SIMULATION-BASED EVALUATION OF THE SMARTER BALANCED SUMMATIVE ASSESSMENTS

JUNE 29, 2015

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Executive Summary

This report describes a study simulating the adaptive administration of the Smarter Balanced summative assessments. The study was conducted in order to examine properties of the simulated tests (such as blueprint fulfillment and item exposure) and the quality of the examinee score estimates (including bias and precision). Simulations were conducted for both English language arts/literacy and Mathematics and in all the tested grade levels (3-8 and high school). For each grade and subject, three item pools were evaluated: the general pool, the Spanish-translated pool, and the Braille pool. The simulated tests included both the computerized adaptive test (CAT) and performance task (PT) components, thus mimicking the operational summative tests.

In order to conduct the simulation study, CRESST designed and programmed a CAT engine that determines the next CAT item to be administered by weighting the item information functions of available items by the current characterization of student proficiency (i.e., the current posterior distribution) and "tuning" parameters related to test blueprint requirements. PT items were delivered as a set, with sets assigned completely at random (i.e., not based on any prior or current information about the examinee's proficiency).

Within each grade band and subject, 1000 simulated examinees were administered the summative assessment. To evaluate the test administration, we examined the extent to which each test instance met the test blueprint. We also examined the proportion of tests in which each item was used (i.e., exposure rate). The item response vectors were scored according to the operational specifications to generate overall and claim scores, with corresponding standard errors of measurement. We examined the extent to which the true (generating) proficiency scores were recovered, as well as the precision of the score estimates.

Overall, we found that the CAT engine used in this study provided good estimation of student proficiency while maintaining very low item exposure rates for the vast majority of items. These results are largely similar to those reported in an earlier simulation study conducted by researchers at American Institutes for Research (AIR). Of some concern, however, is the fact that we observed some difficulty in consistently fulfilling all requirements in the blueprint documents—particularly those requirements related to the number of items in the PT component for each claim.

Introduction

This document presents study design details and results for the CRESST analysis of the Smarter Balanced Summative Assessments in both English language arts/literacy (ELA/L) and Mathematics by the method of Monte Carlo simulation. Simulations were conducted for the general, Spanish, and Braille item pools. At the student level, the summative assessments include a computerized adaptive testing (CAT) portion and a set of items under the performance tasks (PT) portion. A key design document of the summative assessments is the test blueprint, which specifies the number and nature of items to be administered, both individually and in relation to one another, to help ensure high technical quality of the newly developed Smarter Balanced assessment system.

The main components of the study reported here are: 1) the creation of a flexible mechanism for pool assembly; 2) the development of CAT engine that can implement the blueprint specifications plus additional constraints that may arise from design and operational considerations (present and future); 3) simulated administration of the summative assessments, including both CAT and PT portions; and 4) an examination of simulation study results, including characteristics of the summative assessments such as student proficiency estimation and item exposure. This report is primarily concerned with the third and fourth components.

It should be noted that American Institutes for Research (AIR) conducted a prior study of the CAT portion of the summative assessment (2014-2015 Smarter Balanced Summative Simulation Report). Given this existing report, it is natural to ask how the AIR and CRESST studies and results compare with one another. Some of the major distinctions in methodology are presented in the next section. The existence of these differences makes a direct comparison challenging. Nevertheless, to facilitate comparison, many of the same statistical summaries and table formats used in the AIR report are adopted in this report. Overall, it is clear that the results in both reports are in general agreement. For example, the statistical summaries concerning the estimates of student proficiency are quite similar.

The report is organized as follows. In the next section, the main differences in the AIR and CRESST methodologies are presented. Then, a brief outline of the CRESST CAT engine design is presented. Next, the statistical summaries used for the simulation are presented, followed by descriptive statistics of the operational item pool. After these sections, simulation results are presented. These results include summaries of blueprint satisfaction, as well as bias and precision of proficiency estimates based on both CAT

and PT items. (To facilitate a more detailed comparison with results from the AIR report, results of student proficiency estimation based on only the CAT items are presented in Appendix A.) Additionally, for CAT items only, item exposure results are presented. Results specific to the Braille and Spanish tests are presented in Appendices D and E, respectively. We offer conclusions and suggestions at the end.

Differences Between the AIR and CRESST Studies

In this section, we briefly describe some of the key differences between the prior AIR simulation study and the analyses presented in this report. These differences should be kept in mind when making any comparisons across the two study reports.

The first difference is that this study simulates the administration of tests that include both the PT and CAT portions of the assessment. The prior AIR study—due to its focus on the item selection algorithm—only dealt with the CAT portion. One consequence of this difference is that in the CRESST simulations, a greater number of items are administered to each simulee. Additionally, this report includes some results specific to the PT component, such as the number of operational items in the PT pool.

Second AIR and CRESST use different CAT algorithms for selecting items to be administered. The AIR CAT engine selects items based on their difficulty parameters and the current point estimate of a student's proficiency, θ. In contrast, the CRESST engine utilizes item information (a continuous function related to both difficulty and discrimination parameters of an item), as well as the current estimate of the posterior distribution of θ (a density, rather than a point estimate). A related difference in the AIR and CRESST simulation studies is how proficiency estimates are initialized. In the AIR study, students' true scores were used for the initial proficiency estimate. In the current study the population proficiency distribution (for a given grade and subject) was used. The CRESST engine is described in more detail in the next section.

Third, the item pools used by AIR and CRESST were slightly different (see also Appendix B) due to changes (updates) in the lists of active/operational items at the time that each study was conducted. The item pools also differed in that the AIR simulations allowed for item pool expansion (to above- or below-grade items) for students with extreme proficiency estimates, while the CRESST simulated tests did not.[1] We note that AIR's results indicate relatively few instances where such expansions occurred.

[1] We did, however, examine eligibility for pool expansion based on interim score estimates and the eligibility criterion of being either above the level 3/4 cut or below the level 1/2 cut with $p<10^{-7}$ after completing two-thirds of the CAT items or at point following. No simulated ELA examinees

Fourth, there may have been differences in the versions of the blueprint documents used in the two studies (again due to the timing of the two studies; see Appendix B for details concerning the particular blueprint versions used in this study). It's also likely that the blueprints were operationalized within the AIR and CRESST CAT engines in different ways. To the best of our knowledge, CRESST has faithfully implemented the blueprint for our CAT/PT simulations to reflect the SBAC design intentions.

Finally, in the CRESST simulation estimates of θ more extreme than the specified highest and lowest obtainable theta (HOT and LOT) were not excluded from computation of certain statistical summaries. (The number of cases where this occurred is presented in Table 1.) Instead, they were set to either the highest or lowest obtainable score, as appropriate. Additionally, standard errors for these cases were calculated using test information for items administered to the simulee, per the formally adopted Scoring Specification (American Institutes for Research, 2014).

CRESST CAT Engine Design

The purpose of this section is to provide a brief description and list a few key features of the CAT engine designed and created by CRESST.

The engine was written and implemented in R (R Core Team, 2014). For ELA/L, the engine proceeds claim by claim, where the order in which the claims appear is randomized for each student. Further, for Mathematics, within a claim, the engine proceeds "cell" by "cell" (also in a randomized order). Here, "cell" refers to a collection of Assessment Targets for which the Blueprint requires a specified number of items. For instance, in Grade 3 Mathematics, Targets B, C, I, and G define a cell where the Blueprint requires 5-6 items. Given the design of the engine and the Blueprint complexity, ELA/L cannot proceed cell by cell, as some Blueprint requirements or stimuli span multiple cells. For each cell and claim, there are maximum numbers of items/stimuli that may permissibly be administered. In order to satisfy the blueprint requirements, the engine always tries to administer the maximum number of items. For CAT, the algorithm does not allow administration of an item that would result in a maximum requirement being exceeded.

The engine proceeds adaptively in the following manner. Instead of utilizing a current point estimate of the student proficiency, θ, the engine utilizes a current estimate of the posterior distribution of the student proficiency. The engine starts the posterior at the

---

met this criterion. The number of cases eligible in Math ranged from 0 to 16 (1.6%), and all cases eligible for expansion were above the level 3/4 cut.

generating (population) proficiency distribution for the Grade/Subject combination that the simulee comes from. In contrast, the AIR simulation starts the simulated CAT at the simulee's true θ. To select the next item/stimulus, a weighted information value is calculated for all eligible items (i.e., those items within the current cell that have not been used). This "baseline" weight is found by integrating item information function values over the current estimate of the posterior distribution of θ, that is, posterior-weighted item information values. A key advantage of posterior-weighted item information values is that they are more global and less "greedy" measures of optimality than either the item difficulty parameter alone or the Fisher information function. Considerable research (e.g., Chang & Ying, 2008) has indicated the negative consequences of greedy optimization (e.g., maximum Fisher information) algorithms in educational CAT situations. The CRESST engine follows the wisdom from contemporary CAT research.

The weights are sometimes further adjusted or tuned to ensure that the Blueprint is satisfied with sufficient frequency. That is, items/stimuli that meet requirements of the Blueprint that are difficult to meet have weights that are multiplied by a constant (empirically fine-tuned via additional simulations and trials not reported here). Selection of the next item/stimulus occurs by normalizing these weights and treating them as sampling probabilities. Thus, items/stimuli with larger weights have a higher chance of being administered to the simulee. The consequence of not picking the item/stimuli with the largest weight is that CRESST's algorithm may perform better in terms of item exposure, but with a slight decrement in measurement precision. Weights for items that shared a common stimulus were initially summed such that such items shared a common weight and sampling probability and were treated as a single unit – analogous to a single item. If, however, the next unit selected for administration was a stimulus that contained multiple items, the algorithm proceeded adaptively by selecting items within that stimulus until a specified maximum number of items for the stimulus was reached. Upon administration of an item, a response based on the item's parameters and the simulee's true θ was generated, and the posterior for the simulee was updated (for all machine-scored items).

The CRESST CAT terminates after all claims are cycled through. Typically, the Blueprint is satisfied by the sequence of administered items. However, given the design, there are reasons that an administered test may not meet the Blueprint. First, in some cases, within a cell or claim, there are no remaining eligible items that help to satisfy the Blueprint without violating some other aspect of the Blueprint. For example, in ELA/L Grade 5, Claim 1, it is possible to reach the maximum number of items allowed for the "Informational" category before Targets 9 or 11 meet the minimum number of items. In

this rare event, the engine moves on to the next cell or claim. Such a case may occur because weight tuning is ineffective in ensuring that the Blueprint is satisfied. For instance, despite weight tuning, it is also possible that fewer high DOK items are selected than are required by the Blueprint.

In selecting performance tasks (PT items), the algorithm randomly assigned stimuli to meet the stimulus requirements specified in the blueprint. All available items for the administered stimulus were administered to the simulee. In some cases, this may result in min/max item requirements to be violated if a particular stimulus has items associated with it that lead to such violations.

Simulating and Estimating Student Proficiency

True values for student proficiency (θ) were drawn from a normal distribution with parameters specific to grade and subject. These population parameters were the same as those used in the AIR study; their values are presented in Table 1.

Table 1. Characteristics of Simulated and Estimated Proficiencies

Grade	Population Parameters		Obtainable Proficiency Range		Percentage of Winsorized Scores	
	Mean	SD	LOT	HOT	LOT	HOT
			English Language Arts/Literacy			
3	-1.24	1.06	-4.59	1.34	0.7	1.1
4	-0.75	1.11	-4.40	1.80	0.3	2.0
5	-0.31	1.10	-3.58	2.25	1.0	2.1
6	-0.06	1.11	-3.48	2.51	0.8	1.7
7	0.11	1.13	-2.91	2.75	1.4	1.6
8	0.38	1.13	-2.57	3.04	1.5	1.9
11	0.53	1.19	-2.44	3.34	1.5	1.4
			Mathematics			
3	-1.29	0.97	-4.11	1.33	0.5	0.9
4	-0.71	1.00	-3.92	1.82	0.3	1.1
5	-0.35	1.08	-3.73	2.33	1.0	1.6
6	-0.10	1.19	-3.53	2.95	0.8	1.1
7	0.01	1.33	-3.34	3.32	2.2	1.2
8	0.18	1.42	-3.15	3.63	2.8	1.2
11	0.51	1.52	-2.96	4.38	3.3	1.2

These normal distributions were also used as the starting distributions of θ for each simulee. That is, before any items are administered, these population or "prior" distributions are treated as current "posterior" distributions for all simulees. After

proceeding through the CAT and PT components, maximum likelihood (ML) scoring was conducted (to obtain the final score estimates). A complication with ML scoring is that perfect response patterns (i.e., all correct or all incorrect) result in θ estimates of $\pm\infty$, with undefined standard errors. Following the Scoring Specification, estimates more extreme than the HOT and LOT scores were Winsorized to the HOT and LOT scores, as appropriate. Again following the Scoring Specification, corresponding standard errors were calculated using test information for items administered to the simulee, computed at the HOT or LOT score. The HOT and LOT scores are also presented in Table 1. Finally, Table 1 contains the percentage of simulees whose scores were Winsorized.

Statistical Summaries

Generally, the same statistical summaries used in the AIR report (see Statistical Summaries section) are also used in this report to facilitate comparisons. These definitions are included here for completeness.

Given a true value, θ, and its estimate, $\hat{\theta}$, the following summaries are used:

$$bias = N^{-1}\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i),$$

$$MSE = N^{-1}\sum_{i=1}^{N}\left(\theta_i - \hat{\theta}_i\right)^2,$$

and

$$\text{var}(bias) = \frac{1}{N(N-1)}\sum_{i=1}^{N}\left(\theta_i - \bar{\hat{\theta}}\right)^2,$$

where $\bar{\hat{\theta}}$ is the average of the $\hat{\theta}_i$, and N denotes the number of simulees. Statistical significance of the bias is tested using a z-test:

$$z = \frac{bias}{\sqrt{\text{var}(bias)}},$$

where a p-value is reported for a two-tailed test. The average standard error is

$$mean(se) = \sqrt{N^{-1}\sum_{i=1}^{N}se\left(\hat{\theta}_i\right)^2},$$

where $se(\hat{\theta}_i)$ is the standard error of the estimated θ for simulee i. Finally, to find the proportion of simulees falling outside the 95% and 99% confidence intervals (i.e., lack of confidence interval coverage), a t-statistic is computed as

$$t = \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)}.$$

The absolute value of the t statistic is compared to critical values of 1.96 and 2.58 for the 95% and 99% confidence intervals, respectively.

Operational Items

Table 2 displays the total number of items in the operational pool (excluding extended item pool) used in ELA/L CAT simulations, whereas Table 3 displays these values for the PT portion of the test. The number of passages and distinct stimuli also appear in these tables. For Mathematics, results are separated for calculator/no calculator items in Table 4 for the CAT portion of the test, and PT items and stimuli are in Table 5. Note that administration of stimuli for PT results in items administered from across multiple claims.

Table 2. Number of Operational Items in ELA/L Adaptive Test Item Pool

Grade	Number of Items					Number of Passages		
	Total	Claim 1	Claim 2	Claim 3	Claim 4	Claim 1 Literary	Claim 1 Information	Claim 3 Listening
3	591	217	166	118	90	18	17	47
4	567	177	166	127	97	15	11	47
5	546	194	159	108	85	16	13	42
6	548	175	168	116	89	7	21	46
7	508	183	154	117	54	5	24	45
8	499	161	147	131	60	6	18	49
11	1455	499	379	334	243	29	59	121

Note. Item counts current as of 2015-04-03.

Table 3. Number of Operational Items in ELA/L Performance Task Item Pool

Grade	Number of Items					Number of Stimuli
	Total	Claim 1	Claim 2	Claim 3	Claim 4	Across Claims
3	62	0	28	0	34	14
4	85	0	38	0	47	19
5	95	0	40	0	55	20
6	61	0	28	0	33	14
7	79	0	38	0	41	19
8	94	0	42	0	52	21
11	105	0	48	0	57	24

Note. Item counts current as of 2015-04-03.

Table 4. Number of Operational Items in Mathematics Adaptive Test Item Pool

Grade	Calculator	Total	Claim 1	Claim 2	Claim 3	Claim 4
3	No	829	547	76	123	83
4	No	818	519	91	116	92
5	No	807	459	81	146	121
6	Yes	368	151	70	88	59
	No	371	360	0	11	0
7	Yes	459	241	67	97	54
	No	211	211	0	0	0
8	Yes	464	257	43	108	56
	No	148	148	0	0	0
11	Yes	1555	859	159	371	166
	No	156	119	0	37	0

Note. Item counts current as of 2015-04-03.

Table 5. Number of Operational Items in Mathematics Performance Task Item Pool

Grade	Number of Items					Number of Stimuli (Across Claims)
	Total	Claim 1	Claim 2	Claim 3	Claim 4	
3	102	0	40	32	30	18
4	98	0	36	30	32	19
5	85	0	27	28	30	15
6	105	0	36	37	32	18
7	87	0	33	23	31	16
8	86	0	30	30	26	18
11	89	0	31	34	24	17

Note. Item counts current as of 2015-04-03.

Simulation Results: CAT and PT

The simulation for both CAT and PT components of the assessment was conducted for $N = 1{,}000$ simulees per grade and subject. This section reports simulation results for CAT and PT with regards to blueprint satisfaction and student proficiency.

Blueprint Satisfaction

In most cases, overall claim-level requirements (Tables often labeled as "Blueprint Table" in the first few pages of SBAC blueprint documents) were met or exceeded. To save space here, details appear in the four tables in Appendix C, which present the page number in which the requirement appears, the min and max requirement, and percentages of test administration meeting these particular requirements. For CAT, Tables C1 and C3 shows that all ELA/L and Mathematics test instances met the blueprint constraints for passages (100%) and total number of items (100%) within each claim. In the ELA/L PT simulation, most tests met the blueprint requirements for stimuli and total number of items within each claim (see Table C2). However, 4.8% of the test instances in Grade 4 did not meet the minimum number of items for Claim 4. In the Mathematics PT simulation, all the grades except for Grade 3, Grade 6, and Grade 11, met the blueprint constraints exactly for claims. In these exceptions, there were some simulated cases where the number of administered items exceeded some maximum requirement. Specifically, 17.5% of the test instances in Grade 3 exceeded the maximum number of items for Claim 2/4; 5.5% of the test instances in Grade 3 exceeded the maximum number of items for Claim 3; 4.3% of the test instances in Grade 6 exceeded the maximum number of items for Claim 3; 5.9% of the test instances in Grade 11 exceeded the maximum number of items for Claim 3.

Additional blueprint requirements include more detailed requirements about the sampling of items within the above overall requirements (Tables often labeled as "Target Sampling" in the latter pages of SBAC blueprint documents). Due to many such requirements, here we only report on blueprint violations. Tables 6 and 7 list violations for the CAT and PT portions of the simulated tests. These tables show, by subject and grade, the blueprint specification, the page number in the blueprint document in which the requirement appears, the min and max for the requirement, the number of tests in which the blueprint violation occurred, the number of tests in which the blueprint minimum number of items was not met, and the number of tests in which the blueprint maximum was exceeded.

For the target and DOK-level constraints, the ELA/L CAT met the blueprint specifications with the following exceptions (see Table 6): nine Grade 4 tests did not

have the minimum number of items at DOK level 2; one Grade 5 test did not have the minimum number of Claim 1, Target 9 items; one Grade 6 test did not have the minimum number of Claim 1, Target 2 items; and 44 Grade 6 tests did not have the minimum number of items with DOK level 2 or greater. In Mathematics, all CAT portions met the blueprint requirements for targets and DOK.

Table 6. CAT Tests with Blueprint Violations (CAT Portion Only)

			Blueprint Requirement			Number of Tests		
Grade	Subject	Blueprint Specification	Pg. #	Min	Max	Total	Below Min.	Above Max.
4	English	Claim 1, DOK=2	4	6	6	9	9	0
5	English	Claim 1 (Informational), Target 9: Central Ideas	4	1	2	1	1	0
6	English	Claim 1 (Literary), Target 2: Central Ideas	7	1	1	1	1	0
6	English	Claim 2, DOK≥2	7	5	--	43	43	0

In the ELA/L PT simulation, most tests met the blueprint requirements (see Table 7). However, 48 tests in Grade 4 did not meet the minimum number of items for Claim 4; these tests also fell short of the required number of Claim 4 items with DOK level of 3 or greater. We also note that for all grades, only two items were provided within any "Full Write" stimulus in the metadata received from SBAC, whereas the blueprint specifies three item scores per stimulus. This would cause all ELA/L PT tasks (in all grade) to be in violation of the blueprint (having too few Claim 2 items). Based on communication with Smarter Balanced, however, the fact that only two item scores are produced is consistent with the current approach to scoring the Full Write items. Although raters do indeed assign three scores to a student's written response, two of these scores are combined into a single score.[2] Thus, our implemented blueprint specifications were modified to allow either two or three items for ELA/L Claim 2 (Writing).

There was a substantial number of Mathematics PT portions that did not match the blueprint constraints exactly. In some grades, the violation only occurred within a single claim. However, in grades 3 and 11, there were instances of blueprint violations in claims 2, 3, and 4. In some cases, not meeting a minimum blueprint requirement may be a result of the fact that some PT items in the operational pool lack item parameters. That

[2] This combining of scores seems to have been done as a way of dealing with the very high correlation between the separate scores, which resulted in poor calibration results from the unidimensional IRT calibration (which assumes local item independence).

is, some PT items tagged by SBAC as operational do not have associated item parameters and therefore cannot be included in our simulations. In other cases, an insufficient number of items is available in the operational pool for the stimulus administered. Mathematics PT is the only example where a maximum number of items was exceeded.

Table 7. PT Tests with Blueprint Violations (PT Portion Only)

			Blueprint Requirement			Number of Tests		
Grade	Subject	Blueprint Specification	Pg. #	Min	Max	Total	Below Min.	Above Max.
4	English	Claim 4 (Research)	6	2	3	48	48	0
4	English	Claim 4, DOK>=3	6	2	3	48	48	0
3	Math	Claim 2 (Problem Solving)	5	1	2	353	0	353
3	Math	Claim 4 (Modeling and Data Analysis)	5	1	3	164	116	48
3	Math	Claim 3 (Communicating Reason)	5	0	2	55	0	55
4	Math	Claim 2 (Problem Solving)	7	1	2	48	0	48
4	Math	Claim 4 (Modeling and Data Analysis)	7	1	3	44	44	0
5	Math	Claim 2 (Problem Solving)	9	1	2	131	0	131
6	Math	Claim 3 (Communicating Reason)	11	0	2	43	0	43
7	Math	Claim 2 (Problem Solving)	13	1	2	61	0	61
8	Math	Claim 4 (Modeling and Data Analysis)	15	1	3	223	223	0
11	Math	Claim 2 (Problem Solving)	17	1	2	59	0	59
11	Math	Claim 4 (Modeling and Data Analysis)	17	1	3	104	104	0
11	Math	Claim 3 (Communicating Reason)	17	0	2	59	0	59

Student Proficiency

The distributions for student proficiency and for item difficulty of within each grade and subject, are summarized in Table 8. The results are similar to those reported by AIR, and highlight the challenge of administering informative items via the CAT.

Table 8. Summaries of Difficulty of Item Pool and Estimated Student Proficiency

Grade	English Language Arts/Literacy				Mathematics			
	Items		Proficiency		Items		Proficiency	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	-0.36	1.14	-1.30	1.13	-0.80	1.04	-1.34	1.01
4	0.17	1.25	-0.82	1.16	-0.06	1.02	-0.78	1.05
5	0.54	1.16	-0.36	1.16	0.71	1.01	-0.44	1.18
6	1.00	1.30	-0.12	1.15	1.03	1.19	-0.18	1.26
7	1.12	1.27	0.05	1.17	1.77	1.17	-0.07	1.39
8	1.30	1.29	0.30	1.19	2.28	1.40	0.10	1.48
11	1.70	1.33	0.46	1.25	2.70	1.55	0.40	1.61

As discussed above, an estimate of student proficiency and an accompanying standard error was obtained for all simulees in accordance with the Scoring Specification. Tables 9 and 10 present various statistics describing score recovery for the ELA and Math summative assessments, respectively. These quantities—including mean bias and mean square error (MSE)—are defined in the section *Statistical Summaries*.

The mean biases in overall scores are all relatively small, and the null hypothesis that the mean bias is equal to zero in the population cannot be rejected (p-values range from 0.33 to 0.99). However, there is evidence of bias in claim score estimates. This bias appears to be due to the use of the LOT and HOT values for examinees with extreme score estimates for a given claim, including perfect score patterns (i.e., achieving either the minimum score for all items or the maximum for all items) with an infinite ML score estimate. Perfect score patterns are of course far more likely within a claim (based on a relatively small number of items) than for the full set of CAT items. Importantly, we observed that extremely low scores were far more frequent than extremely high scores. As a result, more extreme scores were replaced with the LOT than with the HOT value, producing the positive bias observed in the claim score results.[3] It should be noted,

[3] To illustrate, 245 simulated HS (Grade 11) examinees were assigned the LOT value for Math claim 2/4. All 245 had response patterns with all incorrect scores (with an average of about 9 items administered). In contrast, only 14 examinees were assigned the HOT value for claim 2/4 (4 of whom had patterns will all correct responses). There was, as a result, substantial positive bias (mean bias of +0.39, as shown in Table 10).

however, that the assignment to LOT or HOT values would have little impact on classifications by performance level, which is the way claim scores would be used in practice.

Table 9. Bias of the Estimated Proficiencies: English Language Arts/Literacy

Grade	Mean Bias	SE of Mean Bias	p-value for the *z*-Test	MSE	95% CI Miss Rate	99% CI Miss Rate
			Overall			
3	0.00	0.03	0.89	0.10	4.6	1.3
4	0.01	0.04	0.81	0.11	6.2	1.2
5	-0.01	0.03	0.75	0.10	4.8	1.0
6	0.00	0.04	0.92	0.11	4.5	0.4
7	0.01	0.04	0.87	0.12	4.2	1.2
8	0.02	0.04	0.59	0.11	4.1	0.5
11	0.00	0.04	0.98	0.14	5.7	1.2
			Claim 1: Reading			
3	0.09	0.03	0.01	0.35	6.8	2.8
4	0.06	0.04	0.07	0.38	5.4	1.9
5	0.04	0.04	0.20	0.32	5.7	1.8
6	0.07	0.04	0.04	0.43	4.5	1.5
7	0.06	0.04	0.12	0.42	5.6	1.2
8	0.08	0.04	0.03	0.39	5.5	2.0
11	0.04	0.04	0.34	0.43	5.7	2.0
			Claim 2: Writing			
3	0.01	0.03	0.75	0.32	5.4	1.1
4	0.02	0.04	0.64	0.32	6.7	1.6
5	-0.02	0.03	0.62	0.32	6.2	1.3
6	0.02	0.04	0.55	0.34	4.7	1.3
7	0.05	0.04	0.17	0.38	6.2	2.1
8	0.02	0.04	0.58	0.30	3.2	1.2
11	0.04	0.04	0.32	0.49	6.1	1.5
			Claim 3: Speaking/Listening			
3	0.11	0.03	0.00	0.87	9.3	5.7
4	0.10	0.04	0.01	0.87	8.2	5.0
5	0.09	0.04	0.01	0.83	8.5	5.1
6	0.10	0.04	0.00	0.87	7.9	3.8
7	0.02	0.04	0.56	0.74	5.9	2.8
8	0.05	0.04	0.13	0.83	7.6	4.0
11	0.00	0.04	0.98	0.93	7.2	3.8
			Claim 4: Research			
3	0.17	0.03	0.00	0.87	12.9	7.9
4	0.19	0.04	0.00	0.99	10.9	6.8
5	0.05	0.04	0.13	0.59	8.8	5.1
6	0.19	0.04	0.00	1.01	12.9	8.2
7	0.20	0.04	0.00	0.98	14.6	8.8
8	0.11	0.04	0.00	0.79	11.0	6.8
11	0.18	0.04	0.00	0.88	12.2	7.4

Table 10. Bias of the Estimated Proficiencies: Mathematics

Grade	Mean Bias	SE of Mean Bias	p-value for the *z*-Test	MSE	95% CI Miss Rate	99% CI Miss Rate
			Overall			
3	0.00	0.03	0.99	0.06	4.5	0.9
4	0.01	0.03	0.69	0.08	5.5	1.6
5	0.03	0.03	0.33	0.13	4.5	1.3
6	0.01	0.04	0.80	0.11	4.2	0.8
7	0.00	0.04	0.93	0.19	5.3	1.0
8	0.00	0.05	0.99	0.20	4.3	0.8
11	0.02	0.05	0.72	0.25	4.8	1.2
			Claim 1: Concepts and Procedures			
3	-0.01	0.03	0.83	0.12	5.3	0.8
4	0.03	0.03	0.28	0.15	4.5	0.9
5	0.06	0.03	0.06	0.25	4.9	1.7
6	0.02	0.04	0.54	0.21	4.2	0.6
7	0.06	0.04	0.18	0.37	7.3	1.8
8	0.04	0.05	0.38	0.36	5.8	0.7
11	0.04	0.05	0.42	0.46	4.8	1.4
			Claim 2/4: Problem Solving/Modeling and Data Analysis			
3	0.10	0.03	0.00	0.39	8.4	4.9
4	0.13	0.03	0.00	0.55	10.1	5.2
5	0.29	0.04	0.00	1.03	15.5	9.1
6	0.17	0.04	0.00	0.82	12.4	6.7
7	0.23	0.04	0.00	1.30	15.8	7.3
8	0.36	0.05	0.00	1.64	20.1	10.2
11	0.39	0.05	0.00	1.73	18.2	9.9
			Claim 3: Communicating Reasoning			
3	0.17	0.03	0.00	0.62	12.2	8.2
4	0.15	0.03	0.00	0.55	8.7	5.4
5	0.20	0.03	0.00	0.76	11.1	6.1
6	0.22	0.04	0.00	0.89	11.2	6.0
7	0.29	0.04	0.00	1.29	12.9	6.9
8	0.13	0.05	0.01	0.96	9.5	3.9
11	0.20	0.05	0.00	1.20	9.1	3.6

Tables 9 and 10 show that the estimated standard errors in the overall and claim scores are calibrated well, as indicated by the 95% and 99% confidence interval miss rates. That is, miss rates around 5% and 1%, respectively, show that the estimated standard errors provide an accurate representation of the uncertainty in the score estimates. Finally, MSE values are relatively small, and are slightly improved versus use of the CAT alone.

For example, inclusion of PT items leads to improvement of MSE by approximately .01 to .04 depending on the Grade/Subject (compare with results in Appendix A).

Tables 11 and 12 presents some addition information concerning precision of the score estimates for both ELA and Math. For each score (overall and claim), we present the average number of items administered, the standard deviation of scores, the mean standard error of measurement, the root mean square error, and a marginal reliability coefficient.

Score precision is quite good for the overall scores. As expected (given fewer contributing items), claim score reliability is somewhat lower. For ELA, the ranges of marginal reliability estimates, across grade levels, were 0.75-0.80 for Claim 1, 0.75-0.82 for Claim 2, 0.58-0.63 for Claim 3, and 0.58-0.63 for Claim 4. Marginal reliability for the overall ELA score ranged from 0.91-0.93. For Math, the ranges of marginal reliability estimates, were 0.83-0.89 for Claim 1, 0.58-0.74 for Claim 2/4, and 0.57-0.66 for Claim 3. Marginal reliability for the overall Math score ranged from 0.90-0.94.

Table 13 presents the average standard error by grade and decile of the true overall proficiency score. This table shows general agreement with the results from an analysis the AIR CAT simulation ("CRESST Analysis of AIR CAT Simulation Data", dated 3/31/15). Most of the averages tend to be between 0.20 and 0.35, with a few exceptions. These exceptions are mostly concentrated in Decile 1, and have two primary causes. First, generally, there is a shortage of informative easy items, as indicated by Table 9. Second, all of the LOT cases are in Decile 1, and these cases have relatively high standard errors. The average standard errors for Mathematics in the higher grades are also relatively large, which again is a consequence of the items being relatively difficult.

Finally, Table 14 presents correlations between estimated proficiency and: 1) true proficiency (left column); and 2) overall test difficulty (right column). The overall test difficulty is simply the average item difficulty for a test. The correlations between estimated and true proficiencies are quite high (0.95-0.96), indicating that the administered items are successful in recovering the rank ordering of students; these correlations are on par or slightly higher than those based on CAT alone (see Appendix A). The correlations between estimated proficiency and overall test difficulty are lower. However, this is unsurprising as these correlations are computed with the inclusion of PT items that are not adaptive, and the CRESST CAT algorithm does not depend directly on item difficulty for item selection.

Table 11. Overall Score and Claim Score Precision/Reliability: English Language Arts/Literacy

Grade	Overall ELA/L					Claim 1					Claim 2					Claim 3					Claim 4				
	ave # items	SD($\hat{\theta}$)	mean SEM	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$
3	45.4	1.1	.31	.31	.92	16.0	1.3	.51	.59	.79	12.0	1.3	.54	.57	.80	9.0	1.5	.85	.93	.59	8.4	1.5	.71	.94	.60
4	45.5	1.2	.32	.33	.92	16.0	1.3	.58	.62	.78	12.0	1.3	.53	.56	.81	9.0	1.5	.85	.93	.59	8.5	1.5	.78	.99	.58
5	45.8	1.2	.31	.31	.93	16.0	1.3	.54	.57	.80	12.0	1.3	.53	.57	.81	9.0	1.4	.87	.91	.60	8.8	1.4	.67	.77	.70
6	43.3	1.2	.33	.33	.92	14.0	1.3	.66	.65	.75	12.0	1.3	.55	.58	.80	9.0	1.5	.88	.93	.60	8.3	1.5	.77	1.01	.58
7	43.1	1.2	.35	.35	.91	14.0	1.3	.65	.65	.75	12.0	1.3	.58	.62	.78	9.0	1.4	.87	.86	.63	8.2	1.5	.79	.99	.58
8	43.5	1.2	.34	.34	.92	14.0	1.3	.61	.62	.78	12.0	1.3	.56	.54	.82	9.0	1.5	.90	.91	.61	8.5	1.5	.78	.89	.62
11	45.4	1.2	.37	.37	.91	16.0	1.4	.63	.65	.77	12.0	1.4	.67	.70	.75	9.0	1.5	.95	.96	.58	8.4	1.5	.86	.94	.63

Table 12. Overall Score and Claim Score Precision/Reliability: Mathematics

Grade	Overall Math					Claim 1					Claim 2/4					Claim 3				
	ave # items	SD($\hat{\theta}$)	mean SEM	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$
3	39.7	1.0	.25	.25	.94	20.0	1.1	.35	.35	.89	9.9	1.2	.52	.63	.74	9.8	1.3	.61	.79	.63
4	39.2	1.1	.28	.28	.93	20.0	1.1	.38	.39	.88	9.6	1.3	.57	.74	.69	9.6	1.3	.62	.74	.67
5	39.7	1.2	.35	.36	.91	20.0	1.3	.48	.50	.84	9.8	1.6	.64	1.01	.61	9.9	1.4	.65	.87	.63
6	38.8	1.3	.35	.34	.93	19.0	1.3	.47	.46	.88	9.8	1.6	.67	.91	.67	10.0	1.6	.76	.94	.64
7	39.4	1.4	.44	.44	.90	20.0	1.5	.58	.61	.83	10.0	1.8	.81	1.14	.60	9.4	1.7	.95	1.14	.57
8	38.8	1.5	.46	.45	.91	20.0	1.5	.60	.60	.85	9.1	2.0	.86	1.28	.58	9.7	1.7	.88	.98	.66
11	41.3	1.6	.52	.50	.90	22.0	1.6	.69	.68	.83	9.3	2.1	.95	1.31	.60	10.0	1.9	1.04	1.10	.66

Table 13. Average Standard Errors by Grade and by Deciles of True Proficiency Scores

Grade	Proficiency Score Distribution										Overall
	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
English Language Arts/Literacy											
3	0.49	0.32	0.29	0.27	0.26	0.25	0.25	0.25	0.25	0.28	0.30
4	0.45	0.33	0.30	0.29	0.28	0.28	0.28	0.28	0.28	0.30	0.31
5	0.43	0.31	0.29	0.28	0.27	0.27	0.27	0.27	0.28	0.31	0.30
6	0.48	0.37	0.32	0.30	0.28	0.28	0.28	0.28	0.28	0.31	0.32
7	0.50	0.39	0.35	0.33	0.31	0.30	0.29	0.29	0.29	0.31	0.34
8	0.48	0.37	0.33	0.32	0.31	0.30	0.30	0.30	0.30	0.32	0.34
11	0.53	0.43	0.37	0.35	0.33	0.32	0.31	0.31	0.31	0.33	0.36
Mathematics											
3	0.38	0.28	0.25	0.23	0.22	0.22	0.21	0.21	0.21	0.23	0.25
4	0.45	0.32	0.27	0.25	0.23	0.22	0.21	0.21	0.21	0.23	0.26
5	0.63	0.42	0.34	0.30	0.27	0.25	0.23	0.21	0.21	0.22	0.31
6	0.57	0.41	0.36	0.32	0.31	0.28	0.26	0.25	0.24	0.25	0.33
7	0.72	0.58	0.48	0.42	0.37	0.32	0.29	0.26	0.23	0.23	0.40
8	0.73	0.57	0.50	0.45	0.40	0.37	0.34	0.30	0.27	0.26	0.43
11	0.85	0.67	0.57	0.52	0.45	0.39	0.34	0.31	0.27	0.26	0.47

Table 14. Correlations between True and Estimated Proficiency, and between Estimated Proficiency and Overall Test Difficulty

Grade	True Proficiency and Estimated Proficiency	Estimated Proficiency and Overall Test Difficulty
English Language Arts/Literacy		
3	0.96	0.69
4	0.96	0.72
5	0.96	0.74
6	0.96	0.68
7	0.96	0.65
8	0.96	0.68
11	0.95	0.60
Mathematics		
3	0.97	0.86
4	0.96	0.86
5	0.95	0.81
6	0.96	0.85
7	0.95	0.78
8	0.95	0.82
11	0.95	0.80

Note. Overall test difficulty is the average of item location parameters for all items in the test instance

Simulation Results: CAT Only

Item Exposure

Observed item exposure rates are influenced by the properties of the item, the CAT engine design, and selected tuning weights. Due to the differences in the CRESST and AIR CAT engines, there will be some differences in CRESST and AIR results for item exposure rates. We also expect some amount of sampling variability in these rates, though with $N = 1{,}000$ simulees, this variability should be relatively small.

For each item, we calculated the percentage of simulees who were administered the item. Next, each item was binned according to its percentage, where the bins are defined as follows: 0%-20%, 20%-40%, 40%-60%, 60%-80%, and 80%-100%. In some cases, an item was not administered to any simulees, and can be considered "unused." These exposure rates for both ELA/L and mathematics are presented in Table 15. Across both subjects and all grades, at least 90% of all items were administered to 0%-20% of the simulees. Only a small percentage of the items appeared on a high percentage of the tests.

Table 15. Percent of Items by Exposure Rate

Grade	Total Items	Exposure Rate					
		Unused	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
English Language Arts/Literacy							
3	591	1.35	97.29	1.35	0	0	0
4	567	0.35	97.00	2.65	0	0	0
5	546	5.86	91.58	2.20	0	0	0.37
6	548	4.56	91.42	3.65	0.37	0	0
7	508	5.71	90.16	3.94	0.20	0	0
8	499	1.00	94.79	4.21	0	0	0
11	1455	0.21	99.45	0.34	0	0	0
Mathematics							
3	829	0.48	99.16	0.36	0	0	0
4	818	0.12	99.14	0.73	0	0	0
5	807	0.12	99.38	0.50	0	0	0
6	739	0.14	99.05	0.81	0	0	0
7	670	0.15	98.66	1.19	0	0	0
8	612	0.00	98.04	1.80	0.16	0	0
11	1711	0.70	99.18	0.06	0	0	0.06

To more closely examine the 0%-20% bin, CRESST created histograms of the item exposure rates for the items in this range. These histograms are presented in Figures 1-2 (ELA/L) and 3-4 (mathematics). To clarify how to interpret the histogram, consider grade 3 ELA/L. From Table 15, we know that 97.29% of items in the pool were administered to 0%-20% of simulees. In Figure 1, then, the heights of the histogram bars for this grade and subject ("ELAG3"; top-left panel) sum to 97.29%. The left-most histogram bar, for example, shows that around 6% of items were administered to between 0% and 1% of the simulees.

The histograms in Figures 1-4 make it clear that the item exposure rates are right-skewed, particularly in mathematics. In contrast, as a CAT engine approaches uniform exposure, the histograms would become more sharply-peaked, and more symmetric. The right-skewness observed in the histograms can be interpreted as a departure from uniform exposure, with a small percentage of items exposed at relatively high rates. Overall, however, item exposure under CRESST's CAT algorithm is good, with very few unused items.

Figure 1. ELA/L: Histograms of Item Exposure Rates (Grades 3-6)

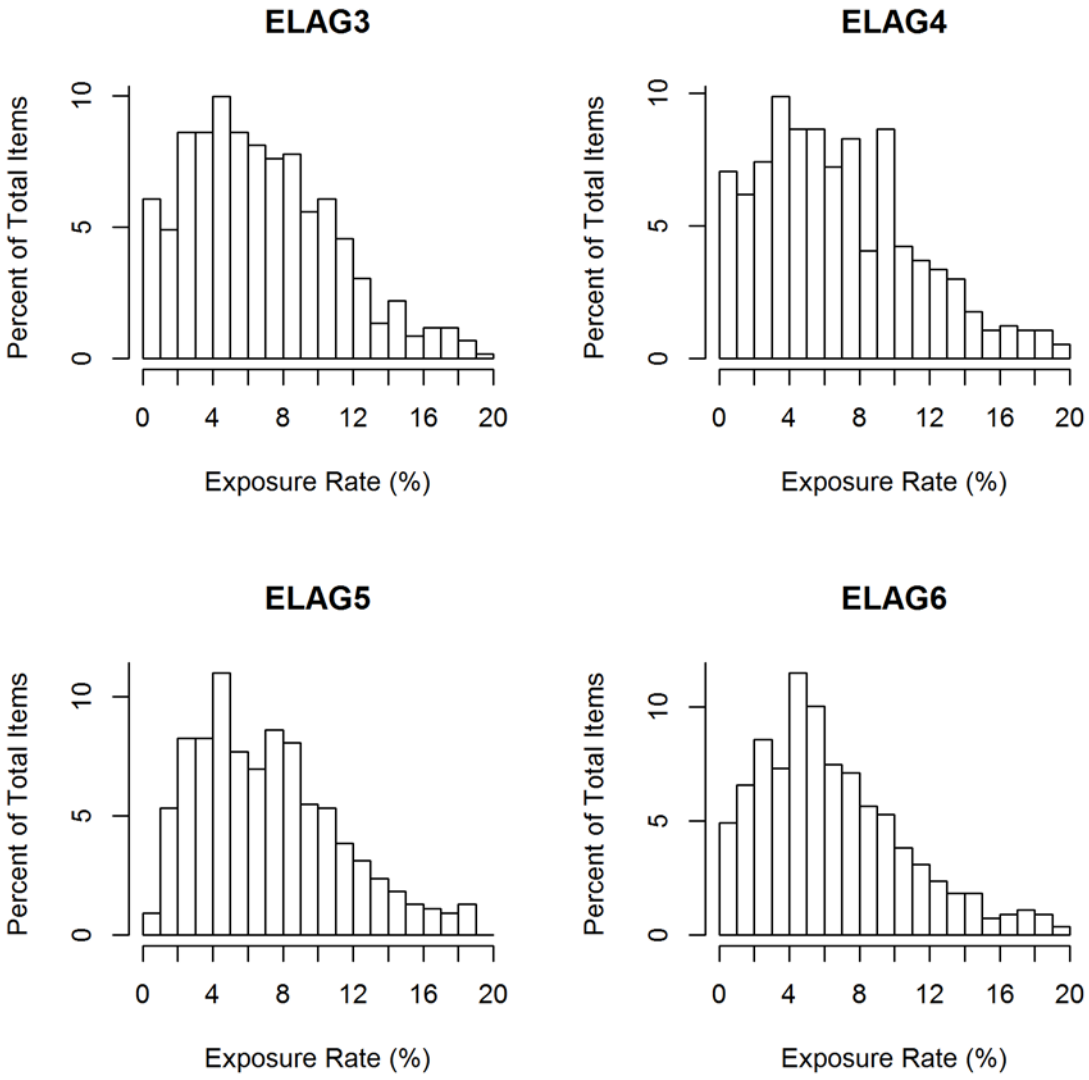


Figure 2. ELA/L: Histograms of Item Exposure Rates (Grades 7, 8, and HS)

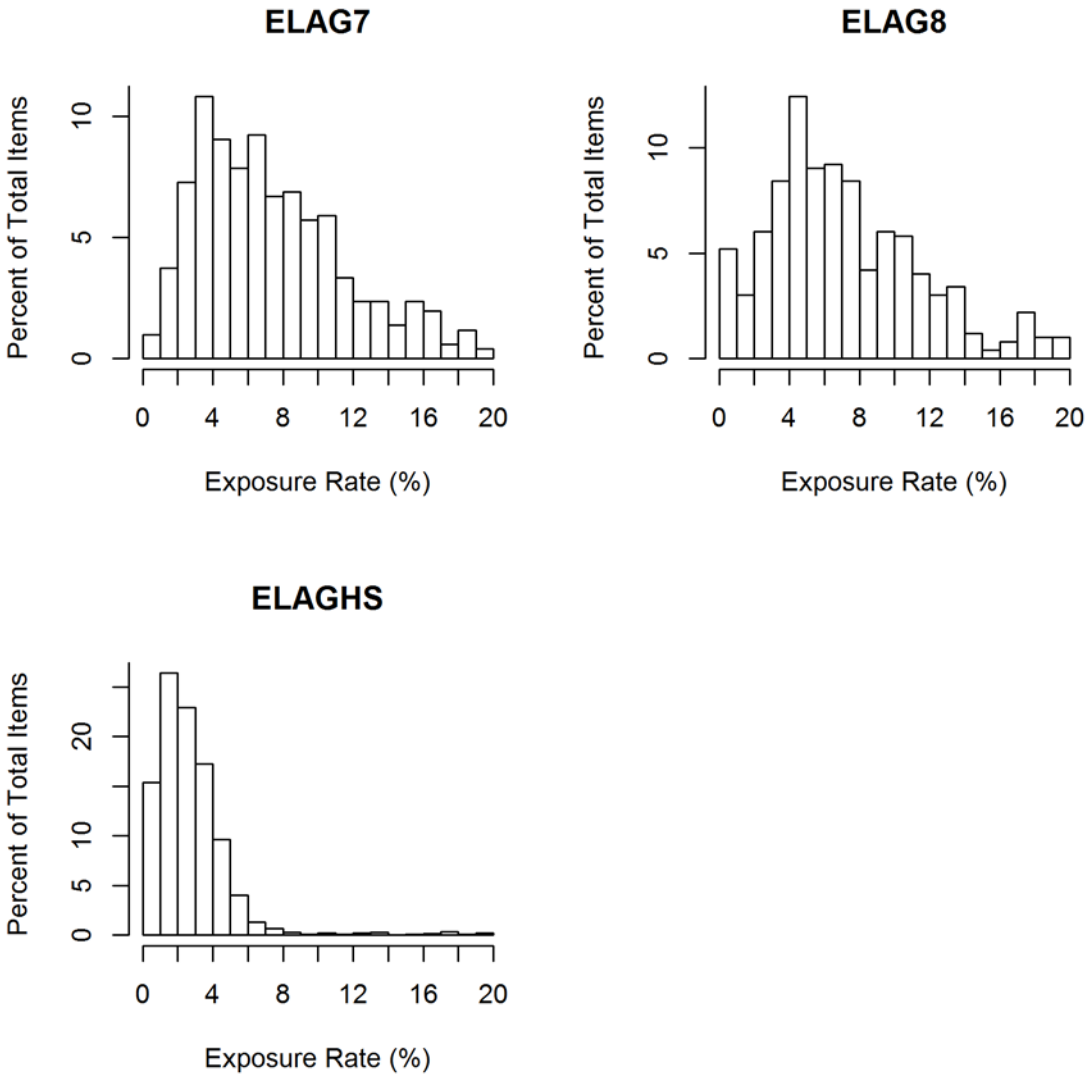


Figure 3. Mathematics: Histograms of Item Exposure Rates (Grades 3-6)

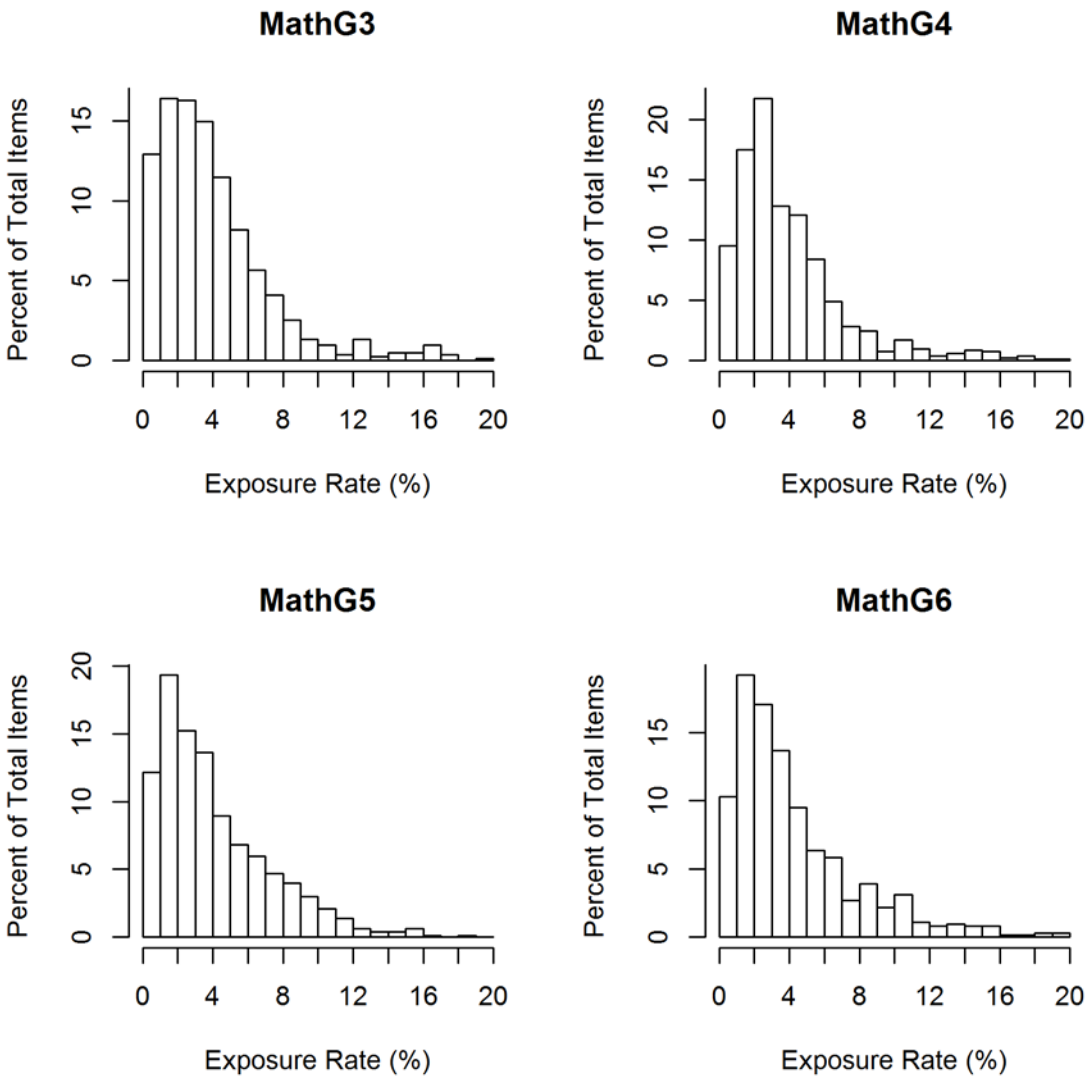
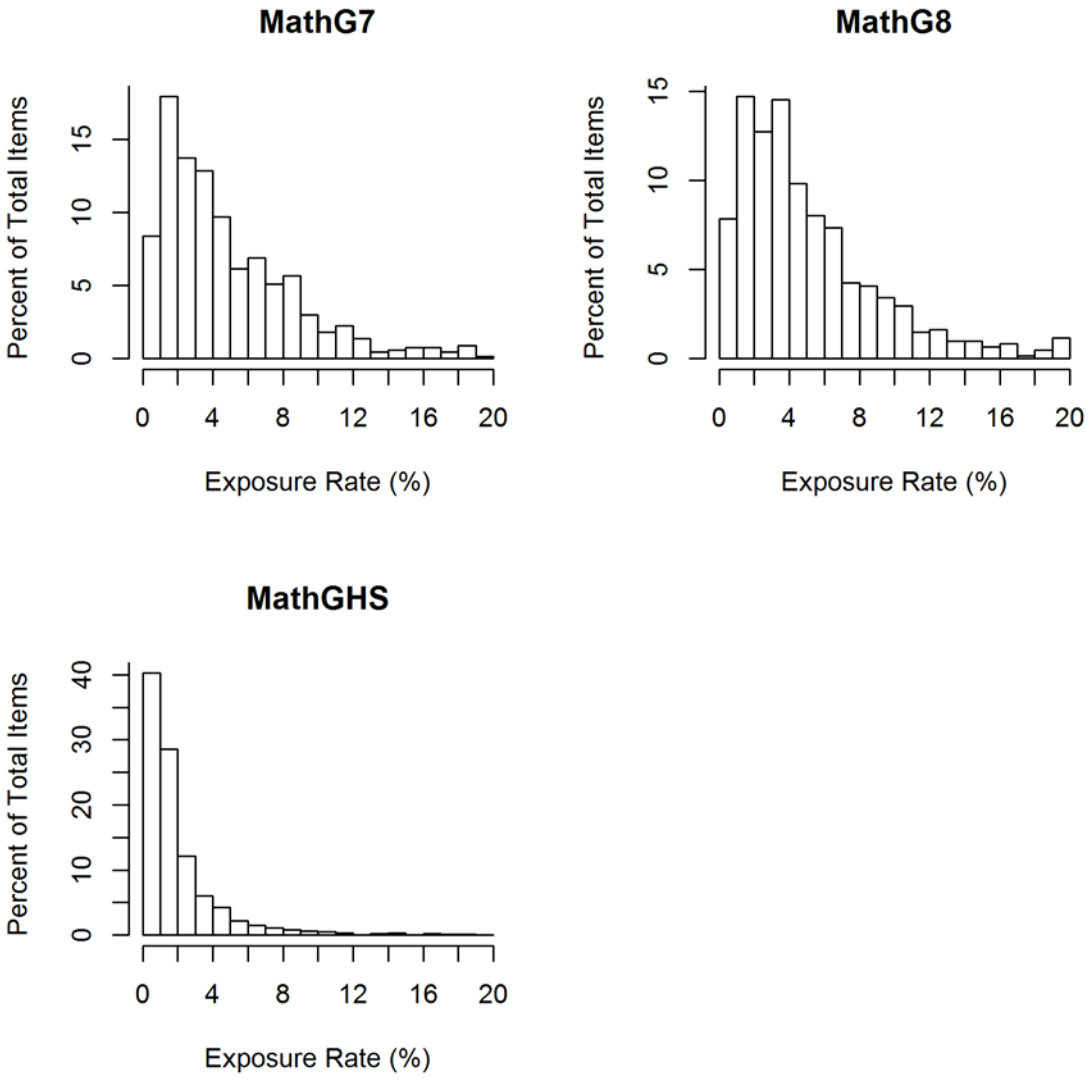


Figure 4. Mathematics: Histograms of Item Exposure Rates (Grades 7, 8, and HS)



Conclusion

The CRESST work presented in this report supports a number of conclusions regarding the Smarter Balanced Summative Assessments. First, for nearly all the simulees, the CAT engine designed by CRESST was able to satisfy the Smarter Balanced blueprint. Satisfying the blueprint with respect to the CAT portion was generally feasible. Satisfying the blueprint with respect to the PT portion, however, was at times more challenging due to a lack of correspondence between stimuli (and the component items) and the blueprint (see Table 7). It is difficult to directly compare the CAT results to those of AIR, due to potential differences in how the blueprint was operationalized and differences in the item pools at the times when these two studies were conducted.

Second, the student proficiency estimation in the simulated CAT/PT administrations (following the the Smarter Balanced Scoring Specifications) was quite good. The process resulted in estimates that showed no significant bias and had reasonable standard errors. Consistent with expectation, standard error averages did vary by decile, and students with the lowest true scores had the highest average standard errors. In some cases, this variation is largely a result of having few items in the pool that are informative in this range of simulee proficiency. Generally, the summary statistics from the AIR and CRESST reports are quite comparable, though we have already noted many differences between the simulations undertaken. While the inclusion of PT items did improve the precision of the score estimates, these gains are relatively small. This, again, is expected, since the PT component of the summative assessment increases the number of items contributing to the final score estimates by a relatively small proportion. Further insight into the incremental addition of PT items can be made by comparison of the above results with the CAT-only results presented in Appendix A.

Finally, exposure rates were low for most items. This is not to say that uniform exposure was achieved, but rather that there were very few items administered to a high percentage of the simulees. There was still, however, variability in the exposure rates, as shown in Figures 1-4. Given the complexity of the blueprint requirements, this sort of variability is likely unavoidable.

Given the potential implementation and methodological risks of item-by-item adaptive testing under a complex test blueprint, CRESST continues to encourage the Consortium to actively consider ways to improve its adaptive testing practices to better utilize the precious resources available in the current item pool, such as augmenting the overall design with multi-stage multi-form adaptive tests.

Appendix A.

Simulation Results for Student Proficiency Based on CAT Only

This Appendix is included to facilitate comparison with AIR's reported results. The tables that follow are based solely on the CAT component and are thus comparable to AIR's reported results on student proficiency. The following Tables (A1-A5) correspond to Tables (1, 8-11) in the main body of the report, and can be interpreted in the same way.

Table A1. Characteristics of Simulated and Estimated Proficiencies

Grade	Population Parameters		Obtainable Proficiency Range		Percentage of Winsorized Scores	
	Mean	SD	LOT	HOT	LOT	HOT
			English Language Arts/Literacy			
3	-1.24	1.06	-4.59	1.34	0.8	1.2
4	-0.75	1.11	-4.40	1.80	0.4	2.0
5	-0.31	1.10	-3.58	2.25	1.1	1.8
6	-0.06	1.11	-3.48	2.51	0.8	1.6
7	0.11	1.13	-2.91	2.75	1.6	1.4
8	0.38	1.13	-2.57	3.04	1.7	2.1
11	0.53	1.19	-2.44	3.34	1.4	1.1
			Mathematics			
3	-1.29	0.97	-4.11	1.33	0.7	0.6
4	-0.71	1.00	-3.92	1.82	0.5	1.1
5	-0.35	1.08	-3.73	2.33	1.0	1.4
6	-0.10	1.19	-3.53	2.95	0.9	1.0
7	0.01	1.33	-3.34	3.32	2.3	1.2
8	0.18	1.42	-3.15	3.63	2.9	1.4
11	0.51	1.52	-2.96	4.38	3.0	1.1

Table A2. Bias of the Estimated Proficiencies

Grade	Mean Bias	SE of Mean Bias	p-value for the *Z*-Test	MSE	95% CI Miss Rate	99% CI Miss Rate
English Language Arts/Literacy						
3	0.01	0.03	0.77	0.11	4.5	1.2
4	0.01	0.04	0.72	0.13	5.8	1.5
5	-0.01	0.03	0.81	0.11	4.1	0.8
6	0.02	0.04	0.65	0.14	4.5	0.4
7	0.01	0.04	0.88	0.15	4.7	0.8
8	0.03	0.04	0.44	0.15	4.3	0.4
11	0.00	0.04	0.99	0.17	5.3	1.2
Mathematics						
3	0.00	0.03	0.99	0.08	4.5	1.1
4	0.02	0.03	0.54	0.09	5.3	1.5
5	0.03	0.03	0.33	0.14	3.6	1.4
6	0.02	0.04	0.69	0.13	4.0	0.9
7	0.01	0.04	0.82	0.21	5.4	1.3
8	0.00	0.05	0.97	0.22	4.1	1.1
11	0.02	0.05	0.68	0.28	4.5	1.1

Table A3. Summaries of Difficulty of Item Pool and Estimated Student Proficiency

Grade	English Language Arts/Literacy				Mathematics			
	Items		Ability		Items		Ability	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3	-0.422	1.145	-1.308	1.135	-0.833	1.067	-1.338	1.016
4	0.126	1.302	-0.822	1.163	-0.065	1.031	-0.782	1.067
5	0.508	1.204	-0.362	1.167	0.675	1.016	-0.438	1.182
6	1.005	1.336	-0.132	1.166	1.064	1.225	-0.181	1.267
7	1.110	1.328	0.046	1.178	1.796	1.183	-0.073	1.395
8	1.301	1.338	0.292	1.197	2.319	1.462	0.096	1.482
11	1.689	1.355	0.464	1.256	2.703	1.571	0.402	1.614

Table A4. Average Standard Errors by Grade and by Deciles of True Proficiency Scores

Grade	Proficiency Score Distribution										
	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Overall
English Language Arts/Literacy											
3	0.51	0.34	0.31	0.29	0.28	0.27	0.27	0.27	0.28	0.32	0.32
4	0.47	0.35	0.33	0.31	0.31	0.32	0.31	0.31	0.32	0.34	0.34
5	0.46	0.34	0.31	0.30	0.30	0.30	0.31	0.31	0.32	0.35	0.33
6	0.52	0.42	0.37	0.35	0.34	0.33	0.33	0.32	0.34	0.35	0.37
7	0.54	0.43	0.40	0.38	0.36	0.35	0.35	0.34	0.34	0.35	0.39
8	0.52	0.42	0.39	0.37	0.35	0.35	0.35	0.34	0.34	0.36	0.38
11	0.57	0.46	0.41	0.38	0.37	0.36	0.35	0.35	0.35	0.37	0.40
Mathematics											
3	0.39	0.29	0.26	0.25	0.24	0.24	0.23	0.23	0.23	0.25	0.27
4	0.47	0.35	0.29	0.27	0.25	0.24	0.23	0.23	0.22	0.25	0.28
5	0.65	0.44	0.36	0.32	0.29	0.26	0.24	0.23	0.23	0.24	0.33
6	0.60	0.44	0.39	0.35	0.33	0.31	0.29	0.27	0.26	0.27	0.36
7	0.75	0.61	0.52	0.45	0.40	0.35	0.31	0.28	0.25	0.25	0.43
8	0.73	0.59	0.52	0.47	0.42	0.39	0.37	0.33	0.30	0.28	0.45
11	0.88	0.70	0.61	0.55	0.48	0.42	0.38	0.34	0.30	0.28	0.50

Table A5. Correlations between True and Estimated Proficiencies, and between Estimated Proficiency and Overall Test Form Difficulty

Grade	True Proficiency and Estimated Proficiency	Estimated Proficiency and Overall Test Difficulty
English Language Arts/Literacy		
3	0.95	0.71
4	0.95	0.73
5	0.96	0.74
6	0.95	0.69
7	0.94	0.66
8	0.95	0.68
11	0.95	0.61
Mathematics		
3	0.96	0.88
4	0.96	0.87
5	0.95	0.82
6	0.96	0.86
7	0.94	0.81
8	0.95	0.82
11	0.94	0.81

Appendix B.

SBAC Files and Documents Used for Analysis

Item pools necessary for simulations were obtained by merging together multiple data files received from SBAC between Feb 9, 2015 and Mar 28, 2015. The main list of items in the operational pool was obtained from *ItemReport_OperSumm_ELA.xlsx* and *ItemReport_OperSumm_MATH.xlsx*, provided via email to CRESST by SBAC on Feb 9, 2015. Additional merging with other files was required to obtain item parameters, complete claim/target information for Claim 2 ELA, stimulus IDs, extended pool indicators, DOK, and min/max number of items per passage/stimuli. Information about deactivated items or items with errors (e.g., in Spanish translation) were also merged with these data files, with the most recent item deactivation on Apr 3, 2015. Items without parameters (some PT items did not have associated item parameters) were not included in the pool of items available for simulated administration.

This analysis used blueprint files dated 02/09/15 that were downloaded from the SBAC website (Smarter Balanced Assessment Consortium, 2015a/b). We required the blueprint for each grade and subject combination to be in a format readable by our CAT/PT simulator. For this, we translated these documents into Excel tables specifying a series of requirements or rules about the min/max number of items or stimuli (e.g., passages) that need to be completed by each subject. Included in these tables were rules about the number of items and stimuli at each Claim, Target (or groups of Targets), and DOK level for CAT and PT separately. Additional requirements for ELA/L included Machine Scored vs. Short Text, Write Brief vs. Revise Brief, and number of Long Passages. Only rules specific enough to yield a minimum and (optional) maximum were included in our translated version of the blueprint. Since requirements for content domains in Mathematics are not clearly specified in the blueprint files mentioned above, these were not included in our translation of the blueprint requirements.

Appendix C.

Claim-level Blueprint Satisfaction

The tables (C1-C4) in Appendix C differ from the Blueprint Violation Tables 6 and 7 in that the C1-C4 Tables describe percentages of tests that meet blueprint requirements at the claim level (in terms of meeting min/max number of items and min/max number of passages for each claim, ignoring more specific target coverage or DOK requirements within a claim), while Tables 6 and 7 describe violations at both the claim level as well as more granular levels (Target and DOK requirements). The C1-C4 tables contain the minimum and maximum requirement of items in the blueprint, the page number in which the requirement appears, and percentages of test administration meeting these particular requirements. Detailed discussions regarding to these tables could be found on Page 9-10 in the main body of the report.

Table C1. Percentage of ELA/CAT Test Administration Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered

Grade	Claim	# Items required		Page	Item Requirement			% Match Passage Requirement
		Min	Max		Under	Match	Exceed	
3	Claim 1: Reading	14	16	1	0.0%	100.0%	0.0%	100.0%
3	Claim 2: Writing	10	10	1	0.0%	100.0%	0.0%	
3	Claim 3: Speaking/Listening	8	9	1	0.0%	100.0%	0.0%	100.0%
3	Claim 4: Research	6	6	1	0.0%	100.0%	0.0%	
4	Claim 1: Reading	14	16	1	0.0%	100.0%	0.0%	100.0%
4	Claim 2: Writing	10	10	1	0.0%	100.0%	0.0%	
4	Claim 3: Speaking/Listening	8	9	1	0.0%	100.0%	0.0%	100.0%
4	Claim 4: Research	6	6	1	0.0%	100.0%	0.0%	
5	Claim 1: Reading	14	16	1	0.0%	100.0%	0.0%	100.0%
5	Claim 2: Writing	10	10	1	0.0%	100.0%	0.0%	
5	Claim 3: Speaking/Listening	8	9	1	0.0%	100.0%	0.0%	100.0%
5	Claim 4: Research	6	6	1	0.0%	100.0%	0.0%	
6	Claim 1: Reading	13	17	2	0.0%	100.0%	0.0%	100.0%
6	Claim 2: Writing	10	10	2	0.0%	100.0%	0.0%	
6	Claim 3: Speaking/Listening	8	9	2	0.0%	100.0%	0.0%	100.0%
6	Claim 4: Research	6	6	2	0.0%	100.0%	0.0%	
7	Claim 1: Reading	13	17	2	0.0%	100.0%	0.0%	100.0%
7	Claim 2: Writing	10	10	2	0.0%	100.0%	0.0%	
7	Claim 3: Speaking/Listening	8	9	2	0.0%	100.0%	0.0%	100.0%
7	Claim 4: Research	6	6	2	0.0%	100.0%	0.0%	
8	Claim 1: Reading	13	17	2	0.0%	100.0%	0.0%	100.0%
8	Claim 2: Writing	10	10	2	0.0%	100.0%	0.0%	
8	Claim 3: Speaking/Listening	8	9	2	0.0%	100.0%	0.0%	100.0%
8	Claim 4: Research	6	6	2	0.0%	100.0%	0.0%	
11	Claim 1: Reading	15	16	3	0.0%	100.0%	0.0%	100.0%
11	Claim 2: Writing	10	10	3	0.0%	100.0%	0.0%	
11	Claim 3: Speaking/Listening	8	9	3	0.0%	100.0%	0.0%	100.0%
11	Claim 4: Research	6	6	3	0.0%	100.0%	0.0%	

Table C2. Percentage of ELA/PT Test Administration Meeting Blueprint Requirements for Each Claim

Grade	Claim	# Items required		Page	Item Requirement		
		Min	Max		Under	Match	Exceed
3	Claim 2: Writing	3	3	1	0.0%	100.0%	0.0%
3	Claim 4: Research	2	3	1	0.0%	100.0%	0.0%
4	Claim 2: Writing	3	3	1	0.0%	100.0%	0.0%
4	Claim 4: Research	2	3	1	4.8%	95.2%	0.0%
5	Claim 2: Writing	3	3	1	0.0%	100.0%	0.0%
5	Claim 4: Research	2	3	1	0.0%	100.0%	0.0%
6	Claim 2: Writing	3	3	2	0.0%	100.0%	0.0%
6	Claim 4: Research	2	3	2	0.0%	100.0%	0.0%
7	Claim 2: Writing	3	3	2	0.0%	100.0%	0.0%
7	Claim 4: Research	2	3	2	0.0%	100.0%	0.0%
8	Claim 2: Writing	3	3	2	0.0%	100.0%	0.0%
8	Claim 4: Research	2	3	2	0.0%	100.0%	0.0%
11	Claim 2: Writing	3	3	3	0.0%	100.0%	0.0%
11	Claim 4: Research	2	3	3	0.0%	100.0%	0.0%

Table C3. Percentage of Math/CAT Test Administration Meeting Blueprint Requirements for Each Claim

Grade	Claim	# Items required		Page	Item Requirement		
		Min	Max		Under	Match	Exceed
3	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
3	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
3	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
4	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
4	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
4	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
5	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
5	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
5	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
6	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
6	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
6	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
7	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
7	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
7	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
8	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
8	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
8	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
11	Claim 1: Concepts and Procedures	19	22	3	0.0%	100.0%	0.0%
11	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	3	0.0%	100.0%	0.0%
11	Claim 3: Communicating Reasoning	8	8	3	0.0%	100.0%	0.0%

Table C4. Percentage of Math/PT Test Administration Meeting Blueprint Requirements for Each Claim

Grade	Claim	# Items required		Page	Item Requirement		
		Min	Max		Under	Match	Exceed
3	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.0%	82.5%	17.5%
3	Claim 3: Communicating Reasoning	0	2	1	0.0%	94.5%	5.5%
4	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.0%	100.0%	0.0%
4	Claim 3: Communicating Reasoning	0	2	1	0.0%	100.0%	0.0%
5	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.0%	100.0%	0.0%
5	Claim 3: Communicating Reasoning	0	2	1	0.0%	100.0%	0.0%
6	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.0%	100.0%	0.0%
6	Claim 3: Communicating Reasoning	0	2	2	0.0%	95.7%	4.3%
7	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.0%	100.0%	0.0%
7	Claim 3: Communicating Reasoning	0	2	2	0.0%	100.0%	0.0%
8	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.0%	100.0%	0.0%
8	Claim 3: Communicating Reasoning	0	2	2	0.0%	100.0%	0.0%
11	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	3	0.0%	100.0%	0.0%
11	Claim 3: Communicating Reasoning	0	2	3	0.0%	94.1%	5.9%

Appendix D.

Simulation Results for Braille

This Appendix reports results from the CAT simulations for the Braille tests. Tables D1-D6 correspond to Tables 2-7 in the main body of the report. Tables D7-D13 correspond to Tables 9-15 in the main body. Similarly, Figures D1-D4 correspond to Figures 1-4 in the main body. Finally, Tables D13-D16, which provide more details on Blueprint violations, correspond to Tables C1-C4 in Appendix C. All tables and figures in this appendix may be interpreted in the same way as the analogous tables and figures in the main body and Appendix C.

The results for the Braille tests are largely similar to the corresponding results for the general tests. The discrepancies are primarily due to the differences in the items pools, as well as sampling variability. As there are fewer items in the pool for the Braille test, there are fewer items available to be selected by the CAT engine. Some consequences of this include: 1) slightly smaller correlations between estimated proficiency and overall test difficulty (see Tables 14 and D12); and 2) fewer items that go unused by the CAT engine (see Tables 15 and D13). One last consequence is that one particular blueprint requirement cannot be met due to a shortage of requisite items (see Table D5, Grade 5).

Table D1. Number of Operational Items in ELA/L Adaptive Test Item Pool

Grade	Number of Items					Number of Passages		
	Total	Claim 1	Claim 2	Claim 3	Claim 4	Claim 1 Literary	Claim 1 Information	Claim 3 Listening
3	299	117	77	69	36	9	10	28
4	297	98	77	76	46	7	8	29
5	317	119	78	71	49	10	8	28
6	288	105	71	75	37	4	12	30
7	272	102	69	76	25	3	14	29
8	258	103	67	61	27	3	12	22
11	524	210	113	135	66	10	23	49

Note. Item counts current as of 2015-04-03.

Table D2. Number of Operational Items in ELA/L Performance Task Item Pool

Grade	Number of Items					Number of Stimuli
	Total	Claim 1	Claim 2	Claim 3	Claim 4	Across Claims
3	23	0	10	0	13	14
4	23	0	10	0	13	20
5	28	0	12	0	16	21
6	22	0	10	0	12	14
7	33	0	16	0	17	19
8	35	0	16	0	19	19
11	35	0	16	0	19	26

Note. Item counts current as of 2015-04-03.

Table D3. Number of Operational Items in Mathematics Adaptive Test Item Pool

Grade	Calculator	Total	Claim 1	Claim 2	Claim 3	Claim 4
3	No	334	207	42	44	41
4	No	265	159	38	32	36
5	No	324	181	39	52	52
6	Yes	192	91	39	39	23
	No	164	162	0	2	0
7	Yes	236	136	35	41	24
	No	87	87	0	0	0
8	Yes	198	124	17	44	13
	No	74	74	0	0	0
11	Yes	321	160	34	83	44
	No	46	33	0	13	0

Note. Item counts current as of 2015-04-03.

Table D4. Number of Operational Items in Mathematics Performance Task Item Pool

Grade	Number of Items					Number of Stimuli
	Total	Claim 1	Claim 2	Claim 3	Claim 4	(Across Claims)
3	47	0	19	15	13	8
4	34	0	15	10	9	7
5	41	0	14	14	13	7
6	34	0	12	12	10	6
7	27	0	11	7	9	5
8	25	0	8	8	9	5
11	29	0	9	10	10	5

Note. Item counts current as of 2015-04-03.

Table D5. CAT Tests with Blueprint Violations (CAT Portion Only)

			Blueprint Requirement			Number of Tests		
Grade	Subject	Blueprint Specification	Pg. #	Min	Max	Total	Below Min.	Above Max.
4	English	Claim 1, DOK=2	4	7	--	34	34	0
4	English	Claim 1 (Literary), Target 4: Reasoning and Evaluation	4	1	2	1	1	0
4	English	Claim 1 (Informational), Target 9: Central Ideas	4	1	2	1	1	0
4	English	Claim 1 (Informational), Target 11: Reasoning and Evaluation	4	1	2	21	21	0
5	English	Claim 1, DOK≥3	4	2	--	1	1	0
5	English	Claim 1 (Literary), Target 2: Central Ideas	4	1	2	5	5	0
5	English	Claim 1 (Literary), Target 4: Reasoning and Evaluation	4	1	2	6	6	0
5	English	Claim 1 (Informational), Target 11: Reasoning and Evaluation	4	1	2	29	29	0
5	English	Claim 2, DOK=2	4	4	--	1000	1000	0
6	English	Claim 1 (Literary), Target 2: Central Ideas	7	1	1	3	3	0
6	English	Claim 2, DOK≥2	7	5	--	1	1	0
7	English	Claim 1 (Literary), Target 2: Central Ideas	7	1	1	1	1	0
7	English	Claim 1 (Informational), Targets 8, 10, 12, 13, 14	7	7	8	1	1	0
7	English	Claim 1 (Informational), Machine Scored Items	7	9	10	1	1	0

Table D6. PT Tests with Blueprint Violations (PT Portion Only)

Grade	Subject	Blueprint Specification	Blueprint Requirement			Number of Tests		
			Pg. #	Min	Max	Total	Below Min.	Above Max.
3	Math	Claim 2/Claim 4	1	2	4	274	0	274
3	Math	Claim 2 (Problem Solving)	5	1	2	387	0	387
3	Math	Claim 4 (Modeling and Data Analysis)	5	1	3	124	124	0
3	Math	Claim 3 (Communicating Reason)	5	0	2	124	0	124
4	Math	Claim 2 (Problem Solving)	7	1	2	139	0	139
4	Math	Claim 4 (Modeling and Data Analysis)	7	1	3	136	136	0
5	Math	Claim 2 (Problem Solving)	9	1	2	279	0	279
7	Math	Claim 2 (Problem Solving)	13	1	2	190	0	190

Table D7. Bias of the Estimated Proficiencies: English Language Arts/Literacy

Grade	Mean Bias	SE of Mean Bias	p-value for the *z*-Test	MSE	95% CI Miss Rate	99% CI Miss Rate
			Overall			
3	0.02	0.03	0.64	0.09	5.2	0.4
4	0.03	0.04	0.40	0.11	5.0	0.6
5	0.00	0.03	0.92	0.09	4.5	0.8
6	0.03	0.04	0.47	0.12	4.1	0.9
7	0.01	0.04	0.79	0.13	5.0	0.8
8	0.01	0.04	0.76	0.12	4.2	0.4
11	0.01	0.04	0.74	0.14	6.0	0.7
			Claim 1: Reading			
3	0.07	0.03	0.03	0.33	6.2	2.0
4	0.10	0.04	0.00	0.46	7.6	2.5
5	-0.01	0.03	0.76	0.29	5.6	1.5
6	0.09	0.04	0.01	0.48	5.5	2.5
7	0.01	0.04	0.77	0.42	5.4	1.3
8	0.04	0.04	0.31	0.40	7.0	2.2
11	0.07	0.04	0.08	0.45	5.6	1.3
			Claim 2: Writing			
3	0.04	0.03	0.29	0.31	4.9	1.4
4	0.03	0.04	0.44	0.31	5.5	1.3
5	0.03	0.03	0.41	0.32	6.1	1.2
6	0.05	0.04	0.16	0.38	6.4	2.3
7	0.03	0.04	0.45	0.36	5.5	1.7
8	0.01	0.04	0.69	0.34	4.8	1.2
11	0.04	0.04	0.33	0.48	5.1	1.5
			Claim 3: Speaking/Listening			
3	0.18	0.03	0.00	0.91	11.3	6.2
4	0.14	0.04	0.00	0.80	7.5	4.9
5	0.09	0.04	0.02	0.83	8.8	4.9
6	0.12	0.04	0.00	0.98	7.8	4.8
7	0.06	0.04	0.08	0.75	6.0	3.3
8	0.09	0.04	0.02	0.89	7.2	4.3
11	0.04	0.04	0.31	0.88	7.3	4.2
			Claim 4: Research			
3	0.16	0.03	0.00	0.85	11.9	7.0
4	0.16	0.04	0.00	0.84	10.7	6.3
5	0.08	0.04	0.03	0.57	8.5	4.5
6	0.28	0.04	0.00	1.16	14.8	9.8
7	0.19	0.04	0.00	1.00	13.2	8.7
8	0.16	0.04	0.00	0.89	11.9	6.5
11	0.13	0.04	0.00	0.88	12.1	6.7

Table D8. Bias of the Estimated Proficiencies: Mathematics

Grade	Mean Bias	SE of Mean Bias	p-value for the *z*-Test	MSE	95% CI Miss Rate	99% CI Miss Rate
Overall						
3	0.00	0.03	0.94	0.07	4.7	1.2
4	0.01	0.03	0.69	0.08	4.4	0.7
5	0.03	0.03	0.42	0.13	4.8	1.3
6	0.01	0.04	0.84	0.11	3.6	0.7
7	0.01	0.04	0.84	0.16	3.6	0.9
8	0.01	0.05	0.75	0.22	5.3	1.4
11	0.05	0.05	0.30	0.27	4.9	0.6
Claim 1: Concepts and Procedures						
3	-0.01	0.03	0.70	0.12	5.1	0.8
4	0.02	0.03	0.55	0.16	5.6	1.1
5	0.06	0.03	0.07	0.27	5.2	1.7
6	0.01	0.04	0.77	0.21	4.5	0.8
7	0.04	0.04	0.38	0.31	5.3	1.2
8	0.04	0.05	0.33	0.35	5.5	1.3
11	0.08	0.05	0.08	0.54	6.0	1.2
Claim 2/4: Problem Solving/Modeling and Data Analysis						
3	0.15	0.03	0.00	0.58	13.1	7.9
4	0.16	0.03	0.00	0.69	12.2	7.1
5	0.32	0.04	0.00	1.16	18.3	10.9
6	0.31	0.04	0.00	1.20	18.5	11.3
7	0.36	0.04	0.00	1.48	18.4	10.3
8	0.38	0.05	0.00	1.64	19.3	10.8
11	0.45	0.05	0.00	1.75	18.8	10.0
Claim 3: Communicating Reasoning						
3	0.17	0.03	0.00	0.58	12.1	7.9
4	0.24	0.03	0.00	0.85	14.2	9.4
5	0.15	0.03	0.00	0.62	9.4	4.6
6	0.16	0.04	0.00	0.73	9.4	4.8
7	0.24	0.04	0.00	1.14	10.4	4.9
8	0.14	0.05	0.00	1.03	9.4	4.3
11	0.16	0.05	0.00	1.06	8.2	3.5

Table D9. Overall Score and Claim Score Precision/Reliability: English Language Arts/Literacy

Grade	Overall ELA/L					Claim 1					Claim 2					Claim 3					Claim 4				
	ave # items	SD($\hat{\theta}$)	mean SEM	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$
3	45.5	1.1	.31	.31	.92	16.0	1.2	.51	.57	.79	12.0	1.3	.53	.56	.80	9.0	1.5	.83	.95	.59	8.6	1.5	.70	.92	.61
4	45.6	1.2	.32	.33	.92	16.0	1.4	.62	.68	.75	12.0	1.3	.53	.56	.81	9.0	1.5	.82	.89	.64	8.6	1.5	.75	.91	.63
5	45.7	1.1	.30	.30	.93	16.0	1.2	.52	.54	.81	12.0	1.3	.54	.56	.81	9.0	1.5	.85	.91	.61	8.7	1.4	.65	.75	.70
6	43.4	1.2	.34	.34	.92	14.0	1.4	.64	.69	.74	12.0	1.3	.56	.61	.79	9.0	1.5	.92	.99	.56	8.4	1.6	.79	1.08	.55
7	43.1	1.2	.35	.35	.91	14.0	1.3	.63	.64	.76	12.0	1.3	.59	.60	.79	9.0	1.4	.87	.87	.63	8.1	1.5	.86	1.00	.55
8	43.4	1.2	.35	.35	.92	14.0	1.3	.60	.63	.77	12.0	1.3	.59	.59	.80	9.0	1.5	.93	.94	.58	8.4	1.5	.84	.94	.61
11	45.4	1.2	.38	.38	.91	16.0	1.4	.65	.67	.76	12.0	1.4	.68	.69	.75	9.0	1.5	.96	.94	.60	8.4	1.5	.86	.94	.62

Table D10. Overall Score and Claim Score Precision/Reliability: Mathematics

Grade	Overall Math					Claim 1					Claim 2/4					Claim 3				
	ave # items	SD($\hat{\theta}$)	mean SEM	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$
3	39.9	1.0	.26	.26	.93	20.0	1.0	.34	.35	.89	10.0	1.3	.53	.76	.66	9.9	1.3	.59	.76	.66
4	38.9	1.1	.30	.29	.93	20.0	1.1	.39	.39	.87	9.4	1.4	.63	.83	.65	9.4	1.4	.66	.92	.59
5	39.9	1.2	.34	.37	.90	20.0	1.3	.48	.52	.83	9.9	1.7	.62	1.08	.58	10.0	1.4	.64	.79	.68
6	38.7	1.3	.35	.33	.93	19.0	1.3	.47	.46	.88	9.7	1.7	.70	1.10	.60	10.0	1.5	.76	.86	.68
7	39.4	1.4	.43	.40	.92	20.0	1.5	.55	.55	.86	10.0	1.9	.80	1.22	.58	9.4	1.7	.95	1.07	.58
8	39.0	1.5	.47	.47	.90	20.0	1.5	.59	.59	.85	9.4	2.0	.88	1.28	.58	9.6	1.7	.95	1.02	.64
11	41.8	1.6	.55	.51	.90	22.0	1.7	.74	.74	.81	9.8	2.1	.90	1.32	.60	10.0	1.8	1.02	1.03	.69

Table D11. Average Standard Errors by Grade and by Deciles of True Proficiency Scores

Grade	Proficiency Score Distribution										Overall
	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
English Language Arts/Literacy											
3	0.49	0.33	0.29	0.27	0.26	0.26	0.25	0.25	0.25	0.28	0.30
4	0.49	0.34	0.31	0.29	0.28	0.28	0.27	0.27	0.28	0.30	0.31
5	0.43	0.31	0.29	0.27	0.27	0.27	0.27	0.27	0.28	0.31	0.30
6	0.52	0.38	0.32	0.30	0.29	0.28	0.28	0.28	0.28	0.30	0.33
7	0.50	0.39	0.35	0.33	0.31	0.30	0.30	0.29	0.30	0.33	0.34
8	0.49	0.38	0.34	0.33	0.32	0.31	0.31	0.31	0.31	0.33	0.34
11	0.55	0.44	0.39	0.36	0.34	0.32	0.32	0.31	0.31	0.33	0.37
Mathematics											
3	0.38	0.28	0.25	0.24	0.23	0.22	0.22	0.21	0.22	0.24	0.25
4	0.49	0.34	0.30	0.28	0.26	0.24	0.23	0.22	0.21	0.24	0.29
5	0.59	0.43	0.34	0.31	0.28	0.26	0.24	0.23	0.22	0.22	0.32
6	0.57	0.42	0.36	0.32	0.30	0.28	0.26	0.25	0.24	0.27	0.33
7	0.71	0.55	0.46	0.41	0.37	0.33	0.30	0.26	0.24	0.24	0.39
8	0.74	0.58	0.52	0.47	0.42	0.38	0.34	0.30	0.27	0.26	0.44
11	0.90	0.72	0.61	0.54	0.47	0.40	0.36	0.32	0.30	0.29	0.50

Table D12. Correlations between True and Estimated Proficiency, and between Estimated Proficiency and Overall Test Difficulty

Grade	True Proficiency and Estimated Proficiency	Estimated Proficiency and Overall Test Difficulty
English Language Arts/Literacy		
3	0.96	0.70
4	0.96	0.69
5	0.96	0.74
6	0.96	0.65
7	0.95	0.63
8	0.96	0.65
11	0.95	0.59
Mathematics		
3	0.97	0.86
4	0.96	0.82
5	0.95	0.80
6	0.96	0.83
7	0.96	0.76
8	0.95	0.80
11	0.95	0.70

Note. Overall test difficulty is the average of item location parameters for all items in the test instance

Table D13. Percent of Items by Exposure Rate

Grade	Total Items	Exposure Rate					
		Unused	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
English Language Arts/Literacy							
3	299	0.33	78.26	20.40	0.67	0.33	0
4	297	0	79.46	17.85	2.36	0.34	0
5	317	0	85.17	12.93	1.26	0	0.63
6	288	5.21	75.69	16.67	1.39	1.04	0
7	272	4.78	73.53	16.91	3.31	1.47	0
8	258	4.26	70.93	20.93	2.71	1.16	0
11	524	1.91	93.89	4.20	0.00	0	0
Mathematics							
3	334	0	87.43	11.68	0.90	0	0
4	265	0.38	83.02	12.83	3.02	0	0.75
5	324	0	87.65	11.42	0.93	0	0
6	356	0	91.01	7.87	1.12	0	0
7	323	0	88.24	9.91	1.86	0	0

Figure D1. ELA/L: Histograms of Item Exposure Rates (Grades 3-6)

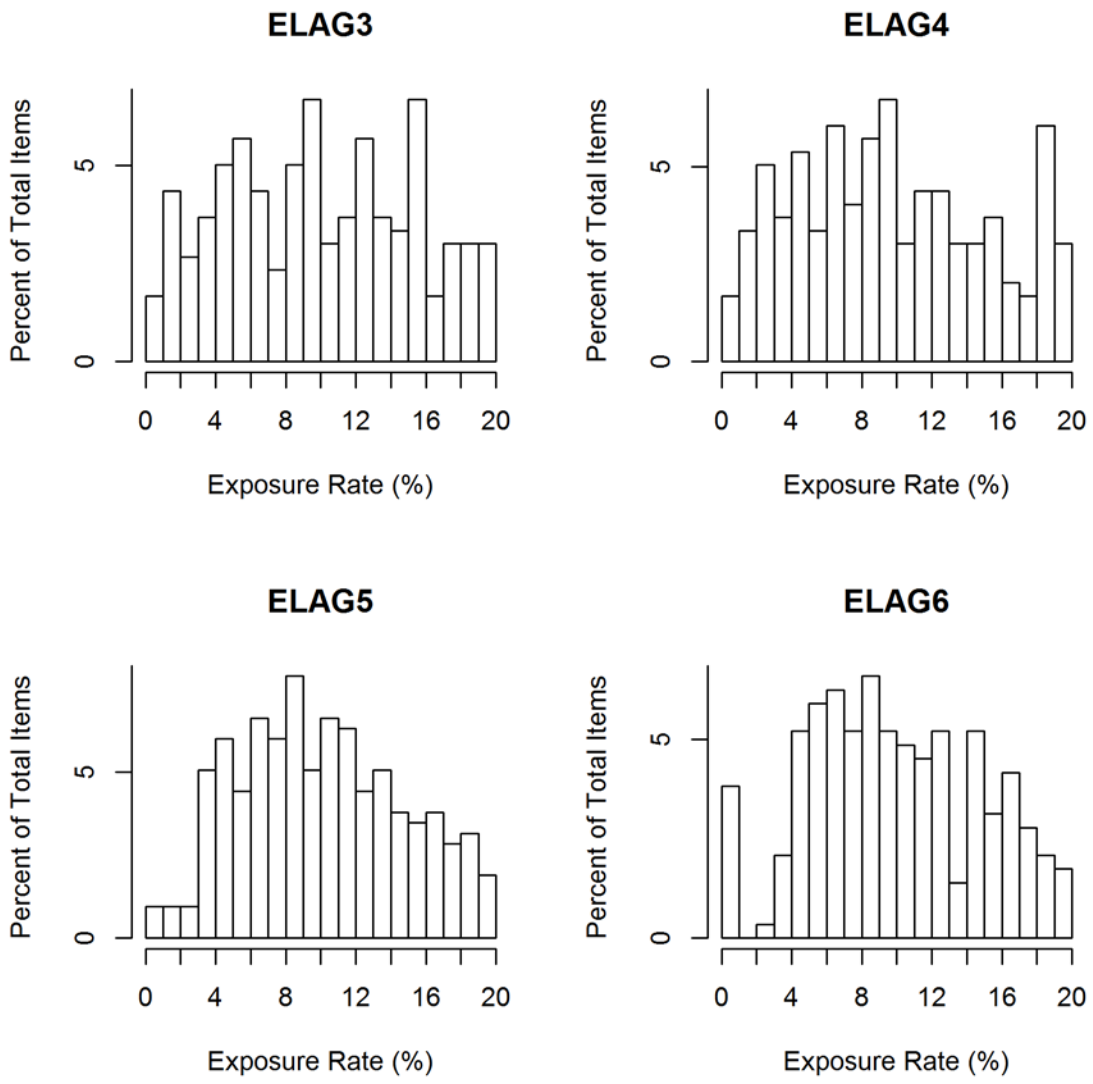


Figure D2. ELA/L: Histograms of Item Exposure Rates (Grades 7, 8, and HS)

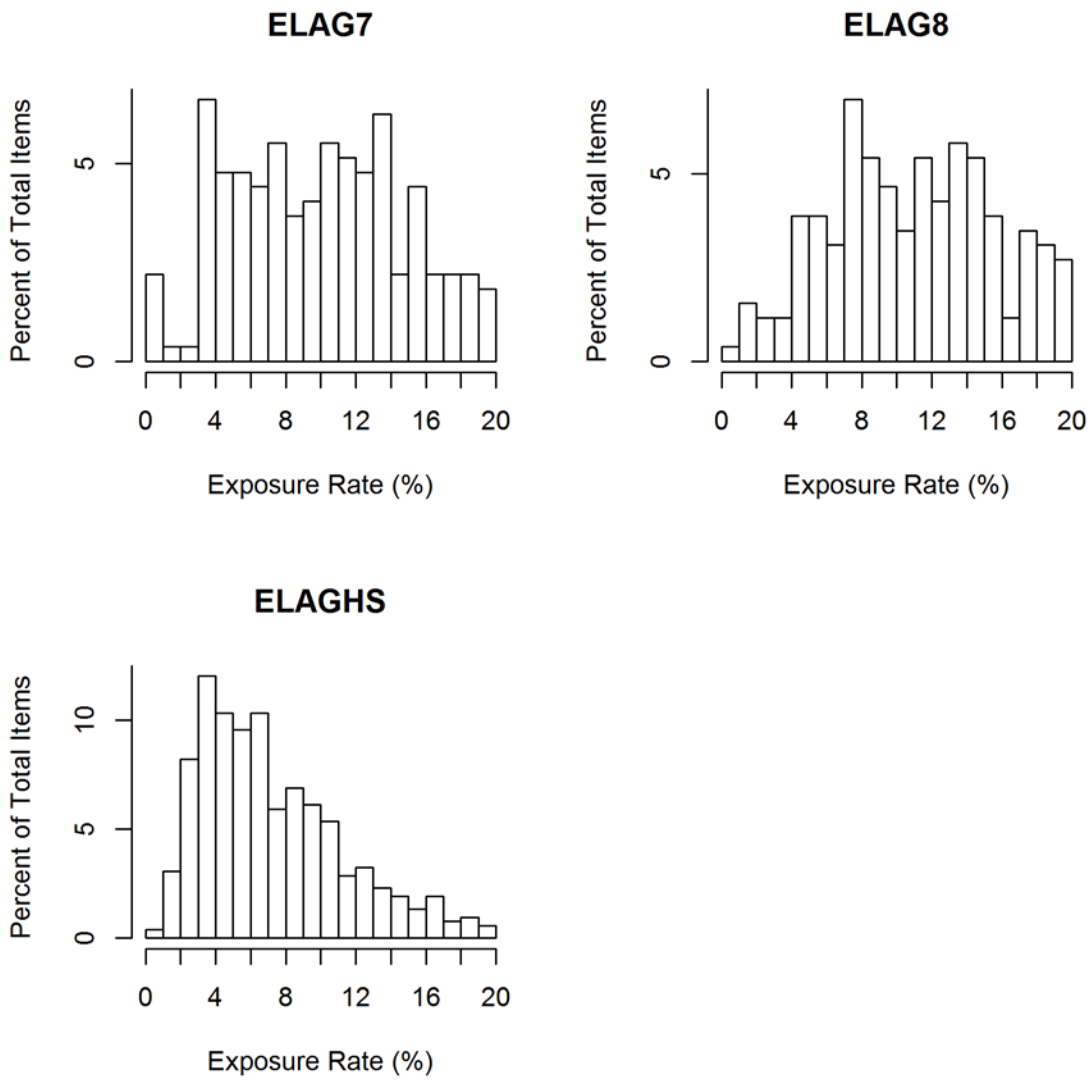


Figure D3. Mathematics: Histograms of Item Exposure Rates (Grades 3-6)

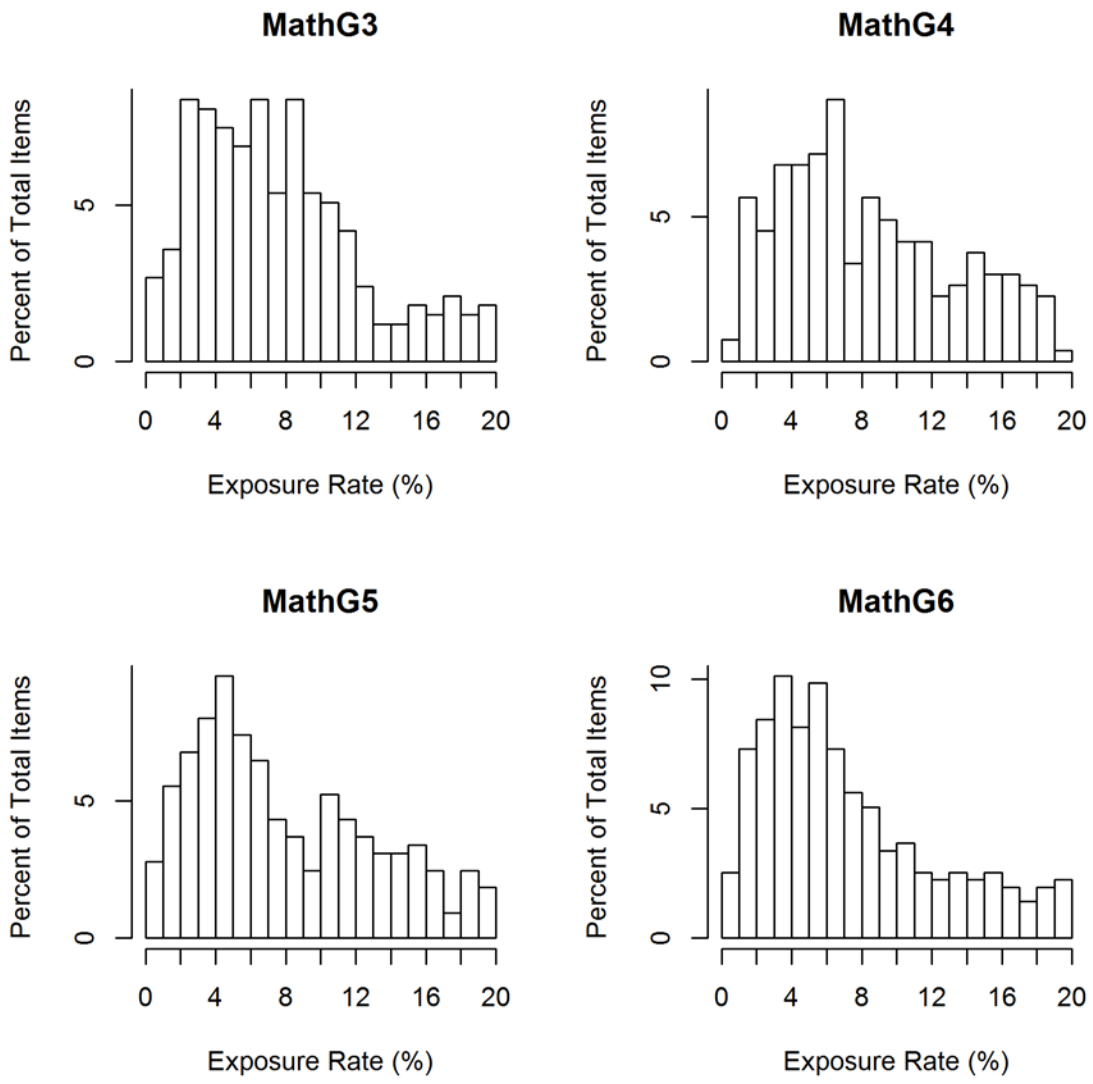


Figure D4. Mathematics: Histograms of Item Exposure Rates (Grades 7, 8, and HS)

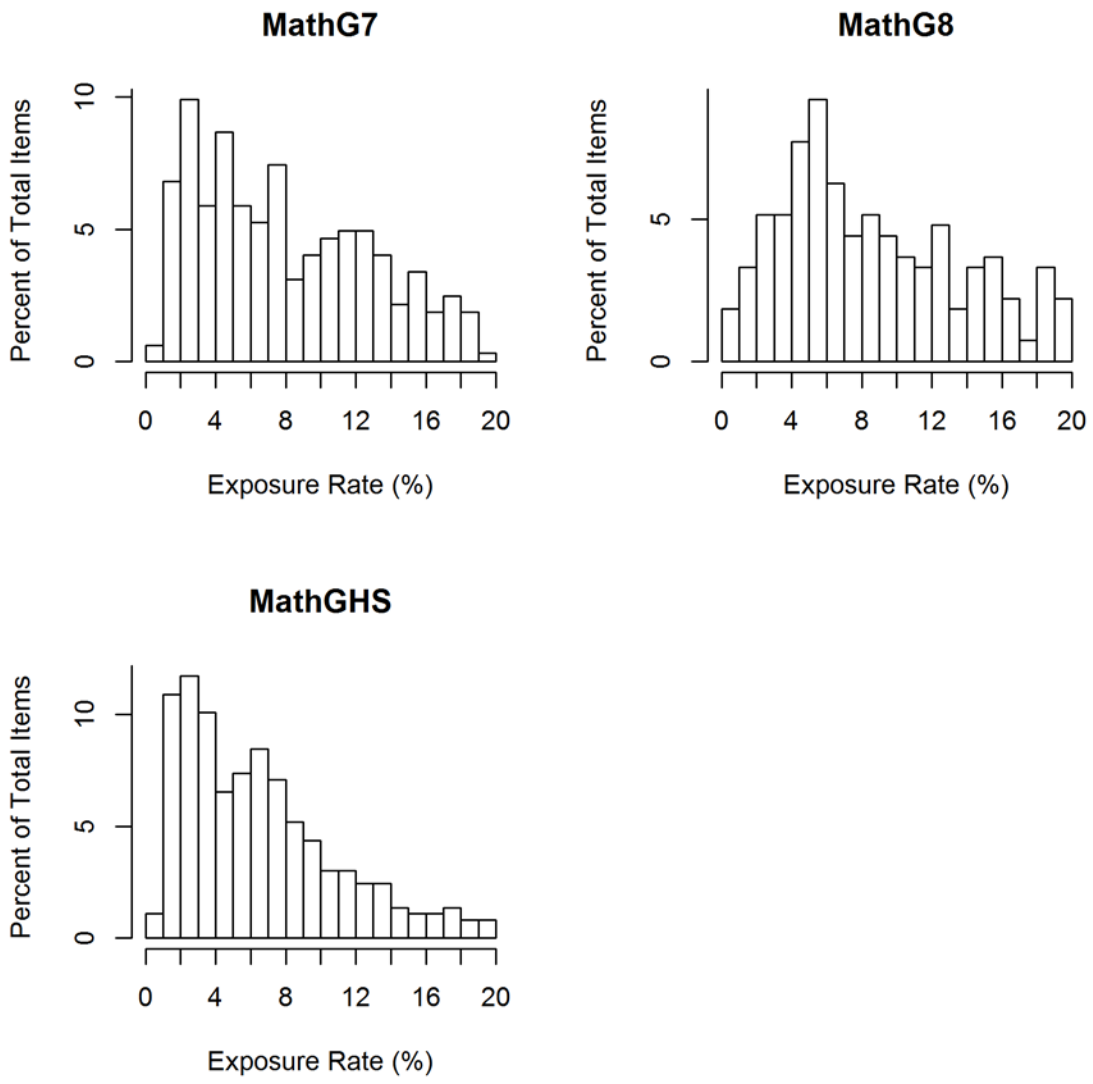


Table D13. Percentage of ELA/CAT Braille Test Administration Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered

Grade	Claim	# Items required		Page	Item Requirement			% Match Passage Requirement
		Min	Max		Under	Match	Exceed	
3	Claim 1: Reading	14	16	1	0.0%	100.0%	0.0%	100.0%
3	Claim 2: Writing	10	10	1	0.0%	100.0%	0.0%	
3	Claim 3: Speaking/Listening	8	9	1	0.0%	100.0%	0.0%	100.0%
3	Claim 4: Research	6	6	1	0.0%	100.0%	0.0%	
4	Claim 1: Reading	14	16	1	0.0%	100.0%	0.0%	100.0%
4	Claim 2: Writing	10	10	1	0.0%	100.0%	0.0%	
4	Claim 3: Speaking/Listening	8	9	1	0.0%	100.0%	0.0%	100.0%
4	Claim 4: Research	6	6	1	0.0%	100.0%	0.0%	
5	Claim 1: Reading	14	16	1	0.0%	100.0%	0.0%	100.0%
5	Claim 2: Writing	10	10	1	0.0%	100.0%	0.0%	
5	Claim 3: Speaking/Listening	8	9	1	0.0%	100.0%	0.0%	100.0%
5	Claim 4: Research	6	6	1	0.0%	100.0%	0.0%	
6	Claim 1: Reading	13	17	2	0.0%	100.0%	0.0%	100.0%
6	Claim 2: Writing	10	10	2	0.0%	100.0%	0.0%	
6	Claim 3: Speaking/Listening	8	9	2	0.0%	100.0%	0.0%	100.0%
6	Claim 4: Research	6	6	2	0.0%	100.0%	0.0%	
7	Claim 1: Reading	13	17	2	0.0%	100.0%	0.0%	100.0%
7	Claim 2: Writing	10	10	2	0.0%	100.0%	0.0%	
7	Claim 3: Speaking/Listening	8	9	2	0.0%	100.0%	0.0%	100.0%
7	Claim 4: Research	6	6	2	0.0%	100.0%	0.0%	
8	Claim 1: Reading	13	17	2	0.0%	100.0%	0.0%	100.0%
8	Claim 2: Writing	10	10	2	0.0%	100.0%	0.0%	
8	Claim 3: Speaking/Listening	8	9	2	0.0%	100.0%	0.0%	100.0%
8	Claim 4: Research	6	6	2	0.0%	100.0%	0.0%	
11	Claim 1: Reading	15	16	3	0.0%	100.0%	0.0%	100.0%
11	Claim 2: Writing	10	10	3	0.0%	100.0%	0.0%	
11	Claim 3: Speaking/Listening	8	9	3	0.0%	100.0%	0.0%	100.0%
11	Claim 4: Research	6	6	3	0.0%	100.0%	0.0%	

Table D14. Percentage of ELA/PT Braille Test Administrations Meeting Blueprint Requirements for Each Claim

Grade	Claim	# Items required			Item Requirement		
		Min	Max		Under	Match	Exceed
3	Claim 2: Writing	3	3	1	0.0%	100.0%	0.0%
3	Claim 4: Research	2	3	1	0.0%	100.0%	0.0%
4	Claim 2: Writing	3	3	1	0.0%	100.0%	0.0%
4	Claim 4: Research	2	3	1	0.0%	100.0%	0.0%
5	Claim 2: Writing	3	3	1	0.0%	100.0%	0.0%
5	Claim 4: Research	2	3	1	0.0%	100.0%	0.0%
6	Claim 2: Writing	3	3	2	0.0%	100.0%	0.0%
6	Claim 4: Research	2	3	2	0.0%	100.0%	0.0%
7	Claim 2: Writing	3	3	2	0.0%	100.0%	0.0%
7	Claim 4: Research	2	3	2	0.0%	100.0%	0.0%
8	Claim 2: Writing	3	3	2	0.0%	100.0%	0.0%
8	Claim 4: Research	2	3	2	0.0%	100.0%	0.0%
11	Claim 2: Writing	3	3	3	0.0%	100.0%	0.0%
11	Claim 4: Research	2	3	3	0.0%	100.0%	0.0%

Table D15. Percentage of Math/CAT Braille Test Administrations Meeting Blueprint Requirements for Each Claim

Grade	Claim	# Items required		Page	Item Requirement		
		Min	Max		Under	Match	Exceed
3	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
3	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
3	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
4	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
4	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
4	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
5	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
5	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
5	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
6	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
6	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
6	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
7	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
7	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
7	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
8	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
8	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
8	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
11	Claim 1: Concepts and Procedures	19	22	3	0.0%	100.0%	0.0%
11	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	3	0.0%	100.0%	0.0%
11	Claim 3: Communicating Reasoning	8	8	3	0.0%	100.0%	0.0%

Table D16. Percentage of Math/PT Braille Test Administration Meeting Blueprint Requirements for Each Claim

Grade	Claim	# Items required		Page	Item Requirement		
		Min	Max		Under	Match	Exceed
3	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.0%	72.6%	27.4%
3	Claim 3: Communicating Reasoning	0	2	1	0.0%	87.6%	12.4%
4	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.0%	100.0%	0.0%
4	Claim 3: Communicating Reasoning	0	2	1	0.0%	100.0%	0.0%
5	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.0%	100.0%	0.0%
5	Claim 3: Communicating Reasoning	0	2	1	0.0%	100.0%	0.0%
6	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.0%	100.0%	0.0%
6	Claim 3: Communicating Reasoning	0	2	2	0.0%	100.0%	0.0%
7	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.0%	100.0%	0.0%
7	Claim 3: Communicating Reasoning	0	2	2	0.0%	100.0%	0.0%
8	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.0%	100.0%	0.0%
8	Claim 3: Communicating Reasoning	0	2	2	0.0%	100.0%	0.0%
11	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	3	0.0%	100.0%	0.0%
11	Claim 3: Communicating Reasoning	0	2	3	0.0%	100.0%	0.0%

Appendix E.

Simulation Results for Spanish

This Appendix reports results from the CAT simulations for the Spanish translations of the ELA/L and Math tests. Tables E1-E5 correspond to Tables 4, 5, 7, 8, and 10 in the main body of the report. Tables E6-E9 correspond to Tables 12-15 in the main body. Similarly, Figures E1 and E2 correspond to Figures 3 and 4 in the main body. Finally, Tables E12-E13, which provide more details on Blueprint violations, correspond to Tables C3 and C4 in Appendix C. All tables and figures in this appendix may be interpreted in the same way as the analogous tables and figures in the main body and Appendix C.

The results for the Spanish tests are largely similar to the corresponding results for the general Mathematics tests. The discrepancies are primarily due to the differences in the items pools, as well as sampling variability. As there are fewer items in the pool for the Spanish test, there are fewer items available to be selected by the CAT engine. Some consequences of this include: 1) slightly smaller correlations between estimated proficiency and overall test difficulty (see Tables 14 and E8); and 2) fewer items that go unused by the CAT engine (see Tables 15 and E9).

Table E1. Number of Operational Items in Mathematics Adaptive Test Item Pool

Grade	Calculator	Total	Claim 1	Claim 2	Claim 3	Claim 4
3	No	350	219	47	45	39
4	No	353	207	46	55	45
5	No	365	196	43	68	58
6	Yes	191	85	32	45	29
	No	175	169	0	6	0
7	Yes	222	131	24	42	25
	No	84	84	0	0	0
8	Yes	220	134	15	47	24
	No	72	72	0	0	0
11	Yes	414	225	45	100	44
	No	50	38	0	12	0

Note: Item counts current as of 2015-04-03.

Table E2. Number of Operational Items in Mathematics Performance Task Item Pool

Grade	Number of Items					Number of Stimuli (Across Claims)
	Total	Claim 1	Claim 2	Claim 3	Claim 4	
3	47	0	19	15	13	8
4	38	0	17	11	10	8
5	41	0	14	14	13	7
6	34	0	12	12	10	6
7	27	0	11	7	9	5
8	29	0	10	10	9	6
11	35	0	11	14	10	6

Note: Item counts current as of 2015-04-03.

Table E3. PT Tests with Blueprint Violations (PT Portion Only)

Grade	Subject	Blueprint Specification	Blueprint Requirement			Number of Tests		
			Pg. #	Min	Max	Total	Below Min.	Above Max.
3	Math	Claim 2 (Problem Solving)	5	1	2	387	0	387
3	Math	Claim 4 (Modeling and Data Analysis)	5	1	3	124	124	0
3	Math	Claim 3 (Communicating Reason)	5	0	2	124	0	124
4	Math	Claim 2 (Problem Solving)	7	1	2	122	0	122
4	Math	Claim 4 (Modeling and Data Analysis)	7	1	3	125	125	0
5	Math	Claim 2 (Problem Solving)	9	1	2	279	0	279
7	Math	Claim 2 (Problem Solving)	13	1	2	190	0	190
8	Math	Claim 4 (Modeling and Data Analysis)	15	1	3	168	168	0
11	Math	Claim 4 (Modeling and Data Analysis)	17	1	3	163	163	0
11	Math	Claim 3 (Communicating Reason)	17	0	2	163	0	163

Table E4. Summaries of Difficulty of Item Pool and Estimated Student Proficiency for Mathematics

	Items		Proficiency	
Grade	Mean	SD	Mean	SD
3	-0.78	1.05	-1.34	1.02
4	0.05	1.01	-0.78	1.07
5	0.69	1.03	-0.44	1.17
6	1.12	1.25	-0.18	1.27
7	1.82	1.26	-0.08	1.39
8	2.35	1.40	0.11	1.49
11	2.95	1.53	0.37	1.62

Table E5. Bias of the Estimated Proficiencies: Mathematics

Grade	Mean Bias	SE of Mean Bias	p-value for the *z*-Test	MSE	95% CI Miss Rate	99% CI Miss Rate
Overall						
3	0.00	0.03	0.99	0.07	4.6	1.4
4	0.02	0.03	0.59	0.08	4.9	0.3
5	0.03	0.03	0.34	0.12	5.3	1
6	0.02	0.04	0.64	0.13	3.8	1.3
7	0.01	0.04	0.74	0.17	4.7	0.9
8	-0.01	0.05	0.75	0.22	4.9	1.5
11	0.05	0.05	0.27	0.30	5.1	1.1
Claim 1: Concepts and Procedures						
3	-0.01	0.03	0.68	0.12	4.1	0.7
4	0.03	0.03	0.33	0.16	3.9	0.6
5	0.07	0.03	0.06	0.24	4.3	0.8
6	0.03	0.04	0.43	0.23	5.3	1.1
7	0.03	0.04	0.44	0.30	5.8	1.7
8	0.02	0.05	0.70	0.37	4.9	0.9
11	0.09	0.05	0.08	0.57	5.2	1.7
Claim 2/4: Problem Solving/Modeling and Data Analysis						
3	0.10	0.03	0.00	0.43	10.7	5.3
4	0.15	0.03	0.00	0.69	12.4	6.3
5	0.25	0.04	0.00	0.94	14.5	8.7
6	0.30	0.04	0.00	1.22	18	10.4
7	0.28	0.04	0.00	1.27	15.2	7
8	0.30	0.05	0.00	1.49	18	8.5
11	0.48	0.05	0.00	1.89	19.9	10.4
Claim 3: Communicating Reasoning						
3	0.25	0.03	0.00	0.79	17.2	12.1
4	0.23	0.03	0.00	0.79	14.9	9
5	0.18	0.03	0.00	0.65	10.1	5.8
6	0.27	0.04	0.00	1.06	14.3	7.5
7	0.55	0.05	0.00	2.11	23.7	14.5
8	0.31	0.05	0.00	1.58	10.7	5.6
11	0.16	0.05	0.00	1.10	8.4	3.9

Table E6. Overall Score and Claim Score Precision/Reliability: Mathematics

Grade	Overall Math					Claim 1					Claim 2/4					Claim 3				
	ave # items	SD($\hat{\theta}$)	mean SEM	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$	ave # items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$
3	39.88	1.0	0.26	0.26	0.93	20.00	1.1	0.35	0.35	0.89	10.03	1.2	0.50	0.65	0.72	9.85	1.4	0.62	0.89	0.59
4	38.62	1.1	0.29	0.29	0.93	19.87	1.1	0.39	0.40	0.87	9.38	1.4	0.61	0.83	0.65	9.37	1.5	0.59	0.89	0.63
5	39.85	1.2	0.35	0.35	0.91	20.00	1.2	0.49	0.49	0.84	9.85	1.6	0.63	0.97	0.62	10.00	1.4	0.63	0.80	0.68
6	38.67	1.3	0.36	0.36	0.92	19.00	1.3	0.47	0.48	0.87	9.67	1.8	0.69	1.10	0.60	10.00	1.7	0.79	1.03	0.61
7	38.99	1.4	0.44	0.42	0.91	20.00	1.5	0.56	0.55	0.86	10.00	1.8	0.83	1.12	0.61	8.99	2.0	0.90	1.45	0.46
8	38.83	1.5	0.49	0.47	0.90	20.00	1.5	0.63	0.61	0.84	9.16	2.0	0.83	1.22	0.61	9.67	1.9	1.23	1.26	0.54
11	41.77	1.6	0.55	0.55	0.89	22.00	1.7	0.74	0.76	0.80	9.45	2.1	0.88	1.38	0.58	10.33	1.9	1.02	1.05	0.68

Table E7. Average Standard Errors by Grade and by Deciles of True Proficiency Scores for Mathematics

Grade	**Proficiency Score Distribution**										
		Decile 1	**Decile 2**	**Decile 3**	**Decile 4**	**Decile 5**	**Decile 6**	**Decile 7**	**Decile 8**	**Decile 9**	**Decile 10**	**Overall**
3	0.37	0.28	0.25	0.24	0.23	0.22	0.21	0.21	0.21	0.23	0.37	0.28
4	0.49	0.34	0.29	0.25	0.24	0.23	0.22	0.22	0.22	0.24	0.49	0.34
5	0.61	0.42	0.35	0.30	0.28	0.26	0.24	0.22	0.21	0.22	0.61	0.42
6	0.62	0.43	0.37	0.33	0.31	0.28	0.27	0.25	0.24	0.26	0.62	0.43
7	0.76	0.56	0.48	0.41	0.37	0.32	0.29	0.26	0.24	0.24	0.76	0.56
8	0.81	0.64	0.54	0.47	0.42	0.38	0.34	0.30	0.27	0.26	0.81	0.64
11	0.89	0.73	0.61	0.54	0.47	0.41	0.37	0.32	0.28	0.27	0.89	0.73

Table E8. Correlations between True and Estimated Proficiency, and between Estimated Proficiency and Overall Test Difficulty for Mathematics

Grade	True Proficiency and Estimated Proficiency	Estimated Proficiency and Overall Test Difficulty
3	0.97	0.86
4	0.96	0.85
5	0.95	0.82
6	0.96	0.82
7	0.95	0.75
8	0.95	0.77
11	0.94	0.74

Note. Overall test difficulty is the average of item location parameters for all items in the test instance

Table E9. Percent of Items by Exposure Rate for Mathematics

Grade	Total Items	Exposure Rate					
		Unused	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
3	350	0	88.29	11.14	0.57	0	0
4	353	0	91.78	7.93	0.28	0	0
5	365	0	92.05	7.40	0.55	0	0
6	366	0	90.71	9.02	0.27	0	0
7	306	0	87.58	10.78	1.63	0	0
8	292	0	83.90	13.70	2.05	0	0.34
11	464	0	92.89	5.39	1.51	0	0.22

Figure E10. Mathematics: Histograms of Item Exposure Rates (Grades 3-6)

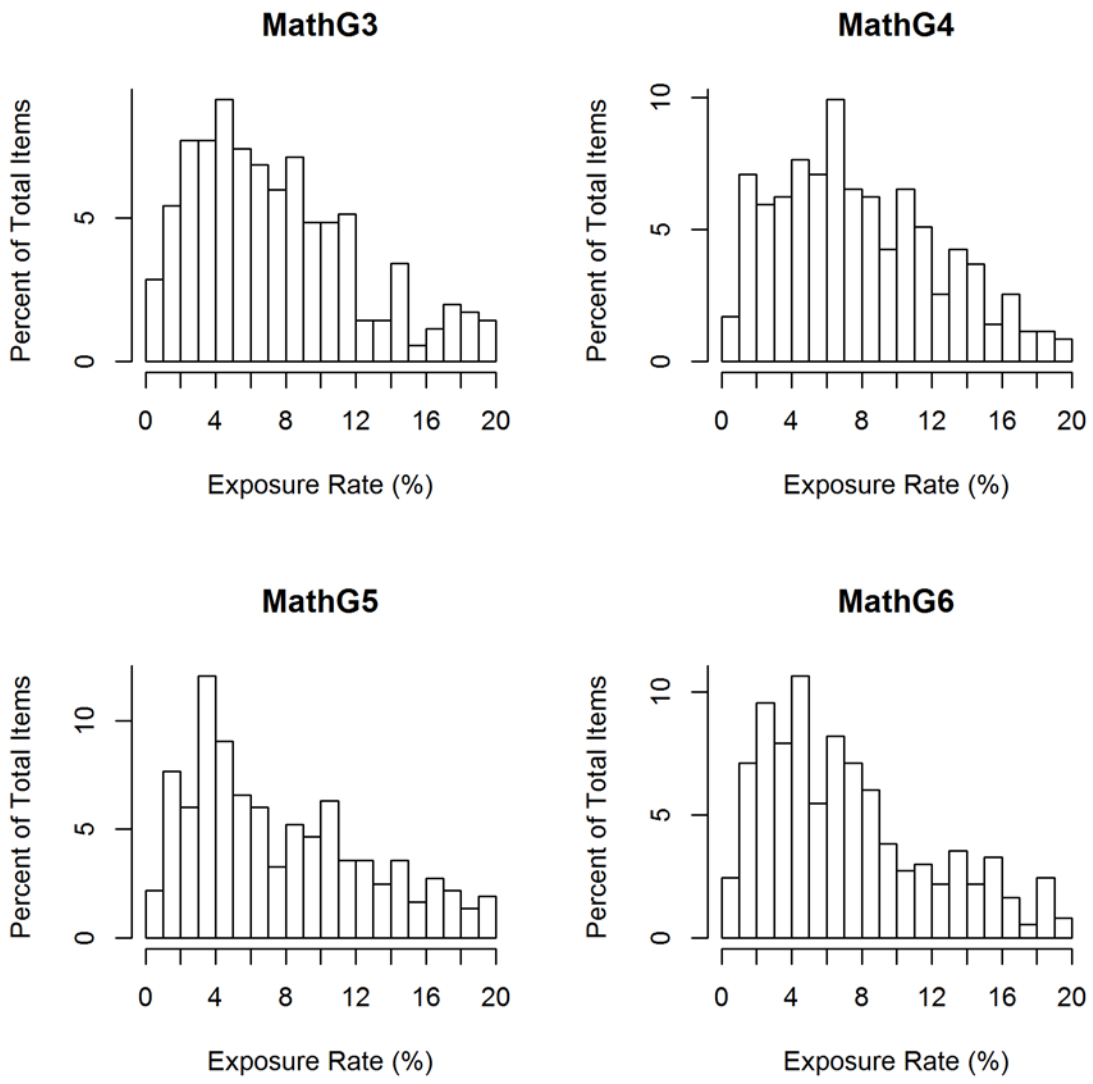


Figure E11. Mathematics: Histograms of Item Exposure Rates (Grades 7, 8, and HS)

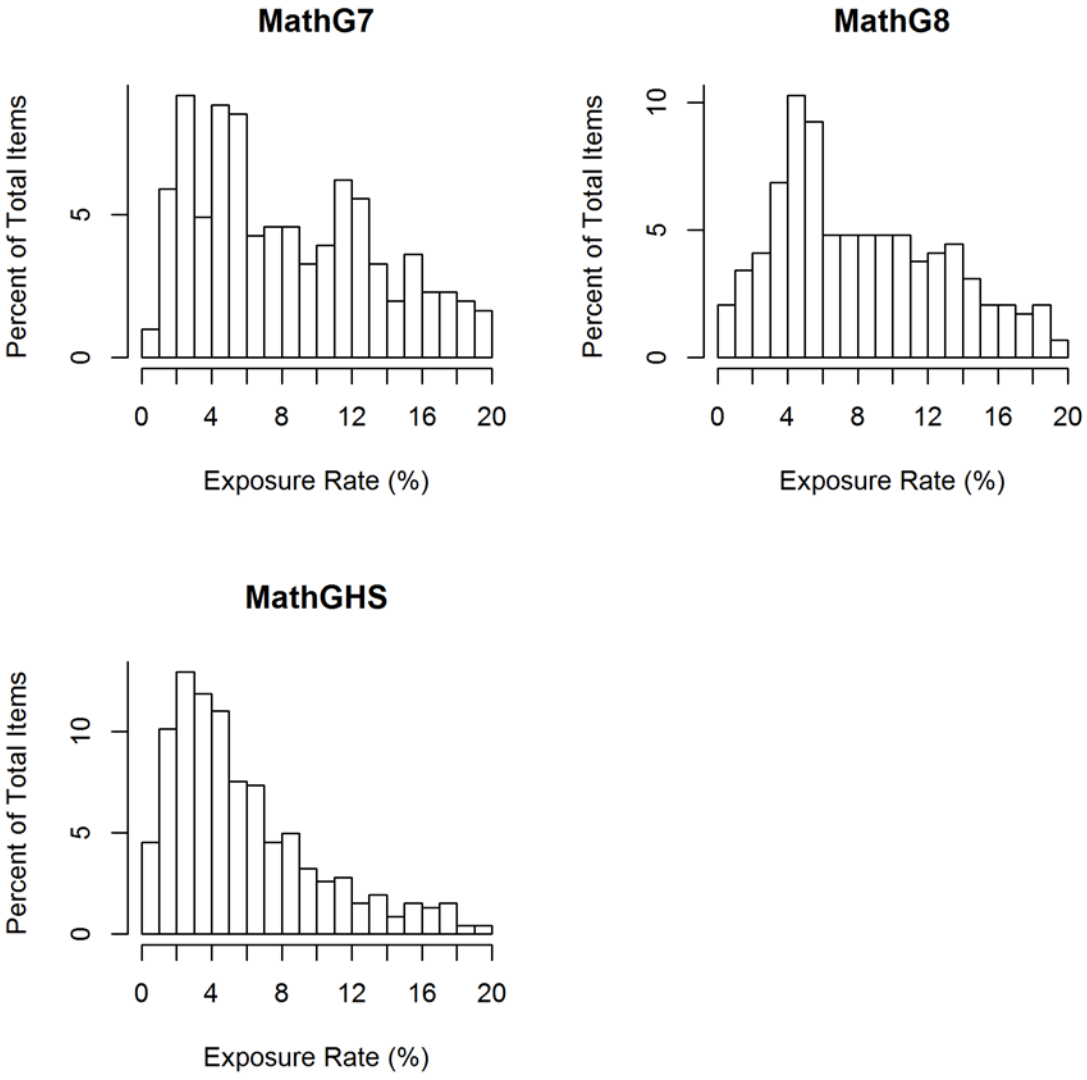


Table E12. Percentage of Math/CAT Test Administration Meeting Blueprint Requirements for Each Claim

Grade	Claim	# Items required		Page	Item Requirement		
		Min	Max		Under	Match	Exceed
3	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
3	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
3	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
4	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
4	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
4	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
5	Claim 1: Concepts and Procedures	17	20	1	0.0%	100.0%	0.0%
5	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	1	0.0%	100.0%	0.0%
5	Claim 3: Communicating Reasoning	8	8	1	0.0%	100.0%	0.0%
6	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
6	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
6	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
7	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
7	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
7	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
8	Claim 1: Concepts and Procedures	16	20	2	0.0%	100.0%	0.0%
8	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	2	0.0%	100.0%	0.0%
8	Claim 3: Communicating Reasoning	8	8	2	0.0%	100.0%	0.0%
11	Claim 1: Concepts and Procedures	19	22	3	0.0%	100.0%	0.0%
11	Claim 2/4: Problem Solving/Modeling and Data Analysis	6	6	3	0.0%	100.0%	0.0%
11	Claim 3: Communicating Reasoning	8	8	3	0.0%	100.0%	0.0%

Table E13. Percentage of Math/PT Test Administration Meeting Blueprint Requirements for Each Claim

Grade	Claim	# Items required		Page	Item Requirement		
		Min	Max		Under	Match	Exceed
3	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.00%	72.60%	27.40%
3	Claim 3: Communicating Reasoning	0	2	1	0.00%	87.60%	12.40%
4	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.00%	100.00%	0.00%
4	Claim 3: Communicating Reasoning	0	2	1	0.00%	100.00%	0.00%
5	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	1	0.00%	100.00%	0.00%
5	Claim 3: Communicating Reasoning	0	2	1	0.00%	100.00%	0.00%
6	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.00%	100.00%	0.00%
6	Claim 3: Communicating Reasoning	0	2	2	0.00%	100.00%	0.00%
7	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.00%	100.00%	0.00%
7	Claim 3: Communicating Reasoning	0	2	2	0.00%	100.00%	0.00%
8	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	2	0.00%	100.00%	0.00%
8	Claim 3: Communicating Reasoning	0	2	2	0.00%	100.00%	0.00%
11	Claim 2/4: Problem Solving/Modeling and Data Analysis	2	4	3	0.00%	100.00%	0.00%
11	Claim 3: Communicating Reasoning	0	2	3	0.00%	83.70%	16.30%

References

American Institutes for Research (2014). *Smarter Balanced Scoring Specification: 2014-2015 Administration.* Washington, DC: Author. Retrieved from http://www.smarterapp.org/documents/TestScoringSpecs2014-2015.pdf

American Institutes for Research (2015). *2014-2015 Smarter Balanced Summative Assessments: Testing Procedures for Adaptive Item-Selection Algorithm.* Washington, DC: Author.

Chang, H. H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*(3), 441-450.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Smarter Balanced Assessment Consortium. (2015a). *ELA/Literacy Summative Assessment Blueprint as of 02/09/15.* Los Angeles, CA: Author. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/02/ELA_Blueprint.pdf

Smarter Balanced Assessment Consortium. (2015b). *Mathematics Summative Assessment Blueprint as of 02/09/15.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/02/Mathematics_Blueprint.pdf