},$$

where reliability α is Cronbach’s alpha estimated above, and σ_x is the standard deviation of the scores. The SEM can be used to determine the confidence interval (CI) that captures an examinee’s true score.

Item Flagging Criteria for Content Data Review. Flagging is used to identify certain statistic characteristics of items that indicate poor functioning. For example if an item is very difficult for a grade level or has very low discrimination, its properties should be reviewed further by content developers before selecting it for inclusion in the item pool. Content developers reviewed items after the Field Test analysis in conjunction with item statistics. Items with poor classical statistics were designated using various types of flags. These flags were used in conjunction with substantive attention to the item content to determine if a problem exists and if any corrective action was required. At a minimum, flagged items underwent additional scrutiny for content appropriateness, bias and sensitivity, and overall statistical performance relative to expectations. Any item with substantial changes was returned to the item bank for further pretesting prior to operational use. Items that were functioning very poorly could either be excluded from further use or be rewritten to improve their performance as new items. Table 9 lists the flagging definitions for selected- and constructed-response items. Note that items were also flagged for differential item functioning (DIF) presented later. If an item was judged to have potential flaws after reviewing the item content, it was flagged for further content review. The item flags for the vertical scaling are listed in Tables 10 and 11 for ELA and Mathematics and Tables 12 and 13 for the item pool calibration. These tables demonstrate that a significant number of items were flagged “A” indicating difficult items particularly dichotomously scored items. The pattern of flagging was not consistent across grades in the case of vertical linking items.

Table 9. Item Flagging Based on Classical Statistics and Judgmental Review.

Flag	Definition
A	High difficulty (less than 0.10)
B	CR items with percentage obtaining any score category less than three percent of total N
C	CR items with higher criterion score mean for students in a lower score-point category
D	SR items with proportionally more high-proficient students selecting a distractor over the key
F	SR items with higher criterion score mean for students choosing a distractor than the mean for those choosing the key
H	Low difficulty (greater than 0.95)
P	SR items with positive distractor biserial correlation
R	Low item-total correlation (less than 0.30)
V	Item more difficult at the higher-grade level for vertical linking items
Z	Item needs content review (judgmental decision)

Table 10. Summary of Vertical Scaling Items with Flags (ELA).

Grade	Grade Assignment	Response Type		Flags										
				A	B	C	D	F	H	P	R	V	Z	
3	3	SR					5	2		29	18	1		
		Other	Dichotomous	26	4						8	1	5	
			Polytomous	16	25	1								
4	3	SR					1	1		23	7	1		
		Other	Dichotomous	4	2						1	1	1	
			Polytomous	4	4									
	4	SR			1			9	4		16	17	1	
		Other	Dichotomous	19	2						9	1	1	
			Polytomous	8	10									
5	4	SR					1			15	5	1		
		Other	Dichotomous	4	1						2	1		
			Polytomous	1	4									
	5	SR			2			8	8		24	18	6	
		Other	Dichotomous	17	2	1					9	3	4	
			Polytomous	6	8						1	1	1	
6	5	SR					4	6		23	10	6		
		Other	Dichotomous	4		1					4	3	2	
			Polytomous	2	1						1	1	1	
	6	SR						11	6		22	19	5	2
		Other	Dichotomous	29	8	2					24	8	6	
			Polytomous	2	10							4		

Grade	Grade Assignment	Response Type		Flags									
				A	B	C	D	F	H	P	R	V	Z
7	6	SR					2	2		14	8	5	
		Other	Dichotomous	10	1						7	8	2
			Polytomous	1	1							4	
	7	SR					10	6		22	22	3	
		Other	Dichotomous	29	9						19	6	2
			Polytomous	2	9							1	
8	7	SR					2			10	9	3	
		Other	Dichotomous	13	3						13	6	4
			Polytomous		2							1	
	8	SR					13	9		38	36	15	
		Other	Dichotomous	31	7	1					24	13	5
			Polytomous	1	8							6	
HS	8	SR					6	4		9	15	15	1
		Other	Dichotomous	8							9	13	1
			Polytomous									6	
	HS	SR					14	9		54	37		1
		Other	Dichotomous	72	15	2					52		3
			Polytomous	2	10	1							

Table 11. Summary of Vertical Scaling Items with Flags (Mathematics).

Grade	Grade Assignment	Response Type		Flags									
				A	B	C	D	F	H	P	R	V	Z
3	3	SR					4	2		10	8		
		Other	Dichotomous	34	7						6	4	
			Polytomous	3	7						2		
4	3	SR								1	1		
		Other	Dichotomous	5							2	4	
			Polytomous		1				1				
	4	SR				4	2		19	12			
		Other	Dichotomous	31	4						5	2	
			Polytomous	3	12	1							
5	4	SR							5	1			
		Other	Dichotomous	11	1					1	2		
			Polytomous	1	4	3							
	5	SR				2			22	12	3		
		Other	Dichotomous	39	5	1					9	5	
			Polytomous	12	11						1	1	
6	5	SR							21	1	3		
		Other	Dichotomous	9	1					4	5		
			Polytomous	2	5						1		
	6	SR				4	2		12	7	2		
		Other	Dichotomous	50	13						11	15	1

Grade	Grade Assignment	Response Type		Flags										
				A	B	C	D	F	H	P	R	V	Z	
			Polytomous	4	5								5	
7	6	SR					3	1		10	3	2		
		Other	Dichotomous	13	2						1	15		
			Polytomous										5	
	7	SR			1			4	3		8	6	3	
		Other	Dichotomous	61	19							15	20	
			Polytomous	9	7								3	
8	7	SR					2	3		4	4	3		
		Other	Dichotomous	18	2						9	20	1	
			Polytomous	2	2								3	
	8	SR						11	6		24	19	3	
		Other	Dichotomous	77	32							15	11	1
			Polytomous	11	15							2	1	
HS	8	SR					2			1	3	3		
		Other	Dichotomous	19	5						1	11		
			Polytomous	2	3								1	
	HS	SR			3			29	10		59	48		
		Other	Dichotomous	99	32							28		2
			Polytomous	15	12									

Table 12. Summary of Item Flags for the Item Pool Calibration (ELA).

Grade	Response Type		Flags									
			A	B	C	D	F	H	P	R	V	Z
3	SR					42	37		116	95		58
	Other	Dichotomous	92	17						56		37
		Polytomous	54	81								6
4	SR		2			29	17		82	85		35
	Other	Dichotomous	88	18						49		33
		Polytomous	29	50						1		2
5	SR		5			43	40		92	90	1	55
	Other	Dichotomous	68	11						56	1	35
		Polytomous	15	23						1		
6	SR					33	28		109	109	3	41
	Other	Dichotomous	107	22						75	5	48
		Polytomous	13	34							5	
7	SR					35	42		100	115	2	50
	Other	Dichotomous	123	29	1					78	6	57
		Polytomous	8	32							3	
8	SR		2			45	50		118	123	12	56
	Other	Dichotomous	113	32	6					94	11	62
		Polytomous	6	27						1	7	2
HS	SR		7			151	169		422	408		156
	Other	Dichotomous	419	114	24					281		184
		Polytomous	7	38	1							2

Table 13. Summary of Items with Flags for the Item Pool Calibration (Mathematics).

Grade	Response Type		Flags									
			A	B	C	D	F	H	P	R	V	Z
3	SR					6	6	1	32	30		11
	Other	Dichotomous	115	26						38	2	35
		Polytomous	15	20						4		6
4	SR					13	13		48	38		18
	Other	Dichotomous	123	19						32	1	41
		Polytomous	30	42	3							12
5	SR		2			22	22		68	63	4	27
	Other	Dichotomous	148	37						24	4	39
		Polytomous	45	52						2		4
6	SR		1			20	12		37	37		24
	Other	Dichotomous	163	45				1		45	11	68
		Polytomous	22	17	3					2	5	7
7	SR		3			19	17		45	40	2	22
	Other	Dichotomous	236	53						54	10	68
		Polytomous	30	23							2	4
8	SR		4			45	31		73	73	1	54
	Other	Dichotomous	254	80	3					44	11	76
		Polytomous	48	56	2					7		19
HS	SR		13			164	132		328	328		156
	Other	Dichotomous	855	268	14			1		182		329
		Polytomous	165	155	5					7		58

Field Test Classical Results

The item and test analyses include the statistics for classical item difficulty (i.e., observed percent of maximum possible score), item discrimination, and reliability. Results are presented primarily for the vertical scaling sample first. This is followed by the presentation of classical item and test results for the calibration sample that represent performance with respect to the entire Smarter Balanced item pool.

Vertical Scaling Results: Classical Item and Test Statistics. The average, item difficulty, and item-total correlation or discrimination are presented in Tables 14 and 15 for the vertical scaling of ELA and mathematics. Item statistics are given for the on-grade and off-grade items sets. Overall, the average item difficulty (observed percentage of the maximum possible score) shows that the items administered were difficult for Field Test administration participants. Most items had item difficulty levels below 0.5. An average item difficulty of 0.5 would indicate that students generally obtained half of the available score points. The most difficult items were in grade 8 and high school (on-grade) in mathematics (0.24). The easiest items were off-grade in grade 4 mathematics (0.51). It also shows the average item discrimination for the on-grade and off-grade items. The average item-test correlations ranged from a low of 0.47 in high school (on-grade) in ELA to a high of 0.62 in grades 4 and 7 (off-grade) in mathematics. The NAEP and PISA items were somewhat easier compared with the Smarter Balanced items. They demonstrated high item-test correlations when the overall IRT ability was used as the criterion.

Figures 1 and 2 compare item difficulty by plotting performance on vertical linking items across grades in ELA and Mathematics. The assumption for the vertical scaling is that in general the items will be easier in the higher-grade level compared with the lower one. The figures show that the items tend to be shifted above the diagonal line indicating that they were easier in the upper grades. There tended to be greater performance differences in grades three and four and less difference in higher-grade levels such as grade 8 and high school. Some items far off the diagonal line indicating performance differences across grades, might be considered as “outliers” and eliminated from the vertical linking. In consultation with the Smarter Balanced Technical Advisory Committee, the decision was made not to eliminate vertical linking items based solely on differences in across-grade item performance. The rationale was that leaving these items in the vertical linking better reflects performance differences across grade levels and how student growth is represented.

Table 16 presents the correlations between the CAT component and the performance tasks for the vertical scaling. The percent of the maximum possible raw score was computed for both the CAT and performance task components. The percent of the maximum possible raw score range is from 0.0 to 1.0. The correlations are across all combinations of the CAT LOFT administrations and different performance tasks for the vertical scaling sample in a grade and content area.

Table 14. Number of Items, Average Item Difficulty, and Discrimination for ELA Vertical Scaling Items.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
On-Grade	No. of Items	261	242	256	232	238	243	410
	Difficulty	0.34	0.35	0.38	0.35	0.34	0.36	0.34
	Discrimination	0.51	0.50	0.52	0.49	0.49	0.49	0.47
Off-Grade Vertical Linking	No. of Items		120	133	131	107	123	107
	Difficulty		0.45	0.45	0.42	0.36	0.38	0.36
	Discrimination		0.54	0.52	0.52	0.51	0.51	0.49
NAEP	No. of Items		28				30	27
	Difficulty		0.55				0.55	0.46
	Discrimination		0.56				0.53	0.54
PISA	No. of Items							33
	Difficulty							0.61
	Discrimination							0.62

Table 15. Number of Items, Average Item Difficulty, and Discrimination for Mathematics Vertical Scaling Items.

Item Purpose		Grade						
		3	4	5	6	7	8	HS
On-Grade	No. of Items	304	306	306	222	239	230	319
	Difficulty	0.39	0.36	0.32	0.30	0.27	0.24	0.24
	Discrimination	0.59	0.58	0.56	0.60	0.59	0.53	0.53
Off-Grade Vertical Linking	No. of Items		104	95	102	71	73	81
	Difficulty		0.51	0.40	0.37	0.32	0.31	0.32
	Discrimination		0.62	0.61	0.58	0.62	0.59	0.56
NAEP	No. of Items		30				33	28
	Difficulty		0.49				0.47	0.41
	Discrimination		0.56				0.57	0.56
PISA	No. of Items							74
	Difficulty							0.41
	Discrimination							0.59

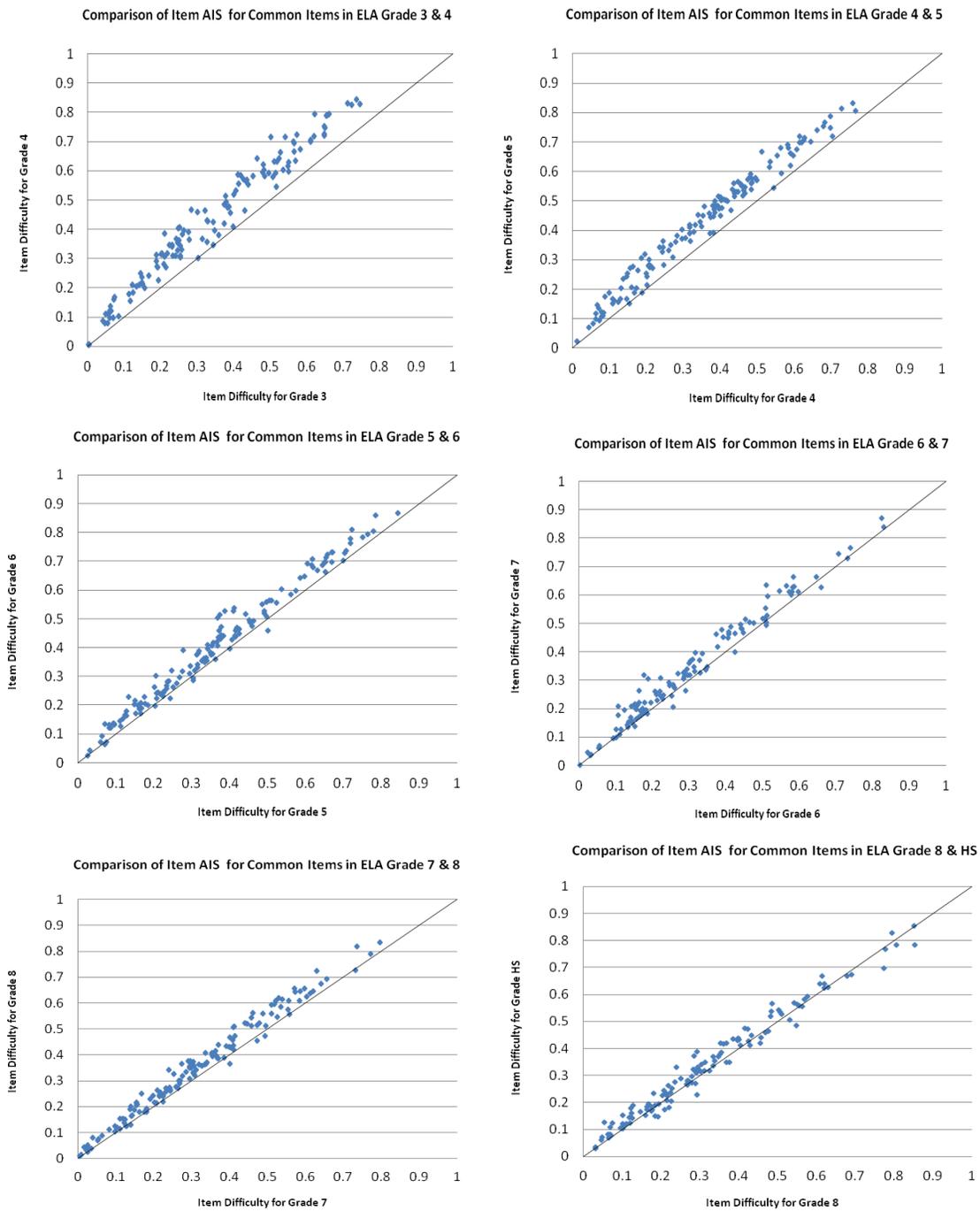


Figure 1. P-value Plots for Vertical Linking Items (ELA) (AIS is used here as association between p-values)

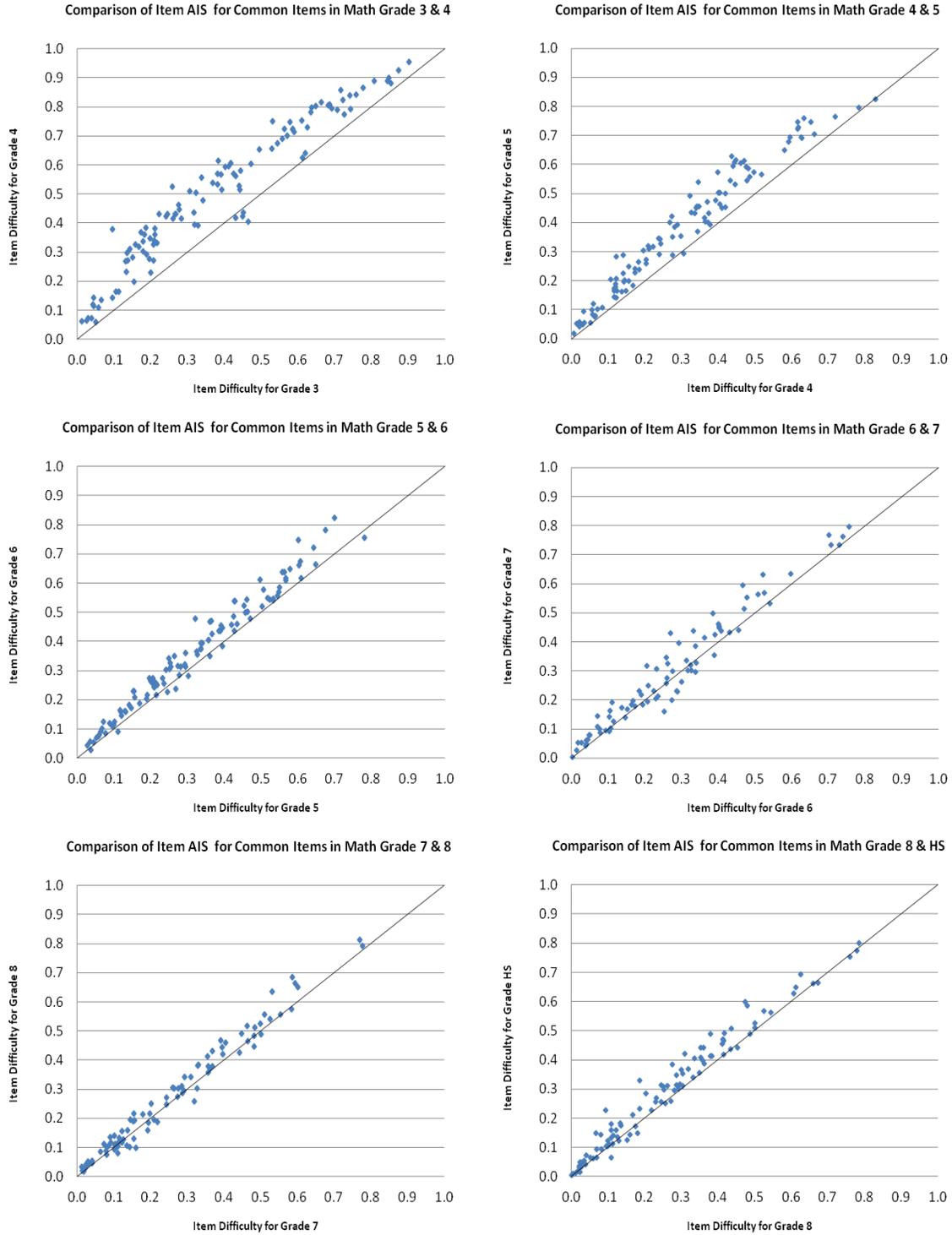


Figure 2. P-value Plots for Vertical Linking Items (Mathematics) (AIS is used here as association between p-values)

Table 16. Pearson Correlation between CAT and Performance Tasks for the Vertical Scaling.

Grade	No. of Students*	Correlation
ELA		
3	18,347	0.55
4	25,613	0.46
5	24,441	0.60
6	24,531	0.61
7	24,248	0.62
8	26,759	0.58
HS	49,392	0.55
Mathematics		
3	20,588	0.60
4	33,025	0.64
5	36,531	0.66
6	25,725	0.60
7	22,230	0.58
8	27,043	0.51
HS	46,877	0.53

*Note: *No. of students refers to the number of students that have valid performance task scores (all students have CAT scores), which may be lower than the counts given in other tables.*

Item Pool Calibration Results: Classical Item and Test Statistics. Table 17 shows the number of items, average item difficulty, and discrimination for the item-pool calibration sample for both ELA and mathematics. This information reflects the item pool combining performance from both performance tasks and CAT items. This Table shows that most students obtained a relatively small portion of the available score points. Items were particularly difficult in mathematics. The average discrimination was high.

Table 17. Number of Items, Average Item Difficulty, and Discrimination for the for Item Pool Calibration.

	Grade						
	3	4	5	6	7	8	HS
ELA							
No. of Items	1,015	948	952	961	1,006	990	3,310
Classical Difficulty	0.33	0.34	0.35	0.32	0.32	0.34	0.33
Classical Discrimination	0.50	0.50	0.50	0.47	0.47	0.46	0.45
Mathematics							
No. of Items	1,162	1,198	1,106	1,107	1,032	1,027	3,311
Classical Difficulty	0.39	0.35	0.28	0.29	0.24	0.22	0.22
Classical Discrimination	0.61	0.60	0.61	0.65	0.66	0.61	0.58

Table 18, which is similar to Table 16, shows the inter-correlations between the CAT and performance task for the item pool calibration. The correlations are of a similar magnitude across grades and content areas and across vertical scaling and item-pool calibration samples.

Table 18. Pearson Correlations between CAT and Performance Tasks for the Item Pool Calibration.

Grade	No. of Students	Correlation
ELA		
3	58,440	0.57
4	56,037	0.55
5	53,280	0.64
6	58,074	0.64
7	56,987	0.66
8	56,960	0.63
HS	114,621	0.58
Mathematics		
3	65,261	0.64
4	66,936	0.65
5	61,457	0.65
6	59,901	0.65
7	60,549	0.61
8	56,133	0.58
HS	116,112	0.56

Test reliability, as expressed by internal consistency, for performance tasks in the item pool calibration is shown in Table 19. Test reliability can be reported since students administered a given performance task all responded to the same set of items. The number of tasks in a grade and content area are presented and the median sample size across that set of performance tasks. The minimum, maximum, average and the standard deviations are presented for Cronbach's alpha and the standard error of measurement (SEM). Reliabilities ranged from 0.07 to 0.79 for ELA and 0.22 to 0.81 for mathematics. Note that there are multiple items or scorable units associated with a given task. In some cases, one or more items might have been dropped from a given task that resulted in very low reliability reported. Note that in the computation of an operational test score both a CAT and performance task will contribute to the overall score. There will likely be a greater number of items

associated with the operational CAT compared with the performance tasks. It is likely that the CAT combined with the performance task will result in sufficient overall levels of reliability.

Table 19. Test Reliability and SEM of Performance Tasks for the Item Pool Calibration.

		Reliability					SEM			
Grade	No. of PT	Median N	Min	Max	Mean	SD	Min	Max	Mean	SD
ELA										
3	19	2,392	0.58	0.72	0.66	0.04	1.04	1.42	1.25	0.11
4	24	1,542	0.07	0.74	0.65	0.13	0.44	1.58	1.36	0.23
5	25	1,401	0.64	0.74	0.69	0.03	1.20	1.51	1.36	0.07
6	20	1,961	0.62	0.79	0.70	0.04	1.15	1.53	1.30	0.11
7	25	1,328	0.64	0.78	0.72	0.03	1.15	1.46	1.26	0.10
8	27	1,251	0.62	0.76	0.70	0.04	1.01	1.63	1.32	0.14
HS	28	2,813	0.61	0.76	0.71	0.04	1.19	1.41	1.31	0.05
Mathematics										
3	24	2,004	0.55	0.79	0.69	0.07	0.93	1.46	1.24	0.14
4	28	1,731	0.53	0.77	0.65	0.06	0.96	1.41	1.18	0.13
5	20	2,234	0.57	0.76	0.66	0.05	0.96	1.61	1.19	0.18
6	30	993	0.58	0.78	0.68	0.05	0.81	1.72	1.19	0.26
7	30	956	0.51	0.78	0.65	0.08	0.53	1.20	0.96	0.15
8	28	946	0.51	0.76	0.64	0.07	0.76	1.51	1.02	0.23
HS	28	2,578	0.22	0.81	0.64	0.14	0.42	1.35	1.01	0.22

Figures 3 and 4 show the distribution of test difficulty for ELA and mathematics for the item pool calibration sample by grade level and across all grades. Using LOFT for the CAT items in a grade and content area, students were administered slightly different numbers and types of items in which the total raw score varied. Students were also administered different performance tasks. In such a design, there are many definitions of a total raw score and test difficulty. As a result, the average percent of maximum is used in a given grade and content area. Since students were administered

different items, test difficulty is the overall raw score divided by the maximum possible score for the collection administered to a given student. This corresponds to the observed percent of the maximum possible raw score including both the CAT and performance task components. A detailed example is given below.

1. A hypothetical student is administered 25 items that contain a mixture of dichotomously and polytomously scored items.
2. The item scores for the student are summed across the 25 items; a total of 45 points were obtained by this student.
3. The maximum possible raw score based on these items is 60 points.
4. The observed percent of maximum for this student is then $45/60$ or 0.75.
5. The distribution of the observed percent of maximum are plotted in the two figures. Each student would have taken essentially a unique set of items that may have varied in item difficulty.

These figures show the tests were difficult for students, which was also reflected by the average item difficulties.

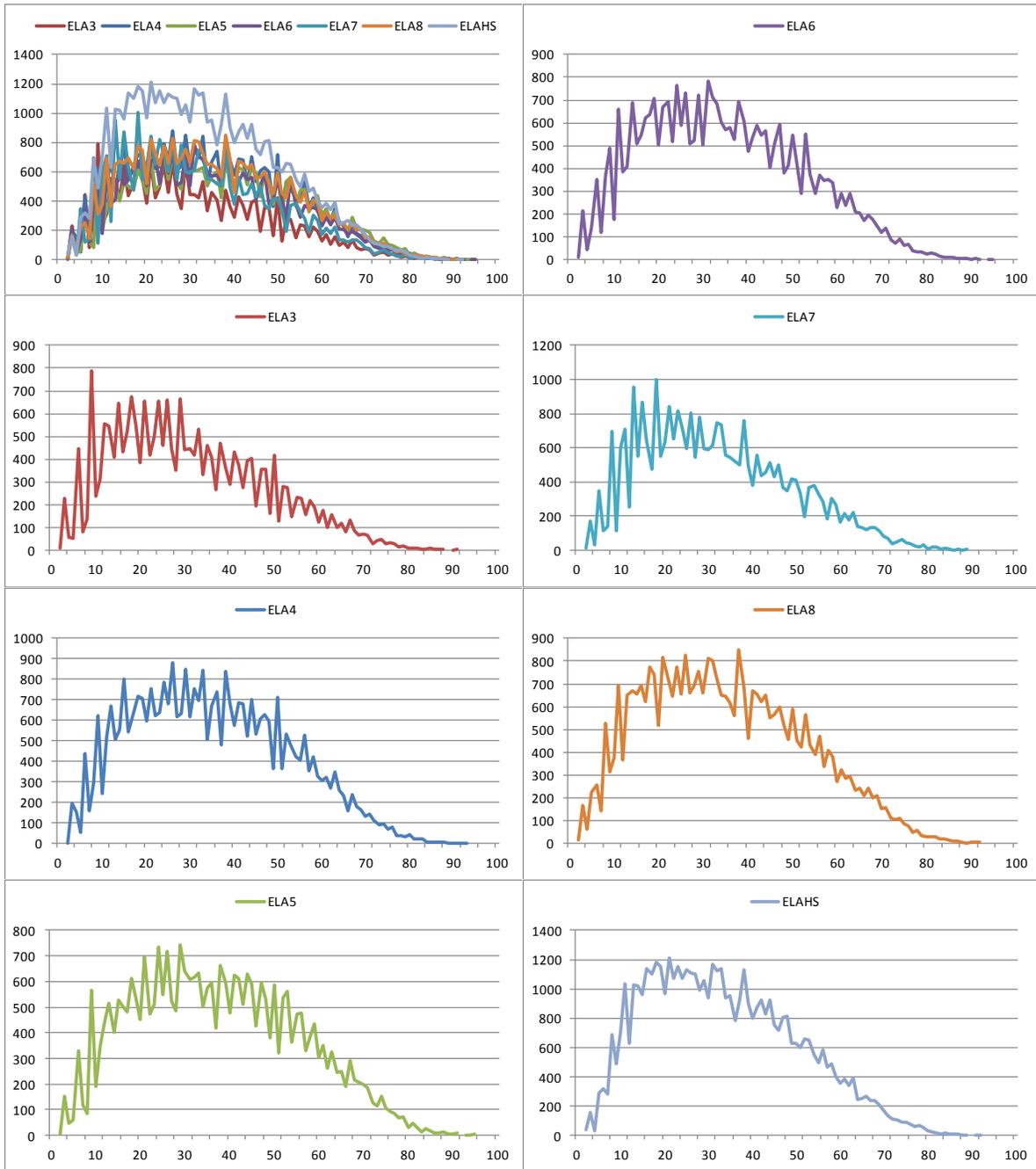


Figure 3. Distributions of Total Raw Scores as a Percentage of the Corresponding Maximum Possible Score for the Item Pool Calibration Sample (ELA)

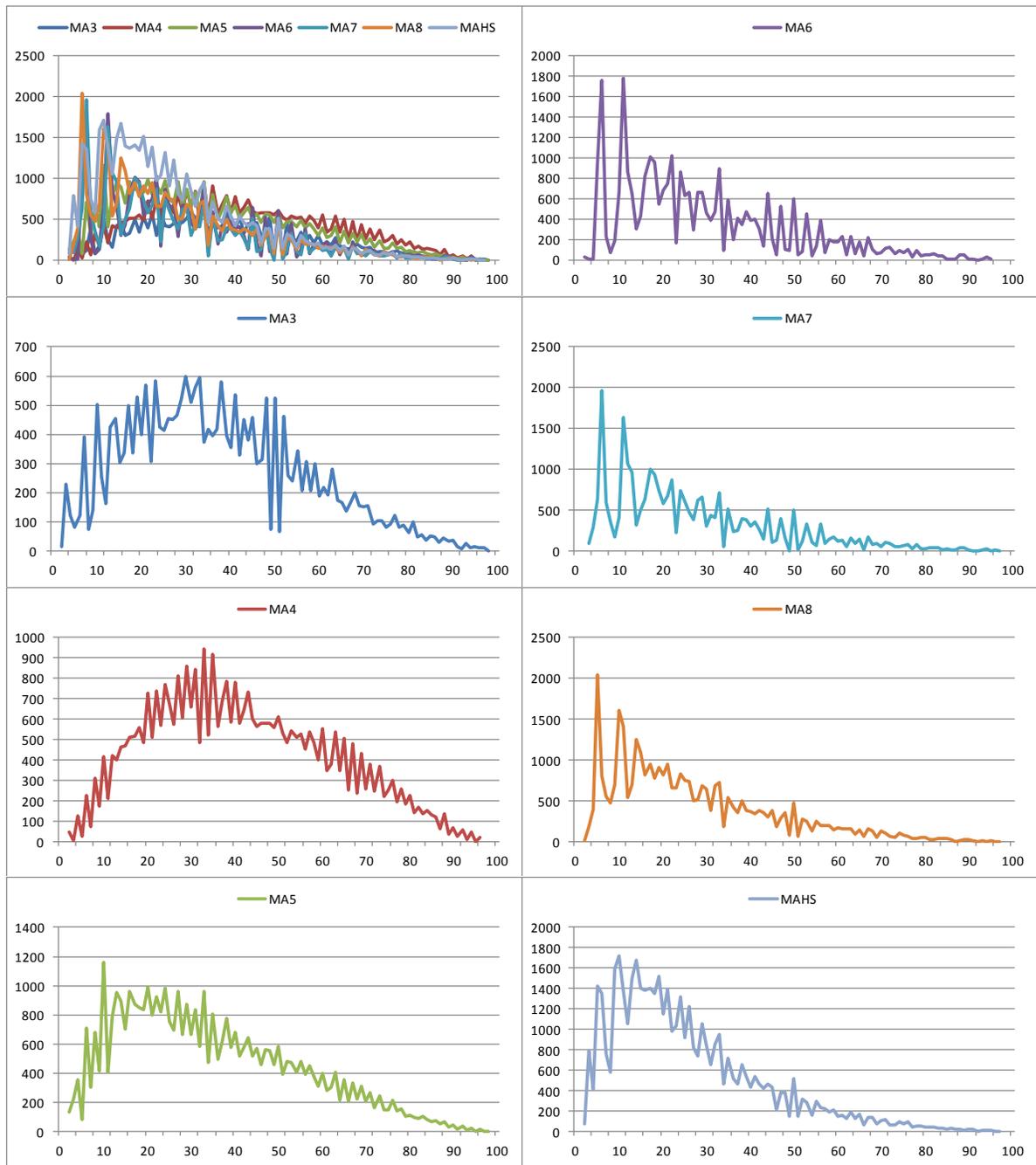


Figure 4. Distributions of Total Raw Scores as a Percentage of the Corresponding Maximum Possible Score for the Item Pool Calibration Sample (Mathematics)

Subgroup Analysis of Test Difficulty for the Item Pool Calibration. Sample size and test difficulty are reported for various subgroups. Test difficulty was defined as the observed percent of the total possible test score. This was computed by taking the overall raw score for a given student and dividing it by the maximum possible raw score. Test difficulty defined here ranges from 0.0 to 1.0. Tables 20 and 21 show average test difficulty for gender, demographic groups, limited English proficiency (LEP), accommodations (Individual Educational Plan: IEP), and Title 1 students for ELA and mathematics.

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

Grade	Subgroup	No. of Students	Average Test Difficulty	SD
3	Female	41,230	0.32	0.16
	Male	42,829	0.29	0.15
	African American	7,799	0.24	0.13
	Asian/Pacific Islander	7,106	0.35	0.17
	Native American/Alaska Native	1,662	0.23	0.12
	Hispanic	24,222	0.25	0.14
	Multiple	4,175	0.30	0.16
	White	39,095	0.34	0.16
	IEP	8,296	0.21	0.13
	LEP	13,886	0.21	0.11
	Title 1	44,640	0.25	0.13
4	Female	45,755	0.35	0.17
	Male	47,767	0.31	0.16
	African American	8,101	0.26	0.14
	Asian/Pacific Islander	7,771	0.39	0.18
	Native American/Alaska Native	2,154	0.24	0.13
	Hispanic	25,158	0.27	0.14
	Multiple	4,795	0.32	0.16
	White	45,543	0.37	0.16

Grade	Subgroup	No. of Students	Average Test Difficulty	SD
	IEP	9,818	0.22	0.14
	LEP	12,720	0.21	0.11
	Title 1	48,323	0.27	0.14
5	Female	42,623	0.38	0.17
	Male	44,167	0.33	0.16
	African American	8,218	0.29	0.15
	Asian/Pacific Islander	6,817	0.42	0.18
	Native American/Alaska Native	1,752	0.26	0.14
	Hispanic	23,262	0.30	0.15
	Multiple	4,418	0.35	0.17
	White	42,323	0.40	0.16
	IEP	9,679	0.22	0.13
	LEP	9,619	0.22	0.11
	Title 1	43,796	0.30	0.15
	6	Female	45,094	0.34
Male		46,652	0.30	0.15
African American		8,976	0.26	0.14
Asian/Pacific Islander		6,820	0.39	0.17
Native American/Alaska Native		2,138	0.24	0.13
Hispanic		25,012	0.26	0.14
Multiple		4,422	0.31	0.16
White		44,378	0.35	0.16

Grade	Subgroup	No. of Students	Average Test Difficulty	SD
	IEP	9,801	0.19	0.11
	LEP	8,808	0.18	0.10
	Title 1	46,642	0.26	0.14
7	Female	44,517	0.34	0.16
	Male	46,019	0.30	0.15
	African American	8,269	0.26	0.13
	Asian/Pacific Islander	7,397	0.39	0.17
	Native American/Alaska Native	2,068	0.24	0.12
	Hispanic	28,357	0.27	0.14
	Multiple	3,983	0.33	0.16
	White	40,462	0.36	0.16
	IEP	9,007	0.19	0.11
	LEP	8,666	0.18	0.09
	Title 1	47,399	0.27	0.14
8	Female	46,616	0.36	0.16
	Male	48,654	0.31	0.16
	African American	9,630	0.28	0.14
	Asian/Pacific Islander	7,284	0.40	0.18
	Native American/Alaska Native	2,163	0.26	0.14
	Hispanic	25,194	0.29	0.14
	Multiple	4,059	0.34	0.16
	White	46,940	0.37	0.16

Grade	Subgroup	No. of Students	Average Test Difficulty	SD
	IEP	9,464	0.20	0.12
	LEP	6,987	0.19	0.10
	Title 1	46,063	0.29	0.15
HS	Female	116,646	0.35	0.16
	Male	117,297	0.30	0.16
	African American	21,824	0.26	0.14
	Asian/Pacific Islander	21,973	0.39	0.18
	Native American/Alaska Native	3,443	0.28	0.14
	Hispanic	71,245	0.28	0.15
	Multiple	8,344	0.33	0.17
	White	107,114	0.35	0.16
	IEP	17,934	0.19	0.12
	LEP	14,881	0.18	0.09
	Title 1	109,507	0.28	0.15

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

Grade	Subgroup	No. of Students	Item Difficulty	SD
3	Female	45,600	0.37	0.19
	Male	46,939	0.37	0.20
	African American	8,445	0.28	0.16
	Asian/Pacific Islander	7,348	0.46	0.21
	Native American/Alaska Native	1,904	0.28	0.16
	Hispanic	27,423	0.30	0.17
	Multiple	4,552	0.37	0.20
	White	42,867	0.42	0.19
	IEP	9,236	0.25	0.18
	LEP	16,361	0.26	0.16
	Title 1	49,037	0.31	0.17
4	Female	52,827	0.36	0.19
	Male	54,969	0.37	0.20
	African American	9,420	0.26	0.16
	Asian/Pacific Islander	8,101	0.45	0.22
	Native American/Alaska Native	3,347	0.26	0.16
	Hispanic	28,703	0.28	0.17
	Multiple	6,228	0.36	0.20
	White	51,997	0.41	0.19
	IEP	11,645	0.23	0.17
	LEP	14,337	0.23	0.14
	Title 1	56,022	0.30	0.17

Grade	Subgroup	No. of Students	Item Difficulty	SD
5	Female	52,355	0.29	0.18
	Male	54,871	0.30	0.20
	African American	8,203	0.20	0.14
	Asian/Pacific Islander	7,877	0.39	0.22
	Native American/Alaska Native	3,162	0.19	0.14
	Hispanic	27,072	0.22	0.15
	Multiple	6,164	0.30	0.19
	White	54,748	0.34	0.19
	IEP	11,851	0.17	0.15
	LEP	11,264	0.16	0.11
	Title 1	53,518	0.23	0.16
	6	Female	56,975	0.27
Male		58,624	0.27	0.19
African American		8,629	0.19	0.14
Asian/Pacific Islander		10,229	0.38	0.22
Native American/Alaska Native		1,472	0.19	0.15
Hispanic		37,395	0.21	0.15
Multiple		6,236	0.28	0.19
White		51,638	0.32	0.19
IEP		12,344	0.14	0.13
LEP		13,138	0.14	0.11
Title 1		60,158	0.21	0.15

Grade	Subgroup	No. of Students	Item Difficulty	SD
7	Female	55,007	0.23	0.17
	Male	56,723	0.23	0.18
	African American	8,577	0.15	0.13
	Asian/Pacific Islander	10,308	0.33	0.21
	Native American/Alaska Native	1,265	0.17	0.13
	Hispanic	39,425	0.17	0.13
	Multiple	5,100	0.24	0.17
	White	47,055	0.27	0.18
	IEP	11,098	0.12	0.11
	LEP	12,188	0.11	0.10
	Title 1	59,209	0.18	0.14
8	Female	54,758	0.22	0.16
	Male	56,054	0.22	0.17
	African American	8,330	0.15	0.12
	Asian/Pacific Islander	10,013	0.31	0.20
	Native American/Alaska Native	1,316	0.17	0.13
	Hispanic	35,537	0.16	0.12
	Multiple	5,168	0.23	0.17
	White	50,448	0.25	0.17
	IEP	10,639	0.12	0.10
	LEP	10,307	0.11	0.09
	Title 1	55,026	0.17	0.13

Grade	Subgroup	No. of Students	Item Difficulty	SD
HS	Female	112,663	0.20	0.15
	Male	112,092	0.21	0.16
	African American	20,772	0.14	0.11
	Asian/Pacific Islander	22,132	0.31	0.21
	Native American/Alaska Native	3,370	0.16	0.12
	Hispanic	70,446	0.15	0.12
	Multiple	8,227	0.20	0.16
	White	99,808	0.23	0.16
	IEP	16,684	0.11	0.09
	LEP	16,621	0.11	0.09
	Title 1	105,246	0.16	0.12

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

In addition to classical item and test analyses, differential item functioning (DIF) analyses were also performed on the Field Test items. DIF analyses are used to identify those items that identify groups of students (e.g., males versus females) with the same underlying level of ability that have different probabilities of answering an item correctly. To perform a DIF analysis, students are separated into relevant subgroups based on ethnicity, gender, or other demographic characteristics. Students in each subgroup are then ranked relative to their total test score (conditioning on ability). Item performance from the focal group to be examined (e.g., females) is compared conditionally based on ability with the reference group (e.g., males). The definitions for the focal and reference groups used are given in Table 22. A DIF analysis asks, “If we compare focal-group and reference-group students of the same overall ability (as indicated by their performance on the full test), are any test items appreciably more difficult for one group compared with another group?” DIF in this context is viewed as a potential source of invalidity.

DIF statistics are used to identify items that are *potentially* functioning differentially. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences. If the item is differentially more difficult for an identifiable subgroup when conditioned on ability, it may be measuring something different from the intended construct to be measured. However, it is important to recognize that DIF-flagged items might be related to actual differences in relevant knowledge or statistical Type I error.

Table 22. Definition of Focal and Reference Groups.

Group Type	Focal Groups	Reference Groups
Gender	Female	Male
Ethnicity	African American	White
	Asian/Pacific Islander	
	Native American/Alaska Native	
	Hispanic	
Special Populations	Limited English Proficient (LEP)	English Proficient
	Individualized Education Program (IEP)	No IEP
	Title 1	Not Title 1

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

DIF Category	Definition
A (negligible)	Absolute value of the MH D-DIF is not significantly different from zero, or is less than one.
B (slight to moderate)	<ol style="list-style-type: none"> 1. Absolute value of the MH D-DIF is significantly different from zero but not from one, and is at least one; or 2. Absolute value of the MH D-DIF is significantly different from one, but less than 1.5. 3. Positive values are classified as “B+” and negative values as “B-”
C (moderate to large)	<ol style="list-style-type: none"> 1. Absolute value of the MH D-DIF is significantly different from 1, and is at least 1.5; and 2. Absolute value of the MH D-DIF is larger than 1.96 times the standard error of MH D-DIF. 3. Positive values are classified as “C+” and negative values as “C-“

Table 24. DIF Flagging Logic for Constructed-Response Items

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

Items are classified into three DIF categories of “A,” “B,” or “C.” DIF Category A items contain negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large values of DIF. Positive values favor the focus group, and negative values are in favor of the reference group. The positive and negative values are reported for C-DIF item flagging. DIF analyses were not conducted if the sample size for either the reference group or the focal group was less than 100 or if the sample size for the two combined groups was less than 400. In subsequent tables, A levels of DIF are not flagged as they are too small to have perceptible interpretation.

Different DIF analysis procedures are used for dichotomous items (items with 0/1 score categories; selected-response items) and polytomous items (items with more than two score categories; constructed-response items). Statistics from two DIF detection methods are computed consisting of the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) and the standardized mean difference (SMD) procedure (Dorans & Kulick, 1983, 1986). Selected-response items are classified into DIF categories of A, B, and C, as described in Table 30.

For dichotomous items, the statistic described by Holland and Thayer (1988), known as Mantel-Haenszel D-DIF, is reported. This statistic is reported on the delta scale, which is a normalized transformation of item difficulty (p-value) with a mean of 13 and a standard deviation of 4. Items that are not significantly different based on the Mantel-Haenszel D-DIF ($p > 0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. For items where the statistical test indicates significant differences ($p < 0.05$), the effect size is used to determine the direction and severity of the DIF. The formula for the estimate of constant odds ratio is

$$\alpha_{MH} = \frac{\left(\frac{\sum_m R_{rm} W_{fm}}{N_m} \right)}{\left(\frac{\sum_m R_{fm} W_{rm}}{N_m} \right)},$$

where

R_{rm} = number in reference group at ability level m answering the item right;

W_{fm} = number in focal group at ability level m answering the item wrong;

R_{fm} = number in focal group at ability level m answering the item right;

W_{rm} = number in reference group at ability level m answering the item wrong; and

N_m = total group at ability level m .

This value can then be used as follows (Holland & Thayer, 1988):

$$MH\ D-DIF = -2.35 \ln[\alpha_{MH}].$$

The Mantel-Haenszel chi-square statistic used to classify items into the three DIF categories is

$$MH\ CHISQ = \frac{(\sum_m R_{rm} - \sum_m E(R_{rm}))^2}{\sum_m Var(R_{rm})},$$

where $E(R_{rm}) = N_{rm}R_{Nm} / N_m$, $Var(R_{rm}) = \frac{N_{rm}N_{jm}R_{Nm}W_{Nm}}{N_m^2(N_m - 1)}$, N_{rm} and N_{jm} are the numbers of examinees in the

reference and focal groups, respectively, R_{Nm} and W_{Nm} are the number of examinees who answered the item correctly and incorrectly, respectively. The classification logic used for flagging items is based on a combination of absolute differences and significance testing. Items that are not statistically different based on the MH D-DIF ($p > 0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. For items where the statistical test indicates significant differences ($p < 0.05$), the effect size is used to determine the direction and severity of the DIF. The classification logic for selected-response items is based on a combination of absolute differences and significance testing, is shown in Table 23.

The standardized mean difference compares item performance of two subpopulations adjusting for differences in the distributions of the two subpopulations. The standardized mean difference statistic can be divided by the total standard deviation to obtain a measure of the effect size. A negative value of the standardized mean difference shows that the item is more difficult for the focal group, whereas a positive value indicates that it is more difficult for the reference group. The standardized mean difference used for polytomous items is defined as:

$$SMD = \sum p_{Fk}m_{Fk} - \sum p_{Rk}m_{Rk},$$

where p_{Fk} is the proportion of the focal group members who are at the k^{th} level of the matching variable, m_{Fk} is the mean score for the focal group at the k^{th} level, and m_{Rk} is the mean item score for the reference group at the k^{th} level. The standardized mean difference is divided by the total item group standard deviation to get a measure of the effect size. The classification logic for polytomous items is based on a combination of absolute differences and significance testing, as shown in Table 24. Items that are not statistically different based on the MH D-DIF ($p > 0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately.

A relatively small number of items showed some performance differences between student groups as indicated by C-DIF flagging criteria. Tables 25 and 26 show the number of items flagged for all categories of DIF for ELA/literacy and mathematics in grades 3 to 11. Note that the item flagging incorporates items that were administered across grades for vertical linking. A relatively small percentage of items were flagged for significant levels of DIF (C-DIF) in the Field Test for the collective item pool. All items had previously undergone bias reviews. Additional inspection of these C-DIF items was conducted by content editors before inclusion in operational tests administrations. Items with A level of DIF are not flagged because the level is too low to interpret meaningfully.

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

Focal Group Category									
Grade	DIF Flag	Female	Asian	African American	Hispanic	Native American	IEP	LEP	Title1
3	C+	4	5						
	C-		1		2			1	
	B	28	45	30	23	6	21	19	3
4	C+	8	7	2			2	1	
	C-	2	6	1	3			3	
	B	36	40	24	23	7	22	21	9
5	C+	18	5	3	1			2	
	C-	2		1	2		3	2	
	B	60	40	24	32	7	17	21	9
6	C+	6	11					1	
	C-	3	7	1	4		2	3	
	B	47	44	21	23	8	14	23	6
7	C+	7	8	2					
	C-	2	2	4	1				
	B	70	48	25	22	12	14	21	4
8	C+	16	12	1	3				
	C-	4	5	2	3		2	5	
	B	70	48	29	39	8	17	32	7
HS	C+	10	15	2	4		3	5	3
	C-	20	19	13	30	1	3	8	11
	B	180	161	77	138	12	60	73	74

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

Focal Group Category									
Grade	DIF Flag	Female	Asian	African American	Hispanic	Native American	IEP	LEP	Title1
3	C+	1	21	5	2		2	4	
	C-		5	1	1		1	3	
	B	14	74	80	58	1	22	39	1
4	C+	1	16	7	3	1		2	
	C-		3	2	2		1	5	
	B	25	73	40	41	14	18	41	4
5	C+		17			3	1		
	C-		5		1	1	2	5	
	B	22	61	43	9	15	21	27	3
6	C+	2	31	4	4				
	C-	2	5	3	1				
	B	29	49	18	19		7	21	7
7	C+	2	24	2			2	2	
	C-		4	1			1		
	B	27	66	19	18		26	19	7
8	C+	1	13	3			5	2	
	C-	1	6	1	2		1		
	B	11	46	22	22		24	22	2
HS	C+	10	46	4	7		1	4	4
	C-	5	2	4				2	
	B	76	60	57	59		26	18	22

Prospective Evidence in Support of Rater Agreement

Since CR items and performance tasks are an integral part of the score, it is important to establish that the results are consistent across raters and task types. This can be accomplished by evaluating the results of the Field Test administration and through careful management of the scoring processes, minimizing all possible sources of variance associated with these procedures. In order to minimize any sources of irrelevant variance, a comprehensive set of plans for evaluating and monitoring the scoring systems was implemented. The procedures described below provide a basis for monitoring whether the score categories and the underlying construct are maintained consistently in the Field Test and subsequent administrations.

Monitoring Scoring Processes. Pre-Field Test scoring procedures consist of range-finding, selection of calibration/benchmark papers, and the establishment of materials for rater training and qualification. Well-developed processes and procedures in the pre-operational phase determine the success of the operational phase. These processes include the following.

- Certification and training. Each qualified rater receives rigorous training in correctly applying the rubric at each specific score point and are required to successfully complete a certification test.
- Automated scoring. Automated scoring was implemented such that the scoring engines have to be established and “trained” to score the targeted items using a requisite number of student responses with known psychometric properties.
- Range-finding and calibration papers. Calibration papers with known psychometric properties are selected by experts and establish the standard for scoring various types of responses—these are also known as “benchmark” papers. These papers are distributed periodically during the course of scoring and are critical to determining the accuracy of scoring. In the pre-operational phase, it is critical that large and robust calibration papers are able to be selected.
- Operational. Operational scoring procedures include monitoring the method of distributing student responses, as well as real-time monitoring of rater accuracy and consistency and supervisory review and auditing.

Monitoring Raters and Associated Statistics. The statistics and methods used for monitoring rater agreement for evaluating the functioning of performance tasks may include, but are not limited to:

- number and proportion of students earning each rubric score;
- number and percentage of exact agreement between two human ratings or between automated and human scores after correcting for chance agreement rates;
- number and percentage of adjacent agreement between two human ratings or between automated and human scores after correcting for chance agreement rates;
- number and percentage of non-adjacent scores between two human ratings or between automated and human scores after correcting for chance agreement rates;
- unweighted Kappa statistics (Cohen, 1960), which characterize the degree of agreement or association between two human ratings or between automated and human scores after correcting for chance agreement rates;
- quadratic-weighted Kappa statistics (Fleiss, 1981), which have similar properties to unweighted Kappa, but are degraded disproportionately by the presence of large

disagreements between ratings of two human raters or between human and automated scores; and

- Pearson correlations, which provide another measure of the degree of agreement or association between two human ratings or between automated and human scores.

Monitoring Automated Scoring and Associated Statistics. Some validation of scoring, even for automated algorithms, is also necessary. Consistent with the procedures used with rater protocols for monitoring reliability, the following statistics can be produced.

- Similarity of human and automated score frequency distributions and means with standard deviations.
- Standardized differences (effect sizes) between human and automated score means. This is computed as the difference between means divided by the standard deviation of the human scores.
- Unweighted Kappa statistics (Cohen, 1960), which characterize the degree of agreement or association between automated and human scores after correcting for chance agreement rates.
- Quadratic-weighted Kappa statistics (Fleiss, 1981), which have similar properties to unweighted Kappa statistics but are degraded disproportionately by the presence of large disagreements between some human and automated scores.
- Pearson correlations provide another measure of agreement or association between automated and human scores.

External Assessments: NAEP and PISA.

Smarter Balanced established achievement levels with respect to the Consortium while also wanting to reference important national or international assessments such as NAEP and PISA. Inferences concerning relative performance on these items relied on assumptions concerning factors like test-delivery-mode effects, item-context effects, the impact of different testing windows and years, and the impact of different test purposes. The NAEP mathematics, reading or writing, and PISA literacy content and skills frameworks are also quite different from Smarter Balanced. Finally, NAEP and PISA data both derive from paper-based administrations, while Smarter Balanced assessments are computer-delivered (NAEP and PISA both plan computer administrations for 2015). Table 27 summarizes some high-level features of Smarter Balanced, NAEP, and PISA programs for purposes of comparison. This is followed by a brief narrative description of each program.

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

Design Feature	Smarter Balanced	NAEP	PISA
Construct Definition	ELA/literacy Claims—Reading, Writing, Listening, & Research Text Types: Literary & Information	Reading Frameworks: (Writing is separate.) Text Types: Literary & Information	Reading Aspects: Text Types: Exposition, Argumentation, Instruction, Transaction, & Description
	Math Claims—Concepts and Procedures, Problem solving, Model and Data Analysis, Communicating Reasoning	Math Frameworks: Number Properties and Operations, Measurement, Geometry, Data Analysis, Statistics and Probability, and Algebra	Math Aspects: Quantity, Uncertainty, Space & Shape, Change & Relationships
Item Context Effects and Test Administration Rules	<ul style="list-style-type: none"> The basic context will be maintained for NAEP and PISA items since they are administered as a set(s). The look and feel of NAEP and PISA item will likely be different from Smarter Balanced. The provision of glossaries, test manipulatives, and accommodation rules differ across programs. Smarter Balanced uses technology-enhanced items, while PISA and NAEP do not. 		
Testing Delivery Modes	LOFT delivery on computer and performance tasks online	Paper 2015: Paper Scale and Computer-based Testing Scale Study	Paper 2015: Computer-based Testing Scale
Testing Window	March–June	February	April/May
Untimed/Timed	Untimed	Timed	Timed
Delivery Design	<ul style="list-style-type: none"> Smarter Balanced Field Test LOFT blueprint(s) took into consideration the embedded set(s) properties such as testing length, reading load, and associated number of items. 		
Constructed-Response Scoring	<ul style="list-style-type: none"> Human scoring for external NAEP/PISA items was required. Approximately 30 percent of the PISA items and 30 to 40 percent of NAEP items were associated with sets requiring rater scoring. Scoring protocols such as training and qualification will need to be followed. Handwritten responses would need to be transcribed for anchors, training and qualification, and calibration papers. 		

Design Feature	Smarter Balanced	NAEP	PISA
Cohort/Population	Sample of 2014 Smarter Balanced Governing States	Based on 2013 U.S. national sample with state-level comparisons	Based on 2012 U.S. sample: 5,000 15-year-old students from 150 schools
Criterion-referenced Inferences	Designated achievement-level scores in 2014	Proficiency cut scores exist	Proficiency cut scores do not exist.
Anticipated Program Changes	No change after 2014 in content; schools still transitioning to the CCSS	Transitioning to computer in 2015 and math content domains will change.	Computer-based in 2015, and assessment framework will change.
IRT Model and Scaling Procedures	Scaling is at the overall content area level using the 2-PL/generalized partial credit model (GPCM).	3-PL and GPCM in reading and math; the main scales are weighted composites of subscales, and calibration is done at the subscale level.	Rasch (calibrated separately with relation to major domain and minor domain).
Anchor Item Requirements	<ul style="list-style-type: none"> • Construct-representative anchor sets were used. • More than one item block (test form) was implemented. 		

PISA Overview. The Program for International Student Assessment (PISA) is an international assessment that measures 15-year-old students' reading, mathematics, and science literacy. PISA also includes measures of general or cross-curricular competencies, such as problem solving. PISA is coordinated by the Organisation for Economic Cooperation and Development (OECD), an intergovernmental organization of industrialized countries, and is conducted in the United States by the National Center for Education Statistics (NCES). PISA emphasizes functional skills that students have acquired as they near the end of compulsory schooling. PISA was first administered in 2000 and is conducted every three years. The most recent assessment was in 2012. PISA 2012 assessed students' mathematics, reading, and science literacy. PISA 2012 also included computer-based assessments in mathematics literacy, reading literacy, and general problem solving, as well as an assessment of students' financial literacy. Results for the 2012 mathematics, reading, science, and problem-solving assessments are currently available.

NAEP Overview. The National Assessment of Educational Progress (NAEP) is a continuing and nationally representative assessment of what our nation's students know and can do. There are two types of NAEP assessments: the main NAEP and the long-term strand NAEP. The main NAEP was utilized for the Smarter Balanced Field Test. It is administered to fourth-, eighth-, and twelfth-grade students across the country in a variety of subjects. National results are available for all assessments and subjects. Each NAEP assessment is built from a content framework that specifies the types of knowledge and skill that students are expected to know in a given grade. When assessing performance for the nation only, approximately 6,000 to 20,000 students per grade from across the country are assessed for each subject. A sampling procedure is used to ensure that those selected to participate in NAEP will be representative of the geographical, racial, ethnic, and socioeconomic diversity of schools and students across the nation. NAEP is not designed to report

individual test scores, but rather to produce estimates of scale score distributions for groups of students.

Results

The subset of students who took NAEP and PISA items also took Smarter Balanced CAT items and performance tasks. A summary of the resulting item performance for NAEP and PISA and all Smarter Balanced students are presented in Table 28 for ELA/literacy and Mathematics showing the number of items administered, mean item difficulty (i.e., p-values) and discrimination. Figure 5 shows the p-values for NAEP items plotted against the ones obtained from the Smarter Balanced vertical scaling sample. The graphs suggest a reasonably linear relationship. The NAEP p-values are relatively higher (i.e., easier) than the ones obtained from the vertical scaling sample. There are other factors such as mode effects (i.e., on-line vs. paper) that might also account for these performance differences. It was not possible to obtain similar sorts of item difficulty information from the PISA program.

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

Item Purpose		Grade						
		3	4	5	6	7	8	HS
ELA/literacy								
NAEP	No. of Items		28				30	27
	Difficulty		0.55				0.55	0.46
	Discrimination		0.56				0.53	0.54
PISA	No. of Items							33
	Difficulty							0.61
	Discrimination							0.62
Mathematics								
NAEP	No. of Items		30				33	28
	Difficulty		0.49				0.47	0.41
	Discrimination		0.56				0.57	0.56
PISA	No. of Items							74
	Difficulty							0.41
	Discrimination							0.59

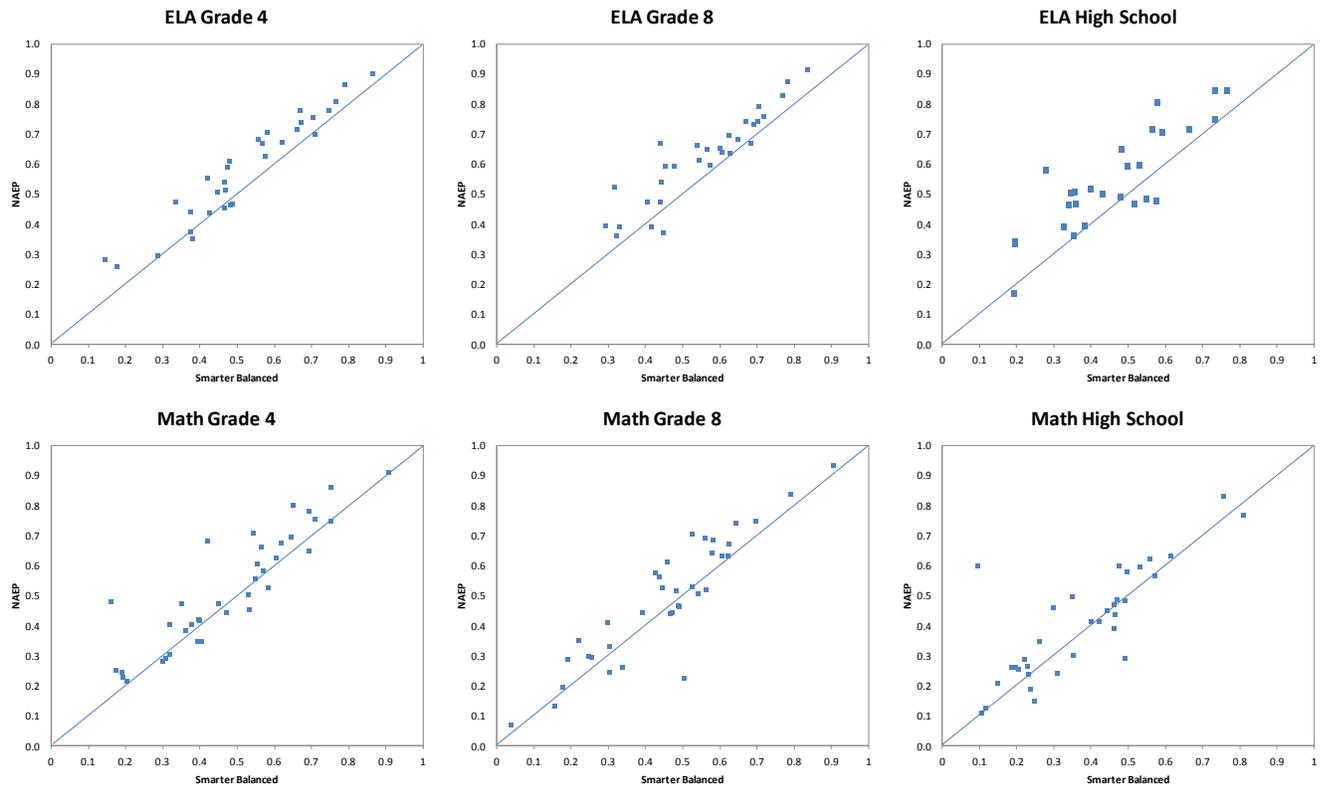


Figure 5. Comparison of NAEP Item Difficulty and Values Obtained from Smarter Balanced Samples

References

- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*, 37-46.
- Dorans, N. J., & Kulick, E. (1983). *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach* (RR-83-9). Princeton, NJ: Educational Testing Service.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.
- Dragow, F. (1988). Polychoric and Polyserial Correlations. In L. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences, 7*, 69-74. New York: Wiley.
- Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. 2nd ed. (New York: John Wiley) pp. 38-46.
- Holland, P. W., & Thayer, D. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Mantel, N., & Haenszel, W. M. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute, 22*, 719-748.
- Olsson, U. (1979). Maximum Likelihood Estimation of the Polychoric Correlation Coefficient. *Psychometrika, 4*, 443-460.