



Answers to Questions about Smarter Balanced 2017 Test Results

March 27, 2018

© Smarter Balanced Assessment Consortium, 2018



Table of Contents

Table of Contents	1
Background	2
Jurisdictions included in Studies	2
How did students perform in 2017 compared to 2016?	3
Overall Trend in Mean Scale Score	3
Trends in Mean Scale Score by Grade	4
Trends in Percent Proficient	5
Were there fewer test questions available?	6
Assessment Structure.....	6
Item Pool Changes over Time	8
Differences between Item Groups	8
Similarities between Item Groups.....	9
Did students receive more difficult test questions in 2017 compared to previous years?	9
Statistical Differences between Item Groups	9
Differences between Old and New CAT Items by Decile of Student Achievement.....	10
Measurement Precision	14
Standard Error of Measurement (SEM).....	14
Expected Scores.....	16
Did the newly-added test questions impact test results?	17
Data and Method	17
Item Counts, Exposure, and Mean Residuals by Item Group and Grade	17
Item Counts, Exposure, and Mean Residuals by Decile within Grade	20
Practical Impact	21
Did students spend less time taking the test?	24
Did students take the test earlier in the school year?	25
Did the student demographics change?	26
Summary and Conclusions	28

Background

This report provides an update on analyses performed to investigate the comparison in student achievement on Smarter Balanced summative tests from 2016 (spring, 2016) to 2017 (spring, 2017). This comparison raised questions among educators about the validity of the 2017 test scores and the possibility that there might have been technical errors associated with changes in the item pool and other components of the administration. Subsequently, Smarter Balanced worked with its Technical Advisory Committee (TAC) to identify analyses that could be performed to address educator’s questions and related technical issues. The analyses addressed the following questions.

- 1) How did students perform in 2017 compared to 2016?
- 2) Were there fewer test questions available?
- 3) Did students receive more difficult test questions in 2017 compared to previous years?
- 4) Did the newly-added test questions impact test results?
- 5) Did students spend less time taking the test?
- 6) Did students take the test earlier in the school year?
- 7) Were the student demographics different?

This report is organized around answers to these questions. In addition, there is a section that describes the measurement precision of the 2016 and 2017 tests.

Jurisdictions included in Studies

The studies were based on four sets of member jurisdictions shown in Table 1. Due to time constraints, differences in data requirements for the various studies, and to differences in when data from various jurisdictions became available, the questions above were answered with varying numbers of jurisdictions, grades per jurisdiction, and student inclusion rules. Grade 11 data from DE was not available for any analysis that included DE. All four samples were representative of the consortium and therefore showed similar trends in student achievement from 2015 to 2017 in each subject overall and within grade.

Table 1. Jurisdictions Used in Studies.

Jurisdiction	Sample A	Sample B	Sample C	Sample D
BIE				
California	1	1	1	1
Connecticut	1			
Delaware	1	1	1	1
Hawaii	1	1	1	1
Idaho	1	1	1	1
Michigan				
Montana	1	1	1	1
Nevada				
New Hampshire	1			
North Dakota	1			1
Oregon	1	1	1	1
South Dakota	1	1	1	1
Vermont	1	1	1	1
Virgin Islands	1	1		
Washington	1	1	1	1
West Virginia	1			
Total number:	14	10	9	10

The samples were used to derive information for this report as follows:

- Sample A
 - Consortium-level mean scale scores and percent proficient by year (2015, 2016, and 2017), and change from year to year. (Tables 2 and 3. Figures 1 and 2).
- Sample B
 - 2017 expected scores by decile (Figure 11)
 - Residual analysis (Tables 10 to 13).
- Sample C
 - 2016 expected scores by decile (Figure 11)
- Sample D
 - 2016 and 2017 means and standard deviations of test scores, differences, and effect size of differences Tables 4 and 5.
 - Percent proficient in 2016 and 2017 and change by grade. (Table 6).
 - Student deciles for old and new CAT item counts and item discrimination in 2017. (Figures 3 through 8).
 - Standard error of measurement by decile in 2016 and 2017 (Figures 9 and 10).
 - Time students spent taking the test in 2016 and 2017 (Table 14).
 - Test start dates in 2016 and 2017 (Table 15).
 - Student demographics in 2016 and 2017 (Table 16).

How did students perform in 2017 compared to 2016?

Overall Trend in Mean Scale Score

Trends in the mean scale score of students in Sample A, grades 3 to 8, over the three years of operational testing from 2015 to 2017, are shown in Figure 1. In mathematics, achievement increased by 6.0 points in 2016 and by 0.7 points in 2017. In English language arts/literacy (ELA/literacy), achievement increased by 7.3 points in 2016 and declined by 1.7 points (-1.7) in 2017. These details are shown in Table 2.

Figure 1. Three-year Trends in Smarter Balanced Test Scores.

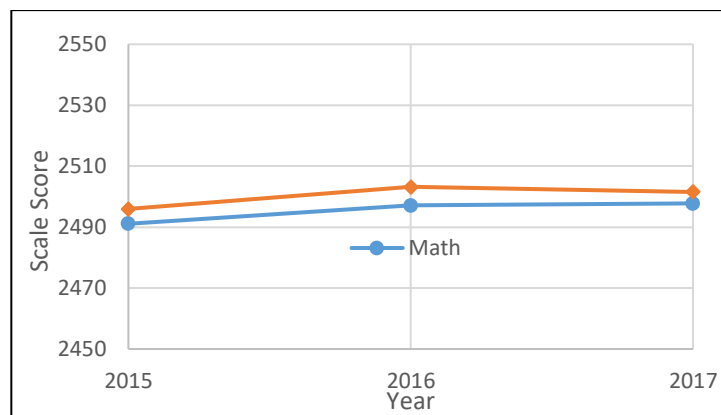


Table 2. Sample A Detail for Points Plotted in Figure 1.

Year	Math Mean	Math Change	ELA/L Mean	ELA/L Change
2015	2491.1		2496.0	
2016	2497.1	6.0	2503.2	7.3
2017	2497.8	0.7	2501.6	-1.7
Net 3-year change:		6.7		5.6

Corresponding changes in the percent of students classified as ‘proficient’ are shown in Table 3. “Proficient” is a term applied to students at achievement levels 3 or 4 on the Smarter Balanced assessments. In 2016, the percent proficient increased by over two points in both ELA/literacy and mathematics. In 2017, the percent proficient increased only slightly in mathematics and decreased by 1.4 points in ELA/literacy. The 2017 change in percent proficient occurred near the 50th percentile—near the mean scale score—for ELA/literacy, and at the 40th percentile—below the mean—for mathematics.

Table 3. Corresponding Changes in Percent Proficient. Sample A Grades 3 to 8.

Year	Math Mean	Math Change	ELA/L Mean	ELA/L Change
2015	37.5		46.0	
2016	40.3	2.8	49.6	3.6
2017	41.2	0.9	49.4	-0.2
Net 3-year change:		3.7		3.4

Student-level data in Sample B were used to confirm the trends in Sample A. Sample B data included high school students, as well as grades 3 to 8. Trends were similar to those in Sample A. In 2017, there was a slight increase in the mean mathematics scale score (0.25) and a slight decrease in the mean ELA/literacy scale score (-0.88). However, the effect sizes (Glass’ Δ using the 2016 standard deviation) show that the changes are extremely small in proportion to the standard deviation of student achievement in 2016.

Table 4. Sample D Changes in Mean Scale Score in Grades 3 to 8 and 11

Year	Math N	Math Mean	SD	Change	Effect Size
2016	4,487,296	2504.72	112		
2017	4,528,851	2504.97	114	0.25	0.002

Year	ELA/Literacy N	ELA/Literacy Mean	SD	Change	Effect Size
2016	4,479,635	2513.23	113		
2017	4,517,885	2512.35	115	-0.88	-0.008

Trends in Mean Scale Score by Grade

Student-level data in Sample B were used to assess changes by grade in Smarter Balanced test scores. Results for change in scale score are shown in Table 5. Change was not uniform across grades. In mathematics, change was positive in grades 3 through 5 and increasingly negative within increasing grades, beginning in grade 6. In ELA/literacy, there were no clear change trends by grade. Change was

positive in high school, slightly positive in grade 7, and negative in other grades. Grade 5 showed the largest decline in ELA/literacy achievement. The effect sizes show that the changes are extremely small in proportion to the standard deviation of student achievement in 2016.

Table 5. Sample D Changes in Mean Scale Score by Subject and Grade.

Subject	Grade	2016 N	2016 Mean	2016 SD	2017 N	2017 Mean	2017 SD	Change	Effect Size
Math	3	671,517	2429.78	82	669,949	2431.49	84	1.71	0.021
	4	681,830	2467.25	84	674,566	2468.52	86	1.27	0.015
	5	668,997	2492.63	92	685,822	2492.74	95	0.11	0.001
	6	661,643	2514.26	107	668,760	2514.13	109	-0.13	-0.001
	7	656,201	2531.76	111	662,688	2531.56	114	-0.21	-0.002
	8	645,760	2547.58	120	656,093	2546.26	124	-1.32	-0.011
	11	501,348	2568.95	124	510,973	2566.33	128	-2.62	-0.021
	Overall:		4,487,296	2504.72	112	4,528,851	2504.97	114	0.25
ELA/Literacy	3	668,971	2420.00	90	666,881	2418.64	92	-1.37	-0.015
	4	679,535	2460.58	95	671,487	2460.03	96	-0.55	-0.006
	5	666,872	2500.66	96	683,226	2495.36	101	-5.30	-0.055
	6	659,976	2523.94	96	666,649	2522.47	97	-1.47	-0.015
	7	654,856	2546.90	99	660,789	2547.04	102	0.15	0.001
	8	645,134	2564.55	99	655,068	2562.29	102	-2.26	-0.023
	11	504,291	2601.08	111	513,785	2603.51	115	2.42	0.022
	Overall:		4,479,635	2513.23	113	4,517,885	2512.35	115	-0.88

Trends in Percent Proficient

Figure 2 plots the percent proficient in Sample A jurisdictions by subject and year. The mean scale score for each point on the plot is the weighted average over jurisdictions, with weights being the number of students per jurisdiction. For ELA/literacy, these were 46, 50, and 49 respectively for 2015 through 2017. For mathematics, the percentages were 38, 40, and 41 respectively. The change from 2016 to 2017 was -0.2 for ELA/literacy and 0.9 for mathematics.

Figure 2. Percent Proficient by Year in 14 Jurisdictions (Sample A).

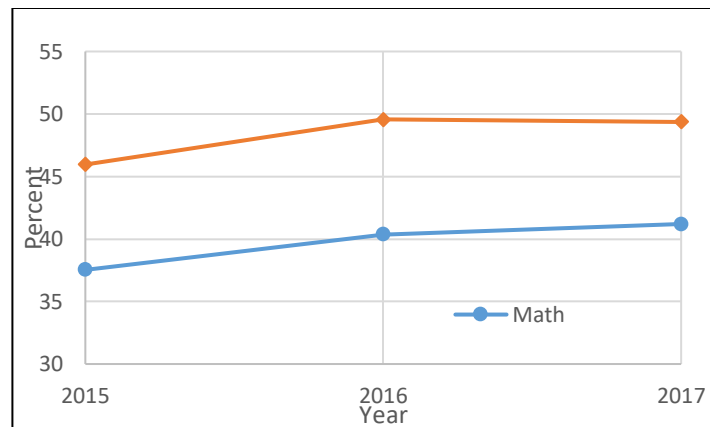


Table 6 shows that overall changes in percent proficient were similar in Sample D and that the pattern of change in percent proficient across grades generally conformed to the pattern of change in the mean scale score. Overall, in the 10 jurisdictions in sample D, there was almost no change in the percent of proficient students in either subject. The percent proficient decreased (-0.1) in ELA/literacy and increased (0.6) in mathematics. Grade 5 ELA/literacy showed the largest decrease in percent proficient. In mathematics, lower grades showed slight improvement while upper grades showed slight declines.

Table 6. Sample D Change in Percent Proficient by Grade and Overall.

Subject	Grade	Percent Proficient 2016	Percent Proficient 2017	Change
Math	3	48.3	48.8	0.5
	4	41.7	43.3	1.5
	5	36.3	36.9	0.6
	6	37.7	38.5	0.8
	7	39.0	39.4	0.4
	8	38.0	38.3	0.3
	11	32.6	32.5	-0.1
	Overall:		39.3	39.9
ELA	3	45.3	45.7	0.4
	4	46.7	47.0	0.3
	5	51.2	49.3	-1.9
	6	49.3	48.7	-0.6
	7	50.2	51.5	1.3
	8	51.2	50.4	-0.8
	11	59.6	60.4	0.8
	Overall:		50.1	50.1

Were there fewer test questions available?

This question is a simplified and highly focused version of the more general concern that the item pool used in 2017 may have differed from the 2016 item pool in ways that might have caused the test, and computer adaptive algorithm in particular, to yield lower estimates of student achievement compared to 2016. The effect, an ‘underestimation’ of student achievement, could conceivably occur more in some regions of the achievement scale than others, such as the region around the proficient cut score. It was noted above that the proficient cut score tends to fall near the 50th percentile ELA/literacy and near the 40th percentile for mathematics. Thus, care must be taken when investigating possible differences in item pools, to assess the effect of those differences over the range of student achievement.

The specific question of whether there were fewer test questions available can be answered simply by tallying the number of items in 2016 and 2017 and comparing the counts overall, and by grade within subject. This comparison is made later in this section. Broader questions about differences between item pools that could cause underestimation of achievement are also investigated and reported in this paper.

The following section attempts to set the context for assessing and understanding item pool changes.

Assessment Structure

Before considering changes in the item pool, it is important to understand the basic structure of the Smarter Balanced assessment. The Smarter Balanced assessments consist of a performance task (PT)

and a computer adaptive test (CAT). The performance task is non-adaptive. For each student, items are randomly selected from an available pool. Important differences between the PT and CAT sections of the test have to do with the role of hand scoring, the number of items and points representing each of these two segments in the blueprint, and how PT items are selected.

- PT items tend to be worth more than one point. This is especially true of the extended writing item (WER item type) on the ELA/literacy PT segment. The writing “item” is represented as two items in the Smarter Balanced, IRT-scoring technology, one worth 2-points and the other worth 4-points.
- The PT section accounts for approximately 11% of the items administered to a student but, in line with the previous bullet, accounts for approximately 20% of the points in the test, and therefore 20% of weight determining the estimate of a student’s achievement.
- The PT items are selected and delivered as a single ‘set’ of items having common stimuli, rather than item-by-item. In mathematics, a set consists of 3 to 5 (grades 5 and 11), or 4 to 6 items (all other grades).
- PT items tend to be hand-scored. In both subjects, a least one item may be machine-scored, but the rest may be hand-scored.
- The ratio of pool size to number of items in the blueprint is smaller for PT than for CAT. PT items comprise approximately 7% of the total item pool, but account for approximately 10% of the items delivered to students (and 20% of the test).
- Compared to the CAT segment, students spend more time per item on the PT.

The information in Table 7 may be important for understanding and suggesting possible follow-up analyses to item pool changes and other studies reported here. Smarter Balanced item types range from the traditional, such as multiple choice items, to the relatively new, technology enabled, such as Equation-Response (ER) and Grid Item Response (GI). In mathematics, short answer text (SA) items are found only in the PT segment. In ELA/literacy, they are found in both segments, but predominantly in the PT. As noted above, Writing Extended Response (WER) items are found only in the PT segment of the ELA/literacy test.

Table 7. Association of Item Types with Subject and Test Segment (PT or CAT).

Note: A '1' indicates that the item type can be found in the test segment

Item Type Abbreviation	Item Type Description	Math CAT	Math PT	ELA/Literacy CAT	ELA/Literacy PT
EBSR	Evidence-Based Selected Response			1	
EQ/ER	Equation Response	1	1		
GI	Grid Item Response	1	1		
HTQ	Hot Text			1	
MC	Multiple Choice	1	1	1	1
MI	Match Interaction	1	1	1	1
MS	Multiple Select	1	1	1	1
SA	Short Answer Text Response		1	1	1
TI	Table Interaction	1	1		
WER	Essay/Writing Extended Response				1

Item Pool Changes over Time

The 2015 and 2016 item pools for Smarter Balanced tests were largely the same. All items in both administrations were calibrated with data from the 2014 stand-alone field test. A small number of mathematics PT-item sets were based on a ‘classroom activity’ that was conducted with students prior to the PT itself. These few sets were used in 2015 but not in later administrations. Other than this, the items added to or dropped from the 2015 assessment in comparison to the 2016 assessment were very small in number and were not systematically associated with blueprint categories or item types.

In 2017, a relatively large number of new CAT items was added to the pool. This created three key groups of items for the analysis of item pool changes:

1. Old CAT items
2. New CAT items
3. PT (old) items.

The term ‘old’ is used here solely for convenience and brevity. It is not meant to imply that the items are outdated. “Old” CAT items are simply the items that were used in one or more previous assessments. The vast majority of old CAT items were used in both the 2015 and 2016 assessments. Likewise, all PT items were used in previous assessments. New CAT items were field tested in the 2015 assessment and used operationally for the first time in 2017. They had relatively limited exposure as embedded field test items in 2015.

Table 8 shows the differences between the 2016 and 2017 item pools with respect to these three groups of items. Other than the addition of new CAT items, the 2017 pool was virtually identical to the 2016 pool (and to the 2015 pool) in both subjects. Practically all of the CAT items from previous administrations were in the old CAT item group. Practically all of the PT items used in 2016 were also used in 2017. In effect, the 2017 item pool was larger than previous item pools by an amount equal to the number of new CAT items. In percentage terms, the ELA/literacy item pool was 50% larger and the mathematics item pool was 33% larger.

Table 8. Item Counts by Group and Year.

Item Group	Math 2016	Math 2017	ELA/L 2016	ELA/L 2017
Old CAT	6754	6588	4590	4231
New CAT	0	2569	0	2965
PT (old)	538	538	573	568
	7292	9695	5163	7764

Differences between Item Groups

The PT items differed from the CAT items in ways described above. But since the PT-component of the assessment was the same in 2017 as in previous administrations, these differences cannot play a direct role in explaining trends.

The new CAT items differed from the other two item groups in how and when they were calibrated. All of the old CAT and PT items were calibrated using data from the 2014 stand-alone field test. The new CAT

items were embedded field test items in an operational assessment—the 2015 administration. They were calibrated to the 2014 base scale by using the old CAT and PT items as ‘anchor’ items.

Group item statistics are generally expected to be comparable across groups unless item writing specifications change. The only known, large-scale change in item writing specifications were that the items field tested in 2015 were intentionally written to be easier. The new CAT items are therefore expected to be easier than the old CAT items.

Similarities between Item Groups

It is also important to note that the items field tested in 2014 were randomly administered, as opposed to adaptively administered, to students. Random administration is a defining characteristic of the base scale. It was essential that the 2015 embedded field test items were randomly administered. Because both old CAT and new CAT items were randomly administered, the item statistics, particularly IRT-item statistics, from these two groups are comparable. Other than the expectation that the new CAT items would be slightly easier, no other differences between these two groups of items is expected.

Did students receive more difficult test questions in 2017 compared to previous years?

Statistical Differences between Item Groups

Table 9 shows item statistics by group within grade and subject. As expected, in the 2017 item pool, new-CAT items were slightly easier than old CAT items at every grade within both subjects. This is seen by the fact that the average b-parameter (b-param) of 2017 new CAT items is lower (less positive or more negative) than the average b-parameter of 2017 old CAT items. Overall, in ELA/literacy, the new CAT items had a mean b-parameter of 0.53 compared to 0.68 for the old CAT items.

Table 9. Item Statistics by Item-Group within Grade within Subject.

Subject	Grade	2016 Old CAT			2016 PT			2017 Old CAT			2017 New CAT			2017 PT		
		N	a-parm	b-parm	N	a-parm	b-parm	N	a-parm	b-parm	N	a-parm	b-parm	N	a-parm	b-parm
Math	3	826	0.85	-0.83	80	0.89	-0.52	808	0.84	-0.83	384	0.83	-0.98	80	0.89	-0.52
	4	818	0.82	-0.07	95	0.85	-0.03	807	0.82	-0.08	472	0.83	-0.31	95	0.85	-0.03
	5	809	0.78	0.67	85	0.76	1.01	776	0.78	0.69	434	0.76	0.20	85	0.76	1.01
	6	739	0.70	1.06	72	0.73	0.80	711	0.70	1.08	357	0.69	0.76	72	0.73	0.80
	7	669	0.71	1.79	87	0.89	1.58	651	0.71	1.79	292	0.70	1.46	87	0.89	1.58
	8	608	0.60	2.31	58	0.88	1.81	584	0.60	2.35	259	0.55	1.34	58	0.88	1.81
	11	2285	0.55	2.53	61	0.66	2.67	2251	0.55	2.54	371	0.65	2.01	61	0.66	2.67
	Overall	6754	0.68	1.33	538	0.81	0.94	6588	0.68	1.34	2569	0.73	0.53	538	0.81	0.94
ELA	3	579	0.67	-0.42	62	0.71	0.21	522	0.66	-0.45	357	0.70	-0.63	62	0.71	0.21
	4	555	0.59	0.13	82	0.64	0.45	493	0.59	0.12	342	0.62	-0.05	81	0.64	0.45
	5	534	0.60	0.51	95	0.71	0.74	473	0.61	0.52	355	0.62	0.27	93	0.71	0.74
	6	536	0.54	1.00	61	0.87	0.92	477	0.54	1.03	292	0.60	0.64	61	0.87	0.92
	7	493	0.53	1.12	79	0.81	1.15	444	0.53	1.14	255	0.57	1.07	79	0.81	1.15
	8	483	0.53	1.30	89	0.72	1.26	448	0.53	1.27	303	0.56	1.08	88	0.72	1.26
	11	1410	0.50	1.70	105	0.59	1.84	1374	0.50	1.73	1061	0.51	1.56	104	0.59	1.84
	Overall	4590	0.55	0.92	573	0.71	1.00	4231	0.55	0.95	2965	0.58	0.78	568	0.71	1.00

As expected, there were no substantial differences between the 2017 and 2016 item pools solely with regard to the old CAT and PT item groups. Both pools contained exactly 538 mathematics PT items. These were likely (but not necessarily) the same 538 items, and they had identical average difficulty (mean b-parm = 0.94) and discrimination (mean a-parm = 0.81). There were slight differences in count and average item statistics across years for other groups of items, but these differences were insubstantial.

Differences between Old and New CAT Items by Decile of Student Achievement

Differences between item pools having greater impact on some students more than others, depending on the students' achievement, might not be revealed by differences in the overall number of items and averages of item parameters. To investigate this, counts and statistics of the old and new CAT item groups were studied by decile of student achievement to see whether the addition of the new items in 2017 could have affected students at some levels of achievement more than others.

Deciles were defined by ranking students by achievement scores and dividing them into ten, approximately equal-sized groups from lowest to highest achievement. Decile 1 contains students in the lowest ten-percent of the population (below the 10th percentile). Decile 10 contains students in the highest ten-percent of the population (at or above the 90th percentile).

Items were classified into these deciles by their difficulty parameter (b-parameter) in the item-response-theory (IRT) models used to estimate student achievement. In IRT models, estimates of item difficulty (b-parameter) and estimates of student achievement are on the same scale. In computer adaptive testing, the items delivered to a student will tend to come from deciles that contain or are near, the student's achievement score.

Two sets of deciles were defined: 1) one for each subject area by combining students across grades, and 2) one for each grade within subject area. Analyses and trends based on deciles for each grade and

subject are expected to fit the overall pattern shown by analyses and trends for the subject area as a whole. This was generally the case, so the results presented in the body of this report are based mostly on subject area deciles.

Figures 3 and 4 show that the new-CAT items fell into every decile of student achievement, both subjects. Relative to the old CAT items, the new CAT items tended to fall more heavily into the middle-to-lower deciles. ELA/literacy cut scores for Level 3 (proficient) tend to lie in the fifth and sixth deciles. Mathematics cut scores tend to lie in the sixth and seventh deciles. Deciles for these figures were based on the Sample B, 2017 distribution of student achievement combined over all grades. Item counts were also combined over all grades.

Figure 3. CAT Item Counts by Old/New and Decile. ELA/Literacy

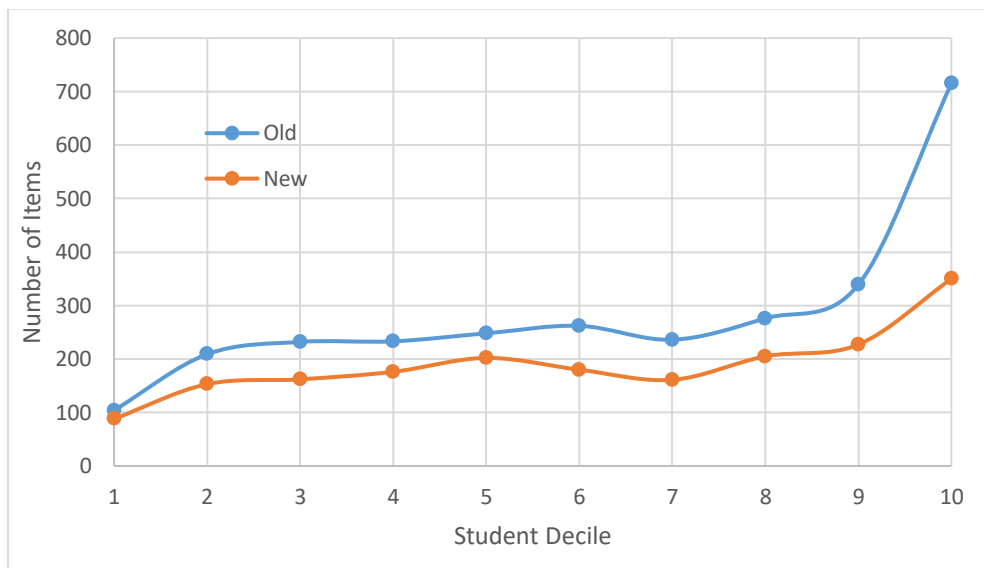
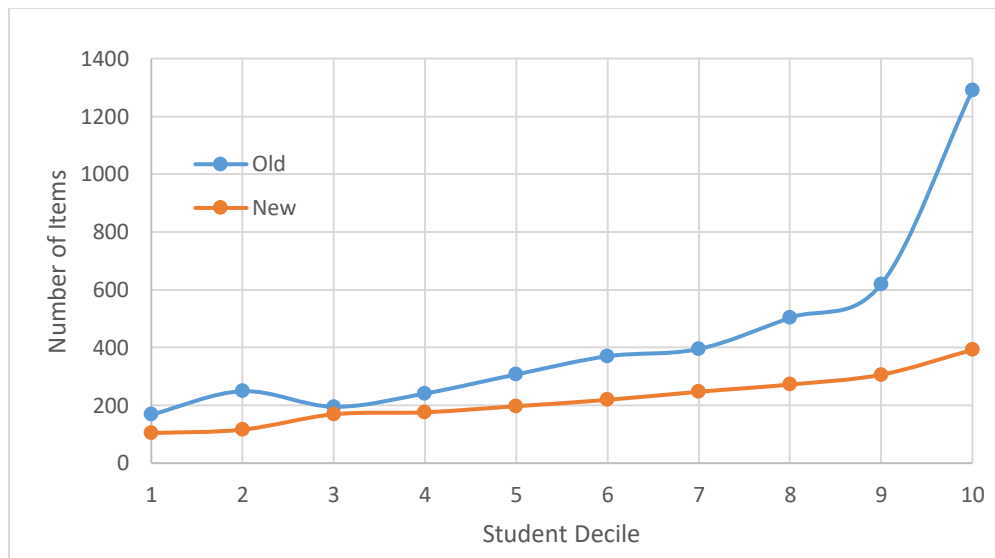


Figure 4. CAT Item Counts by Old/New and Decile. Mathematics.



Figures 5 and 6 show that there was no substantial difference between old and new CAT item groups in average item discrimination by decile. In ELA/literacy, both old and new groups of CAT items show a trend of decreasing discrimination with increasing student achievement. There is no such clear trend in mathematics.

Figure 5. CAT Item Discrimination by Old/New and Decile. ELA/literacy.

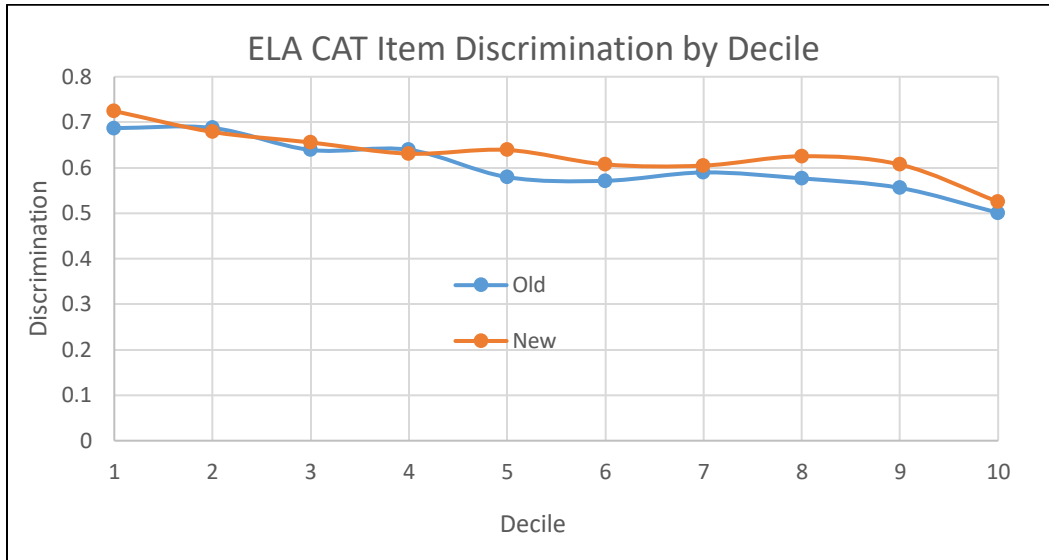
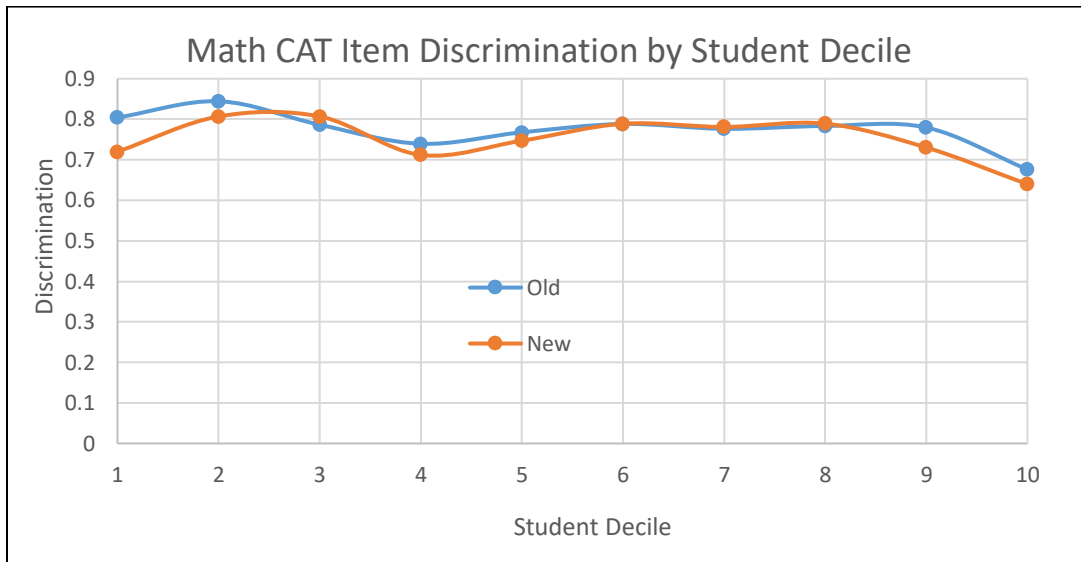


Figure 6. CAT Item Discrimination by Old/New and Decile. Mathematics.



Figures 7 and 8, for Grade 5 ELA/literacy, are similar to Figures 3 and 5, for ELA/literacy overall. New items, being easier overall, tend to fall relatively more often into the lower deciles of student achievement. There are no substantial differences in discrimination between old and new CAT items. In ELA/literacy, item discrimination tends to decrease with increasing student achievement. Plots of item counts and item discrimination by decile for other grades in both ELA/literacy and mathematics showed similar patterns.

Figure 7. Grade 5 ELA/Literacy 2017 CAT Item Counts by Decile.

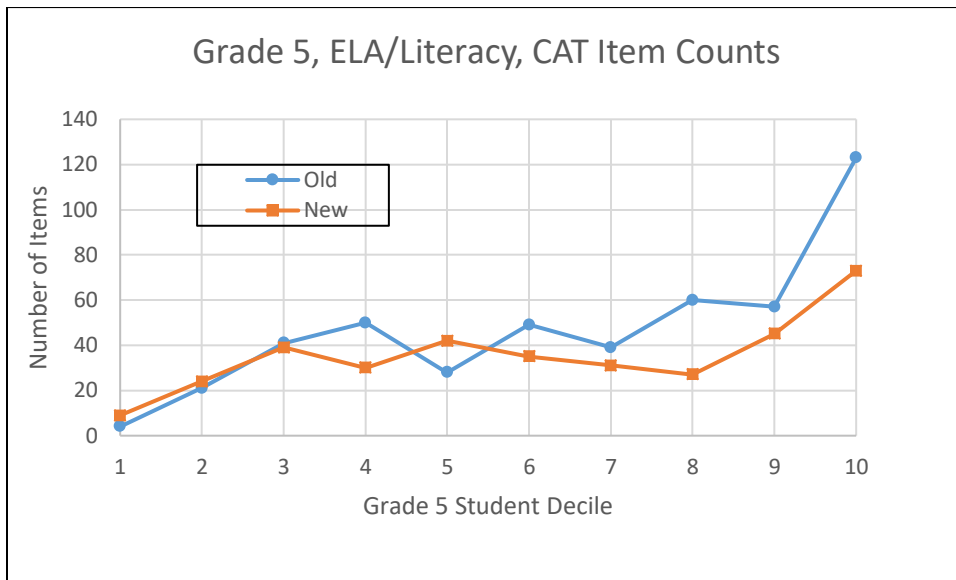
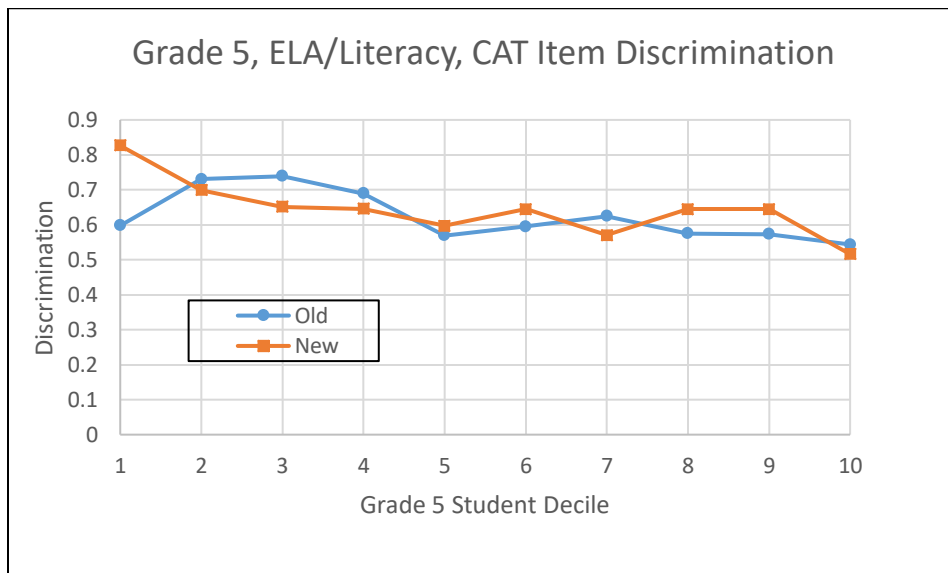


Figure 8. Grade 5 ELA/Literacy 2017 CAT Item Discrimination by Decile.



Measurement Precision

The measurement precision of the test in 2017 was essentially the same as in 2016. Figure 9 shows the average standard error of measurement (SEM) by decile for each subject. Decile lower boundaries were defined using the 2016 student distribution of achievement in each subject. SEMs are based on the entire test (CAT plus PT). On one hand, the inclusion of PT items in these plots somewhat diminishes differences by year stemming strictly from the contrast of old versus new CAT items. On the other hand, the standard errors of student scores in achievement trends are based on the entire blueprint, not just CAT items.

Standard Error of Measurement (SEM)

The average SEMs by year are nearly identical at each decile and show similar trends over deciles. In both years, the SEM is larger in lower deciles, reflecting the fact that the item pools for both subjects contain proportionally fewer items at lower levels of achievement. As the item pool becomes thinner, targeting is less precise for very low achieving students because meeting the test blueprint is a requirement governing state assessments.

Figure 10 shows that results were similar for grades within subject. The deciles used for creating the plot for a given grade and year were defined by the student achievement distribution for that grade and year. Grades 5 and 7 in ELA/literacy and grade 4 in mathematics were selected for presentation in figure 10 because they met both of the following conditions as can be verified with reference to table 6: 1) they exhibited an absolute 2016-to-2017 change of 1 or more points in the percent proficient and 2) they were the most extreme cases within their subject area of positive or negative change in percent proficient. Plots for all grades within subject were created and inspected. The patterns seen in figure 10 were observed in all plots.

Figure 9. Standard Error by Decile for Each Subject

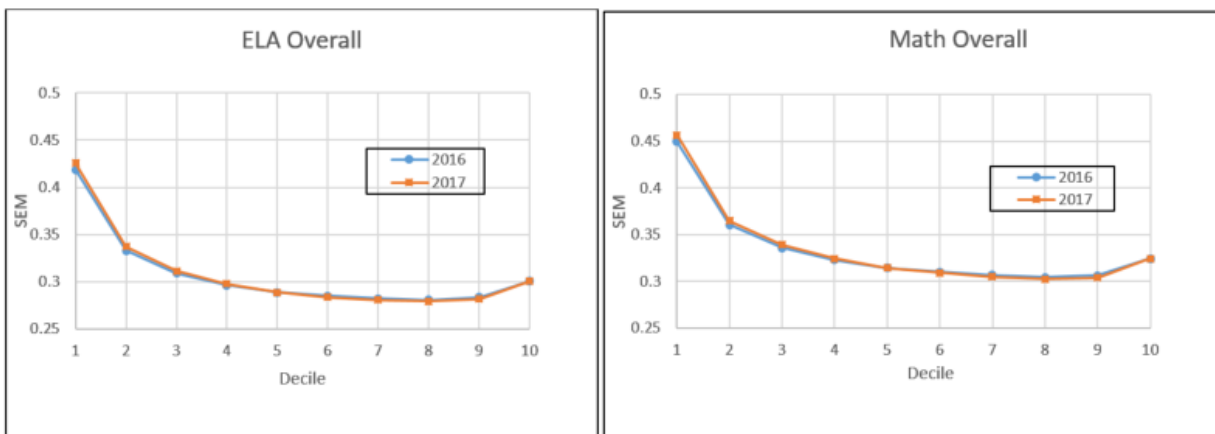
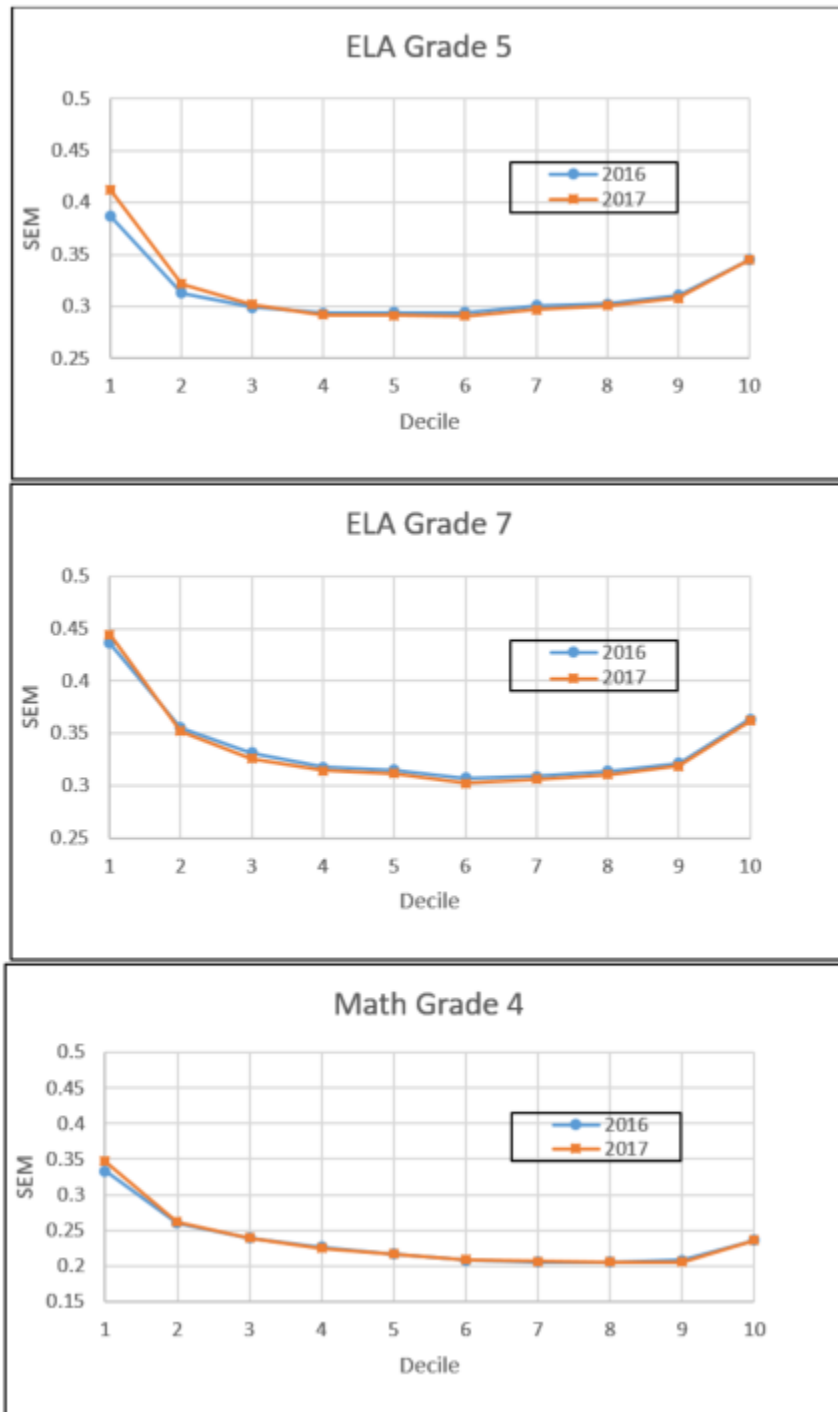


Figure 10. Standard Error of Measurement by Decile within Grade within Subject

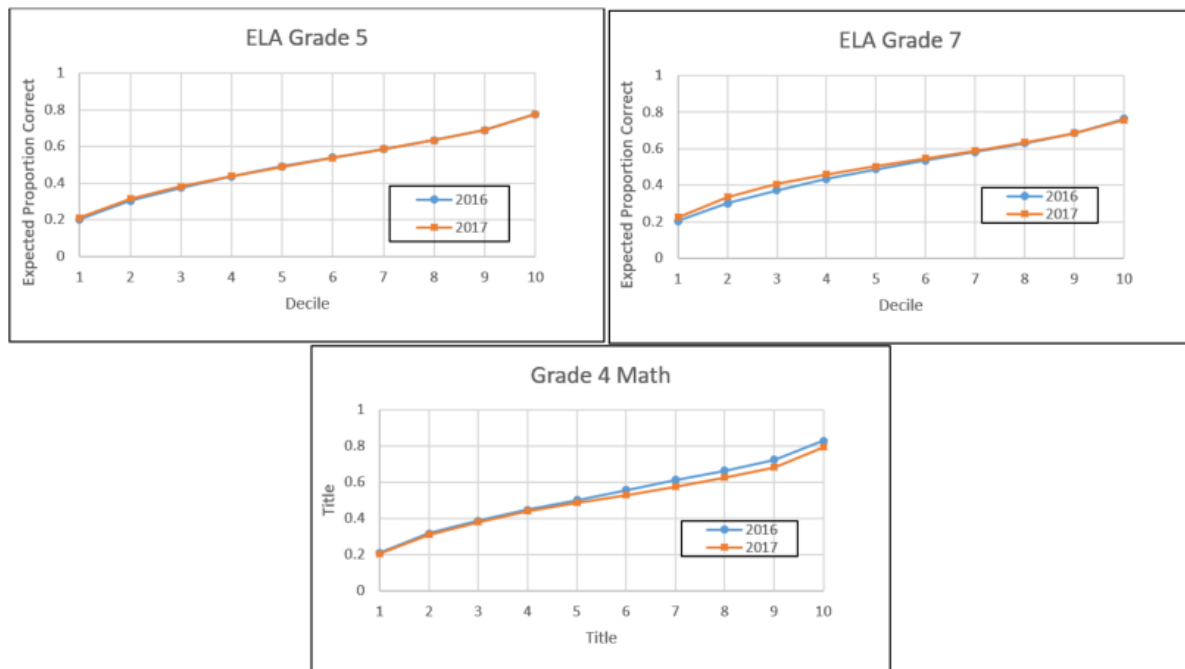


Expected Scores

Figure 11 presents another perspective on measurement precision—how well the CAT test was targeted on student achievement. This comparison used all operational items (CAT plus PT). Results are shown only for the grade/subject combinations selected previously (ELA/literacy grades 5 and 7; mathematics grade 4). For all three grade/subject combinations shown in Figure 11, there were no substantial differences between years (2017 vs. 2016) in the expected score by decile. This result is consistent with the findings in Figure 4, showing no substantial differences between years in measurement precision at any decile of student achievement. Plots for all grades and both subjects showed patterns that were similar to those in Figure 11.

Expected scores in these figures are not uniformly near 0.5, as one would expect in a computer adaptive test that is delivering items solely on the basis of matching item difficulty to student performance. Three reasons for this are as follows: First, the expected scores include scores on both the CAT and the PT (performance test) segments of the test. The PT is not adaptive. Second, the computer adaptive algorithm has to satisfy test blueprint constraints. In the first and second deciles, there are relatively few items, so the adaptive algorithm may have to select more difficult items for students in these deciles in order to meet blueprint constraints. Third, a significant portion of the CAT segment consists of sets of three or four items associated with a common passage or stimulus. The adaptive algorithm cannot target student performance as effectively when delivering items in sets.

Figure 11. Expected Scores by Decile



Did the newly-added test questions impact test results?

In response to the flat trends observed in 2017, it has been suggested that the new items added to the item pool in 2017 were “harder.” An objective way to investigate this claim is to allow the possibility that the new CAT items were harder than *expected* in a way that can be revealed through a residual analysis and that this unexpected difficulty is due to design, procedural, or technical flaws in the field testing and calibration of the new CAT items. Information presented in other sections of this report shows that 1) the new items were not harder, and 2) the difficulty and precision of the tests students received were not affected by the addition of new items to the pool. Even if the new CAT items had been harder, the nature of computer adaptive testing would have assured the second finding.

A residual is the difference between the student’s score on an item and the predicted score. The score is predicted from the item’s statistics in the item-response-theory (IRT) model and an estimate of the student’s achievement. The residuals are expected to deviate from 0 because models are fitted to empirical data with error, but here we focus especially on the sign of the model misfit. A residual is positive if the student’s score was higher than the predicted score or negative if the student’s score was lower than the predicted score. When averaged over all students who saw the item, a positive mean residual indicates that the item was easier than expected and a negative mean residual indicates that the item was harder than expected.

It is only in comparing the mean residual of one item to another, or of one group of items to another, and seeing a difference, that we can say that one group of items is more difficult than expected. This is because 1) the estimate of student achievement used to compute the ‘predicted’ score on an item is based on all of the items taken by the student and 2) the algorithm used to estimate the student’s achievement generally arrives at an estimate where the sum of residuals over the items taken by the student is zero or close to zero. This means that residuals are a zero-sum variable. If the mean residual is positive for one item, it must be negative for another.

Data and Method

The residual analysis focused on the 2017 data from Sample B. The average residual was computed for three groups of items: old CAT, new CAT, and PT (old). Under the hypothesis that new CAT items would yield the same measures of student achievement as old CAT items, the average residual for these two groups of items should be the same. The PT items would not be expected to have the same average mean residual as CAT items, either old or new, due to the many ways these items differ from CAT items as described previously.

As a measure of quality assurance concerning the statistical procedures for computing item residuals and mean residuals for groups of items, Smarter Balanced replicated an analysis of residuals that the American Institutes for Research had performed with 2017 data from one state. Other than differing from AIR in how a few items were classified (old vs. new CAT), Smarter Balanced’s results were exactly equal to AIR’s results in terms of average residuals for each group of items (new CAT, old CAT, and PT items) at every grade within both subjects.

Item Counts, Exposure, and Mean Residuals by Item Group and Grade

Table 10 shows the numbers of items in each group and the percent of times items in each group were administered to students. (The percent of residuals is the same thing as the percent of times items were

administered to students.) The last row for each subject allows one to compare the percent of items in a group to the percent of times items in a group were administered to students. This information is presented to assure the reader that the analyses performed in this study, including computation of averages for groups of items, was correctly based on counts of residuals and not items.

Table 10. Item Counts and Residual Counts by Item Group within Grade within Subject (Sample B).

Subject	Grade	Item Count: Old CAT	Item Count: New CAT	Item Count: PT (Old)	Percent of Residuals: Old CAT	Percent of Residuals: New CAT	Percent of Residuals: PT (Old)
Math	3	809	384	80	54%	32%	14%
	4	819	472	94	57%	30%	13%
	5	820	434	85	53%	33%	14%
	6	742	357	72	55%	30%	15%
	7	668	292	87	56%	30%	14%
	8	599	259	58	53%	34%	12%
	11	2,258	371	61	71%	17%	12%
			68%	26%	5%	57%	30%
ELA/Literacy	3	515	356	62	47%	43%	10%
	4	507	349	81	44%	45%	10%
	5	494	348	93	44%	46%	11%
	6	494	284	61	50%	40%	10%
	7	459	245	79	48%	43%	10%
	8	486	300	88	52%	39%	10%
	11	1342	1051	104	47%	43%	10%
			55%	38%	7%	47%	43%

Table 11 shows the mean residual by grade and subject for each group of items. For each subject, there are two distinct and consistent-across-grade patterns of difference among the three groups of items. These patterns are shown by the values in the last two columns of the table.

First, within a subject, the overall mean residual of the CAT items (old and new combined) has the same sign for every grade. The overall mean residual of CAT items is shown in the last column of Table 11 under the heading “Weighted CAT Mean (WCM)”. The weighted CAT mean is the weighted average of the means in the old-CAT and new-CAT columns. The means are weighted by the percent of residuals that they represent, shown in Table 10. In ELA/literacy, the CAT mean residual is negative for every grade. In mathematics, the CAT mean residual is positive for every grade. A difference in either direction means that student cohorts became more ‘proficient’ in one of these two item groups over time than in the other group. In ELA/literacy, the greater proficiency was developed for PT items. In mathematics the greater proficiency was developed for CAT items.

There are any number of reasons why student cohorts might become more proficient in one group of items than in another over time. Items in one group or the other may be more exposed due to the smaller number of items in the group relative to their weight in the blueprint, or they may be easier to remember. These possibilities seem likely to play a role in students becoming more proficient in PT items than in CAT items over time. In ELA/literacy, the PT section of the test consists of an extended writing item worth

6 points, plus just two or three other items. Students may have learned how to respond to these items—to master the task model for these items—relatively more quickly, especially the 6-point writing item.

Why the opposite would occur in mathematics—students becoming relatively more proficient on CAT items than on PT items over time—is harder to understand. There are no 6-point items in the mathematics PT. Differences in task models and content representation may also play a role. Whatever the reason, it is important to keep in mind that the magnitude of the weighted CAT mean residual is considerably smaller in magnitude in mathematics than in ELA/literacy. That is to say, it might be more appropriate to say that the PT vs. CAT residual difference in mathematics, but not ELA/literacy, is trivial, or practically zero.

Table 11. 2017 Mean Residuals (Sample B).

Subject	Grade	Mean Residuals: Old CAT	Mean Residuals: New CAT	Mean Residuals: PT (Old)	Old CAT minus New CAT	Weighted CAT Mean (WCM)
Math	3	0.012	-0.011	-0.012	0.023	0.003
	4	0.017	-0.013	-0.005	0.030	0.007
	5	0.015	-0.007	-0.012	0.022	0.007
	6	0.003	0.002	-0.012	0.001	0.003
	7	0.010	-0.005	-0.009	0.015	0.005
	8	0.010	-0.004	-0.011	0.014	0.005
	11	0.007	-0.005	-0.009	0.012	0.005
	Overall		0.011	-0.006	-0.010	0.017
ELA/Literacy	3	0.000	-0.020	0.033	0.020	-0.010
	4	-0.005	-0.021	0.039	0.016	-0.013
	5	-0.013	-0.021	0.057	0.008	-0.017
	6	-0.010	-0.021	0.046	0.011	-0.015
	7	-0.008	-0.033	0.059	0.025	-0.020
	8	-0.003	-0.035	0.055	0.032	-0.017
	11	-0.018	-0.029	0.067	0.011	-0.023
	Overall:		-0.008	-0.026	0.051	0.018

The second consistent pattern in Table 11 is that the new CAT items are more difficult than expected compared to the old CAT items. This is shown by positive values in the second-to-last column of Table 11, labeled “Old CAT minus New CAT” at every grade, in both subjects. In other words, the mean residual of new CAT items is less positive, or more negative, than the mean residual of old CAT items. On average, across grades, the residual for an old CAT item is 0.017 more positive in mathematics and 0.018 more positive in ELA/literacy. The direction of this difference is the same in all grades and both subjects.

Although this difference indicates that the new CAT items would yield lower measures of student achievement than the old CAT items, one cannot interpret this finding out of context and without considering the magnitude and practical effect of the difference. With regard to context, the old CAT items were used in two previous operational administrations, plus the stand-alone 2014 field test. The new CAT items had no previous operational use and only very limited exposure as embedded field test items two years before the 2017 administration. Items generally become slightly easier over time through exposure—

a phenomenon that is controlled by adding new items to the assessment. The effects of exposure on item residuals and measures of student achievement, however, are generally quite small.

Before addressing the question of practical effects in more detail, however, the next section addresses the possibility that means, and differences between means, may not represent, or may be suppressed by, more complicated effects possibly interacting with levels of student achievement.

Item Counts, Exposure, and Mean Residuals by Decile within Grade

It is common wisdom in applied statistics that the mean of a distribution can often fail to represent important differences or effects taking place throughout the distribution, especially in regions far from the average. It is of particular interest in this study to know what effects and differences are taking place in the part of the student distribution near the proficient cut score. The null hypothesis in this case is that the differences between the three item groups—old CAT, new CAT, and PT—as shown by the overall mean residuals in Table 11, apply to students at every level of achievement, including those near the proficient cut score.

To evaluate this hypothesis, students were classified into deciles by their estimated achievement and the mean residuals for the three groups of items were computed and tabulated by decile. This was done for all grades within subject. The difference between old and new CAT item residuals varied over deciles. With few exceptions, however, the difference in the decile containing the proficient cut score is close in value to the overall mean difference.

Table 12 shows results for each decile that contains the proficient cut score by grade and subject. With few exceptions, the Old CAT minus New CAT difference within the “proficient decile” is reasonably close to the overall difference in Table 11. Exceptions are Grades 6, 8, and 11 in mathematics. The old CAT minus new CAT difference for these grades is zero or negative at the decile containing the proficient cut score. The overall mean differences across grades in Table 12 (proficient deciles): .021 for ELA/L and .015 for mathematics, are reasonably close to the mean differences across grades in Table 11: 0.017 for ELA/L and 0.018 for mathematics.

The results in this section show that with few exceptions (grades 6, 8, and 11 in mathematics), means and differences between means are fairly consistent with what is taking place at the deciles containing the proficient cut score (Table 10). At grades 6, 8, and 11 in mathematics (the exceptions), there is virtually no difference between new and old CAT item residuals in the decile containing the proficient cut score.

Table 12. Residual Means at Deciles Containing Proficient Cut Score by Grade and Subject (Sample B).

Subject	Grade	Theta Cut	Decile	Lower Bound	Mean Residuals: Old CAT	Mean Residuals: New CAT	Mean Residuals: PT (Old)	Old CAT minus New CAT
Math	3	-0.99	5	-1.24	0.005	-0.018	0.006	0.023
	4	-0.38	6	-0.50	0.011	-0.016	0.008	0.027
	5	0.17	7	0.09	0.012	-0.016	-0.016	0.028
	6	0.47	7	0.44	-0.006	-0.004	-0.007	-0.002
	7	0.66	7	0.64	0.010	-0.017	-0.019	0.027
	8	0.90	7	0.82	0.003	0.003	-0.019	0.000
	11	1.43	7	1.06	-0.003	-0.002	-0.015	-0.001
								0.015
ELA/Literacy	3	-0.89	6	-1.11	-0.009	-0.021	0.042	0.012
	4	-0.41	6	-0.57	-0.004	-0.029	0.048	0.025
	5	-0.07	6	-0.14	-0.019	-0.029	0.077	0.010
	6	0.27	6	0.19	-0.011	-0.031	0.057	0.021
	7	0.51	5	0.20	-0.016	-0.037	0.069	0.021
	8	0.69	6	0.68	-0.001	-0.044	0.070	0.043
	11	0.87	5	0.84	-0.024	-0.042	0.095	0.018
								0.021

Practical Impact

In assessing the practical impact of differences between old and new CAT item residuals, it may be informative to consider the magnitude of differences between means for different groups of items in the last row for each subject in Table 11. The difference between old and new CAT item residuals is .017 for mathematics and .018 for ELA/literacy. These differences are less than a third the size of the difference between CAT and PT item residuals in ELA/literacy (.067) and are similar in magnitude to the CAT vs. PT difference in mathematics (.015). The CAT and PT item groups differ in exposure, differential learning, and other phenomena that generally occur in assessment programs. It therefore seems reasonable to conclude that differences between old and new CAT item residuals are well within the range of residual differences one might expect to see in a typical assessment program. The effects of such differences on student measures are generally thought to be quite small from year to year.

Given differences among grades in how much student achievement changed from 2016 to 2017 (see Tables 5 and 6), another reasonable question to ask is, “did the change in test scores from 2016 to 2017 at a given grade correspond to how differently the newly-added items in 2017 performed compared to old CAT items?” Table 13 shows for each grade and subject, two measures of change from 2016 to 2017 and two measures of difference between old and new CAT items. Each measure of change is logically paired with a measure of difference between residuals. Measures of change are based on Sample D.

Differences between residuals are based on Sample B. These samples are highly similar in the patterns of gains across grades for each subject.

- Change in the mean scale score (Mean SS) is expected to be negatively associated with the old CAT minus new CAT difference in mean residuals computed over the entire distribution (overall)

of student achievement. Both the mean scale score and the *overall* difference in mean residuals are based on the entire student distribution.

- Change in the percent proficient (% Prof) is expected to be negatively associated with the old CAT minus new CAT difference at the decile containing the proficient cut score (Proficient Decile). Change in the percent proficient is most likely to be affected by differences between old CAT and new CAT mean residuals within the decile containing the proficient cut score.

A negative relationship between members of these pairs is based on the assumption that a positive, old CAT minus new CAT mean residual difference means that students are not performing as well as expected on the new CAT items as they are on the old CAT items.

The correlations at the bottom of Table 13 do not confirm these expectations. They are close to zero at best and positive at worst. Given the small number of observations for each coefficient (seven) none of the correlations may be statistically significant, meaning one could not reject the hypothesis that there is no relationship association between change from 2016-to-2017 how much more difficult than expected the new CAT items were compared to the old CAT items.

This is not to say that the new CAT items were not more difficult than expected, but rather, that magnitude of this differences has no clear relationship to how much growth students at a given grade showed, compared to students at other grades, from 2016 to 2017.

Table 13. Old-Minus-New CAT Differences and Change in Student Achievement

Subject	Grade	2017 Change: Mean SS (a)	2017 Change: % Prof. (b)	Old minus New Difference: Overall (c)	Old minus New Difference: Proficient Decile (d)
Math	3	1.7	0.5	0.014	0.014
	4	1.3	1.5	0.020	0.018
	5	0.1	0.6	0.014	0.018
	6	-0.1	0.8	0.001	-0.001
	7	-0.2	0.4	0.010	0.017
	8	-1.3	0.3	0.009	0.000
	11	-2.6	-0.1	0.010	-0.001
		-0.2	0.6	0.011	0.009

Correlation (a,c) = 0.48

Correlation (b,d) = 0.48

Subject	Grade	2017 Change: Mean SS (a)	2017 Change: % Prof. (b)	Old minus New Difference: Overall (c)	Old minus New Difference: Proficient Decile (d)
ELA/L	3	-1.4	0.4	0.020	0.006
	4	-0.5	0.3	0.016	0.012
	5	-5.3	-1.9	0.008	0.005
	6	-1.5	-0.6	0.011	0.011
	7	0.1	1.3	0.025	0.011
	8	-2.3	-0.8	0.032	0.022
	11	2.4	0.8	0.011	0.009
		-1.2	-0.1	0.018	0.011

Correlation (a,c) = 0.08

Correlation (b,d) = 0.00

Another perspective on whether the observed differences between old and new CAT mean residuals could account for lack of gains in mean achievement scores from 2016 to 2017 can be gained by considering the standard deviation of change in mean scale score across grades. The standard deviation of the values in the (c) columns of Table 13 is approximately 2 points (2.4 for ELA/literacy and 1.5 for mathematics). Now suppose that a difference of .005 between old and new mean residuals (column (a)) suppressed change in mean scale scores by 1 scale score point, and a difference of .01 suppressed change in mean scale scores by 2 scale score points, and so on in proportion. If this were the case, the variation of difference in ELA/literacy mean residuals (column (c) of the ELA/literacy section of Table 13) would add 1.7 points to the standard deviation of change in column (a). But the standard deviation of change in column (a) is only 2.4 to begin with. If 1.7 of this 2.4 were due to variation among the values of column (c), the correlation between the values in columns a and c (Correl(a,c)) would be negative as expected, in fact, close to -1.

Still another perspective on the observed differences between old and new CAT mean residuals is gained by considering the difference between the overall new CAT mean residual and the overall CAT mean residual at the bottom of each subject section of Table 11. This difference is .011 for mathematics (.005 minus -.006) and .010 for ELA/literacy (-.016 minus -.026). These differences show the direction and magnitude that each new CAT item residual would have to change in order for the new CAT and old CAT mean residuals to show no difference and to be equal to the overall CAT mean residual. The effect of

these changes on a student's total score is proportional to the number of points represented by the new CAT items, which is approximately proportional to the percentage of residuals represented by the new CAT items. Table 12 shows that this percentage over all grades is 30% for mathematics and 43% for ELA/literacy. Given a test length of approximately 40 items, these percentages translate to approximately 12 items or points, for mathematics, and 16 items, or points, for ELA/literacy due to new CAT items. So adjusting this number of items by the magnitude and direction of the differences computed above (0.011 for mathematics and .01 for ELA/literacy), would amount to adding 0.132 raw score points to students' mathematics total score and .16 raw score points to students' ELA/literacy score. These differences do not translate to an appreciable impact on 2017 test results and change from 2016 to 2017.

Based on these analyses that addressed a variety of perspectives, it seems doubtful that the new CAT items actually had a substantial effect in suppressing achievement scores. It therefore seems doubtful that the flat or negative achievement gains in 2017 can be attributed to effects involving the new CAT items.

Did students spend less time taking the test?

Analyses of the amount of time students spent on the test in 2017 versus 2016 are still being refined. The time students spend viewing and answering an item is not recorded by the test delivery system individually for each item, but rather, is recorded by 'page'. A page may contain more than one item, including embedded field test items. It is important to avoid attributing the time students spend viewing embedded field test items to the time they spend taking operational items because the 2017 administration contained more short-answer embedded field test items than the 2016 administration and short-answer items take students longer to answer than other item types.

Table 14 shows the current set of results for the amount of time, in minutes, that students spent taking the test in 2016 versus 2017. The last three columns under the general heading of "change" show that students spent more time on both the CAT and PT sections of the test in 2017 compared to 2016. There is no evidence in these results that students spent less time on the test in 2017 compared to 2016. On average, across grades, students spent about 10 minutes longer taking the test in 2017 compared to 2016 (9.6 minutes for mathematics and 11.2 minutes for ELA/literacy). On the PT section, students spent 2 minutes longer in mathematics and 7 minutes longer in ELA/literacy. The PT section is not affected by the time students spend on embedded field test items because there are no embedded field test items in this segment.

Table 14. Test Durations (Minutes) by Test Segment (CAT and PT) and School Year (Sample D).

Subject	Grade	2015–16 CAT	2015–16 PT	2015–16 Total	2016–17 CAT	2016–17 PT	2016–17 Total	Change CAT	Change PT	Change Total
Math	3	72.3	42.7	115.0	77.6	45.1	122.7	5.3	2.4	7.7
	4	74.1	40.4	114.5	82.2	42.0	124.2	8.1	1.6	9.7
	5	78.7	62.2	140.9	85.7	64.5	150.2	7.0	2.3	9.3
	6	83.0	53.0	136.0	97.0	54.9	151.9	14.0	1.9	15.9
	7	82.3	31.0	113.3	87.2	32.5	119.7	4.9	1.5	6.4
	8	86.7	37.8	124.5	95.2	39.7	134.9	8.5	1.9	10.4
	11	67.9	31.2	99.1	74.5	32.1	106.6	6.6	0.9	7.5
							7.8	1.8	9.6	
ELA/L	3	90.3	109.5	199.8	96.4	118.5	214.9	6.1	9.0	15.1
	4	97.5	119.4	216.9	99.5	126.8	226.3	2.0	7.4	9.4
	5	95.6	119.2	214.8	98.1	124.9	223.0	2.5	5.7	8.2
	6	101.7	114.9	216.6	102.7	123.6	226.3	1.0	8.7	9.7
	7	86.9	101.4	188.3	93.7	108.2	201.9	6.8	6.8	13.6
	8	89.5	101.5	191.0	91.0	107.1	198.1	1.5	5.6	7.1
	11	78.0	88.5	166.5	88.1	93.4	181.5	10.1	4.9	15.0
							4.3	6.9	11.2	

Did students take the test earlier in the school year?

Table 15 summarizes results of analyses of test starting times. At most grades, the 2016 ELA/literacy test was started an average of 111 days into the year. In 2017, the start date was very much the same. Likewise, the start date for the mathematics test did not change appreciably from 2016 to 2017.

Table 15. Test Start Dates (Days past December 31st) (Sample D).

Subject	Grade	2016 Min	2016 Max	2016 Mean	2017 Min	2017 Max	2017 Mean	Difference Between Means
Math	3	52	158	121	65	158	121	0.1
	4	66	158	121	53	158	121	0.8
	5	69	159	121	66	158	122	0.5
	6	68	159	119	52	158	121	1.7
	7	68	159	118	64	157	118	0.8
	8	65	160	118	63	157	119	0.8
	11	58	152	114	50	154	111	-3.5
ELA/L	3	63	154	111	63	156	112	0.5
	4	68	158	111	51	155	113	1.1
	5	68	155	113	64	154	113	0.2
	6	68	156	111	51	156	112	0.6
	7	68	159	111	63	155	111	0.2
	8	64	160	111	63	157	111	0.2
	11	36	153	110	49	154	106	-4.1

Did the student demographics change?

The demographic analysis used data from states that had delivered data files for the 2017 assessment, containing all necessary data elements for the analysis and for whom matching data elements were found in the 2016 assessment files. Sample B in Table 1 shows the jurisdictions used for this analysis. Standard selection criteria for inclusion in accountability reports were applied to student records.

Descriptive statistics were generated for the overall population of students with eligible data and by examinee subgroup (gender, ethnicity, LEP, IDEA, Section 504, economic disadvantage). The descriptive statistics include the proportion of subgroup counts, the mean scale scores and proportion of proficient students (achievement levels 3 & 4). These statistics were computed for each year (spring, 2016 and spring, 2017). Changes from 2016 to 2017 were computed as the spring 2017 value minus the spring, 2016 value.

The number of students in each of the six race/ethnicity subgroups (American Indian Or Alaska Native, Asian, Black Or African American, Native Hawaiian Or Other Pacific Islander, Hispanic Or Latino Ethnicity, and White) was based on students for whom one and only one race/ethnicity group was checked. The category, “Two Or More Races” included only students for whom two or more race/ethnicity groups were checked. The category, “Unknown Race” included only students for whom no race/ethnicity group was checked.

Table 16 presents demographic group proportions, mean scale scores, and percent proficient by year, and changes from 2016 to 2017. There were no demographic changes in the student population from 2016 to 2017 that would have had an appreciable impact on the difference in achievement between 2016 and 2017. The largest change was a decrease in the percent of students with ‘unknown’ race. This change is likely due to differences between years in how race was assigned to students, with more accurate classifications of students in 2017.

Table 16. Demographic Changes by Subject Area from 2016 to 2017 (Sample D).

Subject	Demographic Group	Proportion 2016	Proportion 2017	Mean Scale Score (SS) 2016	Mean Scale Score (SS) 2017	% Prof. 2016	% Prof. 2017	Change Proportion	Change Mean SS	Change % Prof.
Math	Overall	n.a.	n.a.	2504.7	2505.0	39.3	39.9	0.0	0.3	0.6
	Female	48.9	48.9	2506.8	2506.8	39.2	39.7	0.0	0.0	0.5
	Male	51.1	51.1	2502.7	2503.2	39.4	40.0	0.0	0.5	0.6
	American Indian or Alaska Native	1.2	1.2	2453.0	2450.3	19.6	19.5	0.0	-2.8	-0.1
	Asian	8.3	8.6	2584.1	2588.6	70.1	70.9	0.2	4.5	0.8
	Black or African American	5.2	5.1	2454.3	2452.1	20.2	20.8	-0.1	-2.2	0.6
	Native Hawaiian or Other Pacific Islander	1.2	1.1	2476.5	2474.5	27.5	27.1	0.0	-2.0	-0.3
	Hispanic or Latino	41.5	42.2	2470.6	2471.2	24.7	25.8	0.7	0.6	1.1
	White	33.9	33.7	2534.1	2534.3	52.4	52.4	-0.2	0.1	0.0
	Two or More Races	5.8	6.2	2505.4	2507.8	42.0	43.2	0.4	2.4	1.3
	Unknown Race	3.4	2.2	2540.1	2539.5	48.6	51.6	-1.1	-0.6	3.0
	LEP Status	18.2	17.4	2418.2	2416.9	13.2	13.2	-0.8	-1.3	0.1
	IDEA Indicator	10.0	11.1	2407.3	2406.7	11.5	12.0	1.1	-0.6	0.6
	Section 504 Status	1.6	1.7	2517.5	2516.3	38.4	38.5	0.1	-1.1	0.1
	Economic Disadvantage Status	57.1	55.7	2471.0	2470.2	25.6	26.2	-1.4	-0.8	0.6
ELA	Overall	n.a.	n.a.	2513.2	2512.3	50.1	50.1	0.0	-0.9	-0.1
	Female	48.9	48.9	2527.1	2526.1	55.8	55.5	0.0	-1.0	-0.3
	Male	51.1	51.1	2499.9	2499.2	44.7	44.9	0.0	-0.7	0.2
	American Indian or Alaska Native	1.2	1.2	2461.1	2456.5	28.2	27.2	0.0	-4.6	-1.0
	Asian	8.2	8.5	2572.6	2575.1	73.6	74.1	0.3	2.4	0.4
	Black or African American	5.2	5.1	2471.9	2469.2	32.6	32.5	-0.1	-2.7	-0.1
	Native Hawaiian or Other Pacific Islander	1.2	1.1	2483.8	2479.8	37.0	35.5	0.0	-4.0	-1.5
	Hispanic or Latino	41.2	42.1	2483.0	2483.4	36.7	37.3	0.9	0.4	0.6
	White	33.9	33.8	2540.7	2539.3	63.0	62.4	-0.1	-1.3	-0.5
	Two or More Races	5.8	6.2	2515.2	2515.5	53.3	53.6	0.5	0.3	0.3
	Unknown Race	3.6	2.2	2550.9	2550.8	62.6	64.4	-1.4	-0.2	1.8
	LEP Status	18.0	17.2	2415.0	2411.7	13.9	13.2	-0.8	-3.3	-0.7
	IDEA Indicator	10.0	11.1	2417.6	2415.7	14.2	14.7	1.1	-1.9	0.5
	Section 504 Status	1.6	1.6	2526.0	2524.0	50.4	49.8	0.1	-2.0	-0.5
	Economic Disadvantage Status	57.1	55.6	2481.2	2479.7	36.5	36.5	-1.4	-1.5	-0.1

There was an increase from 2016 to 2017 in the proportion of students classified under IDEA (Individuals with Disabilities Education Act), for whom mean achievement is significantly below that of non-IDEA students. If this increase was caused by more IDEA students having moved in to the Smarter Balanced test-taking population in 2017, they would have caused a decrease in the overall population mean scale score of only 0.1 point from 2016 to 2017. It seems possible that IDEA students increasingly take the Smarter Balanced test it becomes increasingly recognized for its accessibility to special populations. But it is also possible that increase in IDEA students was due only to a change in the classification of students who were in the Smarter Balanced test-taking population both years.

Summary and Conclusions

The 3-page summary that accompanies this report by the same title, *Answers to Questions about 2017 Test Results*, was developed in consultation with Smarter Balanced's technical advisors. Please refer to that summary for plain-language answers to the questions posed at the beginning of the report. More technical, but brief summary statements and conclusions follow here:

Analyses of the measurement properties of the assessment in 2016 and 2017 show that the test measured students as precisely in 2017 as in 2016. The newly-added items in 2017 were slightly easier than their older, CAT-pool counterparts as intended, but this did not mean that students received an easier test or that student measures were adversely affected. The item response theory (IRT) models used in Smarter Balanced assessments assure that students' measures of achievement are comparable even if they receive tests of different difficulty. The computer adaptive algorithm further ensures that the items students receive are at a level of difficulty that is appropriate and does not change with overall changes in the difficulty of the item pool.

The most critical analysis in this report was the residual analysis, performed to answer the question of whether students' scores in 2017 were impacted by the newly-added items. This analysis showed that students performed slightly better than expected on the older CAT items, in contrast to the newly-added CAT items. This result is not surprising or unusual. The old CAT items were used in three previous administrations, and items tend to become easier over time through exposure. The magnitude of difference between new and old CAT items was very small (0.018 for ELA/literacy and 0.017 for mathematics) and well within the range seen with item drift due to differential learning and item exposure. The difference between ELA/literacy CAT and PT item residuals in this study (0.067) may also be attributed to item drift and was over three times as large (as the difference between old and new CAT items in ELA/literacy (.018) and mathematics (.017)). Effects of item exposure are typically mitigated in high stakes assessments by periodically adding new items to the assessment and retiring old items. At Smarter Balanced, this practice is guided by input from members and technical advisors and informed by studies that include analyses of residuals.

Given the magnitude of difference between old and new CAT item mean residuals and the fact that these item pools differed mainly in their exposure, it is extremely unlikely these differences were due to procedural or technical errors in the field testing and equating of the new CAT items.

It also seems unlikely that the new CAT items had a substantial effect on 2017 test results. This assertion is based primarily on the fact that the difference between old and new CAT item residuals at each grade does not correspond well with the change in student achievement from 2016 to 2017. At grade 5 in ELA/literacy, the difference in mean residuals was small (.008), but the decline in test scores was large (-5.3 scale score points). At grade 7 in ELA/literacy, the difference in mean residuals was relatively large (.025), but there was actually a slight gain in student achievement from 2016 to 2017 (0.1 point). Overall, the grade-level difference between mean residuals (old and new CAT) did not correlate with grade-level change in achievement from 2016 to 2017.

It is conceivable that some of the gains between 2015 and 2016 were in part based on students becoming more familiar with the format of certain parts of the test. Smarter Balanced will use the recommendations from the National Center for the Improvement of Educational Assessment and in

collaboration with its members and its TAC, establish additional procedures to monitor and potentially discriminate better between growth due to increased familiarity with the test and true improvement in student learning.