UCLA | CRESST NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

SIMULATION-BASED EVALUATION OF THE 2014-2015 SMARTER BALANCED SUMMATIVE ASSESSMENTS: ACCOMMODATED ITEM POOLS

CRESST Psychometrics Team
May 27, 2016

TABLE OF CONTENTS

Introduction

This report describes a study simulating the adaptive administration of the Smarter Balanced summative assessments using the accommodated item pools utilized during the 2014-2015 school year. The study was conducted to examine properties of the simulated tests (such as blueprint fulfillment and item exposure) and the quality of the resulting 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). At each grade level, one ELA item pool was evaluated: the American Sign Language (ASL) pool. For Math, the evaluation in each grade level included the ASL pool, as well as a translated glossaries pool. The simulated tests included both the computerized adaptive test (CAT) and performance task (PT) components, thus mimicking the operational summative tests.

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). Final score estimates were obtained based on the combined performance across the CAT and PT components.

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.

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.

This report is organized as follows. In Chapter 1, we briefly describe the CRESST CAT engine design. Chapter 2 provides an overview of the various ways in which the performance of the item pools and simulated tests was evaluated. We include descriptions of the specific indices used to examine score recovery precision. Chapters 3 through 5 present the simulation results for each pool. These results include summaries of blueprint satisfaction, as well as bias and precision of the proficiency estimates that were obtained, based on the CAT and PT items administered. Item exposure rates for the CAT portion are also presented. Chapter 3 presents results for the English language arts/literacy tests based on the ASL pool. Chapters 4 and 5 present the results for the simulated Mathematics tests based on the ASL and translated glossaries pools respectively.

Chapter 1. Adaptive Test Engine Design

In this chapter, we provide a brief description and list a few key features of the CAT engine CRESST developed for the simulation study.

General Approach to Item Selection

The CAT engine was written and implemented in R (R Core Team, 2014). For both ELA/L and Mathematics, the test proceeds claim by claim, and the order in which the claims appear is randomized over students. In addition, the Mathematics test proceeds "cell" by "cell" within a claim. The order in which the cells are presented is randomized. 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 for which the test 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. To satisfy the blueprint requirements, the engine always tries to administer the maximum number of items. Within the CAT portion of the test, the algorithm does not allow administration of an item that would result in a maximum requirement being exceeded.

The test engine proceeds adaptively in the following manner. Instead of utilizing a current point estimate of the student proficiency, the engine utilizes a current estimate of the posterior distribution of the student proficiency, given the observed pattern of item responses up to that point in the test. To select the next item/stimulus to be administered, a posterior-weighted item information index is calculated for all eligible items (i.e., those items within the current cell that have not been used). This "baseline" weight is obtained by integrating item information function values over the current posterior distribution. Before the first item is administered, the engine uses the generating (population) proficiency distribution as the current estimate (see Table 1.1 below). A key advantage of using a posterior-weighted item information index is that it takes more global and less "greedy" perspective of item optimality than either the item difficulty parameter or the Fisher information function alone. Considerable research (e.g., Chang & Ying, 2008) has indicated the negative consequences of greedy item selection algorithms (e.g., approaches based on maximum Fisher information).

Item 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. By not picking the item/stimuli with the largest weight, the algorithm may perform better in terms of item exposure, but with a slight decrement in the resulting measurement precision. Weights for items that shared a common stimulus are initially summed such that these item sets shared a common weight and sampling probability and were treated as a single unit (i.e., as a single item). However, once a set containing multiple items is selected for administration, the algorithm proceeds adaptively by selecting items within that set until a specified maximum number of items for the stimulus is reached. Upon administration of an item, a response is randomly generated based on the item's parameters and the simulee's true score, and the posterior for the simulee is then updated (for all machine-scored items; hand-scored items only contribute to the final score estimate).

The CRESST CAT terminates after cycling through all required claims in the blueprint. 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 assigns stimuli to meet the stimulus requirements specified in the blueprint. All available items for the administered stimulus are administered to the simulee. In some cases, this can 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 were drawn from a normal distribution with mean and standard deviation parameters specific to grade and subject.[1] These population parameters are presented in Table 1.1.

Table 1.1. Population Proficiency Distributions and Obtainable Score Ranges, by Grade and Subject

Grade	Population Parameters		Obtainable Score Range	
	Mean	SD	LOT	HOT
		English Language Arts/Literacy		
3	-1.24	1.06	-4.59	1.34
4	-0.75	1.11	-4.40	1.80
5	-0.31	1.10	-3.58	2.25
6	-0.06	1.11	-3.48	2.51
7	0.11	1.13	-2.91	2.75
8	0.38	1.13	-2.57	3.04
11	0.53	1.19	-2.44	3.34
		Mathematics		
3	-1.29	0.97	-4.11	1.33
4	-0.71	1.00	-3.92	1.82
5	-0.35	1.08	-3.73	2.33
6	-0.10	1.19	-3.53	2.95
7	0.01	1.33	-3.34	3.32
8	0.18	1.42	-3.15	3.63
11	0.51	1.52	-2.96	4.38

As noted above, these normal distributions were also used as the initial proficiency estimate. That is, before any items are administered, these population or "prior" distributions are treated as current "posterior" distributions for all simulees for the purpose of selecting the first item to be administered.

Final Score Estimates

Once the CAT and PT components were completed, maximum likelihood (ML) scoring was used 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 score estimates of $\pm\infty$ and undefined standard errors. Following the SBAC Scoring Specification, any all-correct response vectors were assigned the Highest Obtainable T-score (HOT), while any

[1] The same population distribution was used for all item pools (general, Braille, Spanish, Translated Glossary, ASL) within each grade and subject.

all-incorrect response vectors were assigned the Lowest Obtainable T-score (LOT). LOT and HOT values are shown in Table 1.1, above. Standard errors for these cases were calculated from the test information function for the specific items administered to the simulee, evaluated at the HOT or LOT. Response patterns that produced finite ML estimates outside the obtainable range were also assigned the LOT or HOT, as appropriate.

Chapter 2. Summary Indices

In this chapter, we describe the statistical summaries used in this report to evaluate the performance of the item pools and the adaptive test administration. Given an examinee's true score, θ_i, and final score estimate, $\hat{\theta}_i$, average *bias* in the score estimates is defined as

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

and the error variance of the estimated bias is

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

where $\bar{\hat{\theta}}$ is the average of the $\hat{\theta}_i$ and N denotes the number of simulees (N =1000 for all conditions). Statistical significance of the bias is tested using a z-test,

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

for which we report the p-value for a two-tailed test. The mean squared error (MSE) in the estimated scores is

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

and its square root is the root mean squared error ($RMSE$). Marginal reliability of the simulated tests is estimated as

$$\bar{\rho} = 1 - \frac{MSE}{var(\hat{\theta})}.$$

The average standard error of the score estimates is

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

where $SE\left(\hat{\theta}_i\right)$ is the standard error of the estimated score for simulee i. Miss rates for the 95% and 99% confidence intervals are computed by computing the percentage of cases

for which the intervals computed from the score estimates and standard errors do not contain the true score confidence interval coverage). Specifically, a t-statistic is computed for each case:

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

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.

Table Overview

In the following chapters, results are broken down by subject (ELA/L and Math) and item pool (ASL for ELA/L; ASL and Translated Glossaries for Math):

- Chapter 3. English Language Arts/Literacy ASL Pool Results
- Chapter 4. Mathematics ASL Pool Results
- Chapter 5. Mathematics Translated Glossaries Pool Results

Within each of these chapters, 13 tables and two figures are provided. An overview of the information in each table is provided in the summary below.

The first four tables in each chapter describe the item pool. Within each table, results are reported for each of the tested grade levels (3-8 and high school). Note that in describing the tables and figures here, we use "C" as a stand-in for the actual chapter number (3-7). Table C.1 contains the percentage of simulees whose scores were assigned the LOT of HOT values due to having either infinite ML score estimates (a result of a response pattern with either all correct, or all incorrect responses) or finite estimates that were outside the acceptable range. Table C.2 displays the total number of items in the operational pool (excluding extended item pool) used in CAT simulations, while Table C.3 displays these values for the PT portion of the test. Table C.4 provides the means and standard deviations of item difficulty for the CAT and PT items. This table also reports the means and standard deviations in the final score estimates.

The second section of results focuses on score precision and bias. Table C.5 presents mean *bias* and *MSE* in the student proficiency estimates, as well as the 95% and 99% confidence interval miss rates. Table C.6 reports the average standard errors, *RMSE*, and

marginal reliability. Table C.7 shows the average standard errors within true score deciles. Table C.8 presents correlations between true and estimated scale scores, as well as the correlations between the estimated score and the average difficulty of items administered to the examinee.

The third section within each chapter provides details concerning fulfillment of the test blueprint. The 2014-15 operational item pools used in these simulations were current as of September 29, 2015. Fulfillment was judged by comparing the items administered to each simulated examinee against the Smarter Balanced Assessment Consortium ELA and Mathematics Blueprints dated February 9, 2015 (Smarter Balanced Assessment Consortium 2015a/b). Table C.9 provides a list of blueprint violations for the CAT component; Table C.10 lists violations within the PT component. This table shows, by grade, the blueprint specification, the page number in the blueprint document in which the requirement appears, the minimum and maximum number of items specified, the number of tests in which a 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. In Tables C.11 and C.12, we describe the percentage of tests that met the blueprint constraints for passages and the total number of items within each claim. Results for the CAT component are shown in Table C.11; results for PT are shown in Table C.12. These tables present the page number in which the requirement appears, the min and max requirement for each claim, and percentages of test administrations meeting these particular requirements.

In the final section of each chapter, we examine exposure rates for items in the CAT component. We calculated the percentage of simulees to whom each item was administered. Next, items were binned according to exposure rate. The bins were defined as follows: unused, 0%-20%, 21%-40%, 41%-60%, 61%-80%, and 81%-100%. These exposure rates are presented in Table C.13. Finally, we present histograms of the item exposure rates (focusing on the range from 0% to 20%). Figure C.1 shows these results for grades 3-6; Figure C.2 shows the results for grades 7, 8, and high school.

Chapter 3. Results for the English Language Arts/Literacy ASL Pool

In this chapter, we present the results of the simulated administration of the ELA ASL pool. Within each of the seven grade levels, true values for student proficiency (θ) were drawn for 1,000 simulated examinees from a normal distribution with population parameters shown in Table 3.1, below. At the completion of the simulated test administration, some examinee score estimates were infinite (due to having achieved the minimum score on all items or achieving the maximum score on all items) or outside the specified range of obtainable scores. These estimates were assigned the lowest or highest obtainable T-score (LOT and HOT, respectively). Table 3.1 shows the percentage of cases assigned to the LOT or HOT within each of these possible options.

Table 3.1. Characteristics of Simulated and Estimated ELA Proficiencies

Grade	Population Parameters		Obtainable Range		% Infinite ML Scores		% Outside Range	
	Mean	SD	LOT	HOT	LOT	HOT	LOT	HOT
3	-1.24	1.06	-4.59	1.34	0.4	0.0	0.1	1.4
4	-0.75	1.11	-4.40	1.80	0.2	0.0	0.5	1.9
5	-0.31	1.10	-3.58	2.25	0.0	0.0	0.8	1.9
6	-0.06	1.11	-3.48	2.51	0.2	0.0	0.6	1.6
7	0.11	1.13	-2.91	2.75	0.1	0.0	1.1	1.5
8	0.38	1.13	-2.57	3.04	0.0	0.0	1.0	1.5
11	0.53	1.19	-2.44	3.34	0.1	0.0	2.0	1.3

The number of operational CAT and PT items in each grade are summarized in Tables 3.2 and 3.3. In grades 3-8, the number of operational CAT items ranges from 426 to 1244. The Grade 11 CAT item pool is substantially larger (with 1244 items). Across grade levels, Claim 1 items represent 35-41% of the CAT pool.

Table 3.2. Number of Operational Adaptive Items – ELA ASL 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	533	217	166	60	90	18	16	24
4	507	177	166	67	97	15	11	25
5	496	194	159	58	85	16	13	23
6	495	175	168	63	89	7	21	25
7	453	183	154	62	54	5	23	24
8	426	161	147	58	60	5	18	21
11	1244	499	379	123	243	27	58	45

Note. Item counts current as of 2015-09-29.

The numbers of PT stimuli ranges from 14 (with 62 items) in grade 3 to 24 (with 105

items) in grade 11. As seen in Table 3.3, PT items do not contribute to scores for Claims 1 and 3.

Table 3.3. Number of Operational Performance Task Items – ELA ASL Pool

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

Note. Item counts current as of 2015-09-29.

The overall performance of the adaptive test algorithm depends, in part, on the availability of items that are informative at the levels of proficiency found in the examinee population. Table 3.4 shows the means and standard deviations of item difficulty for the CAT and PT portions in each grade band. The means and standard deviations of the final proficiency estimates for the simulated examinees are also shown.

Table 3.4. ELA ASL Pool - Item Difficulty and Estimated Student Math Proficiency

Grade	CAT Items		PT Items		Proficiency	
	Mean	SD	Mean	SD	Mean	SD
3	-0.47	1.13	0.21	0.96	-1.34	1.11
4	0.13	1.30	0.45	0.80	-0.84	1.19
5	0.48	1.22	0.74	0.85	-0.38	1.16
6	1.01	1.31	0.92	0.85	-0.13	1.17
7	1.13	1.34	1.15	0.86	0.04	1.19
8	1.40	1.34	1.27	1.01	0.31	1.19
11	1.75	1.34	1.84	0.82	0.46	1.25

The mean difficulty of items increases across grade levels. However, it is notable that average item difficulty is consistently higher than the average proficiency of the examinees, which could affect the efficiency of the adaptive test.

Table 3.5 presents a summary of average bias in the overall and claim score estimates, along with the 95% and 99% confidence interval miss rates.

Table 3.5. Bias of the Estimated Proficiencies – ELA ASL Pool

Grade	Bias	SE(bias)	*p*-value	MSE	95% CI Miss Rate	99% CI Miss Rate
			Overall English Language Arts/Literacy			
3	0.01	0.03	0.88	0.01	5.0	1.2
4	0.03	0.04	0.47	0.12	4.6	0.6
5	0.01	0.03	0.80	0.10	5.1	0.7
6	0.01	0.04	0.75	0.12	5.8	0.9
7	0.02	0.04	0.67	0.12	5.0	0.5
8	-0.01	0.04	0.71	0.11	3.8	1.3
11	0.00	0.04	0.95	0.14	4.6	1.1
			Claim 1: Reading			
3	0.04	0.03	0.22	0.31	6.4	1.4
4	0.10	0.04	0.00	0.39	5.4	2.1
5	0.05	0.04	0.14	0.32	5.3	1.9
6	0.05	0.04	0.16	0.41	4.6	1.3
7	0.07	0.04	0.07	0.45	5.8	2.0
8	0.05	0.04	0.14	0.39	5.6	1.8
11	0.03	0.04	0.46	0.39	5.8	1.1
			Claim 2: Writing			
3	0.04	0.03	0.25	0.28	3.7	0.9
4	0.04	0.04	0.24	0.29	4.1	1.3
5	0.01	0.03	0.76	0.29	5.2	1.5
6	0.05	0.04	0.18	0.32	5.5	1.2
7	0.04	0.04	0.32	0.35	6.2	1.6
8	0.02	0.04	0.50	0.32	4.8	1.5
11	0.03	0.04	0.36	0.45	5.4	1.6
			Claim 3: Speaking/Listening			
3	0.11	0.03	0.00	0.86	11.2	5.9
4	0.06	0.04	0.08	0.78	8.5	4.2
5	0.08	0.04	0.03	0.83	8.8	5.5
6	0.11	0.04	0.00	0.93	8.3	5.4
7	0.03	0.04	0.40	0.90	8.9	5.1
8	0.05	0.04	0.16	0.96	9.0	5.4
11	-0.02	0.04	0.66	0.88	6.9	4.7
			Claim 4: Research			
3	0.21	0.03	0.00	0.93	12.8	8.1
4	0.16	0.04	0.00	0.90	10.3	6.4
5	0.10	0.04	0.01	0.65	9.8	5.9
6	0.23	0.04	0.00	1.15	15.1	10.5
7	0.17	0.04	0.00	0.87	11.9	7.8
8	0.10	0.04	0.00	0.72	10.4	5.3
11	0.14	0.04	0.00	0.87	12.2	7.2

Mean bias in the overall score estimates is small for all grade levels (with a range of 0.00 to 0.03 on the logit scale), and the null hypothesis that the mean bias in the overall scores is equal to zero in the population cannot be rejected (*p*-values are 0.55 or greater).

On the other hand, there is evidence of bias in claim score estimates. This bias appears to be due to the assignment of the LOT and HOT values for examinees with extreme score estimates for a given claim—in particular, those examinees with an infinite ML score estimate due to a perfect score patterns (i.e., achieving either the minimum score for all items or the maximum for all items). Such score patterns are of course far more likely within a claim (based on a relatively small number of items) than for the full test. Importantly, patterns in which all item scores received the minimum were far more frequent than patterns with all items receiving the maximum score. The fact that more infinite scores were replaced with the LOT than with the HOT value resulted in the observed average bias in the claim score results. It should be noted, however, that the assignment to LOT or HOT values would have little impact on the claim-level classifications currently used in practice (below standard, at or near standard, above standard).

Confidence interval miss rates for overall scores are very close to their expected levels. The overall score miss rate for the 95% confidence interval—expected to be 5%—ranges from 3.8% to 5.8%, while the miss rate for the 99% confidence interval—expected to be 1%—ranges from 0.5% to 1.3%. Taken together with the results concerning average bias, these confidence interval miss rates suggest that the standard errors of measurement for the overall score estimates are well-calibrated (i.e., correctly reflecting the level of score uncertainty).

The confidence interval miss rates for the claim scores are less consistent and—for Claims 3 and 4, in particular—show evidence of poor calibration. This is not surprising, however, given the bias observed in these score estimates. It is likely that the deviations of the miss rates from their expected values are due to the assignment of the LOT and HOT for examinees with perfect item score patterns. Because such patterns are relatively common for the small number of items in a claim, the LOT or HOT is a poor estimate of the true score for many examinees. This makes it less likely that the confidence interval around the LOT/HOT will include the true score, increasing the miss rate.

Table 3.6 summarizes the standard deviation in score estimates, average standard error,

actual error (RMSE), and marginal reliability for the overall and claim scores.

Table 3.6. Overall Score and Claim Score Precision/Reliability – ELA ASL Pool

Grade	mean # Items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$
		Overall English Language Arts/Literacy			
3	45.3	1.1	.29	.30	.93
4	45.4	1.2	.31	.33	.92
5	45.5	1.2	.30	.31	.93
6	43.2	1.2	.32	.34	.91
7	42.9	1.2	.34	.35	.91
8	43.2	1.2	.34	.34	.92
11	45.4	1.3	.36	.37	.91
		Claim 1: Reading			
3	16	1.3	.48	.56	.80
4	16	1.4	.55	.63	.79
5	16	1.3	.53	.56	.80
6	14	1.3	.62	.64	.76
7	14	1.3	.63	.67	.74
8	14	1.3	.60	.62	.77
11	16	1.4	.61	.63	.79
		Claim 2: Writing			
3	12	1.2	.52	.53	.81
4	12	1.3	.52	.54	.83
5	12	1.3	.52	.54	.82
6	12	1.3	.51	.57	.81
7	12	1.3	.56	.60	.79
8	12	1.3	.55	.57	.81
11	12	1.4	.64	.67	.77
		Claim 3: Speaking/Listening			
3	9	1.5	.78	.93	.60
4	9	1.5	.78	.88	.63
5	9	1.5	.82	.91	.62
6	9	1.5	.87	.96	.57
7	9	1.5	.89	.95	.60
8	9	1.5	.93	.98	.55
11	9	1.5	.93	.94	.61
		Claim 4: Research			
3	8.4	1.5	.68	.96	.59
4	8.4	1.5	.76	.95	.60
5	8.7	1.4	.66	.81	.67
6	8.3	1.6	.72	1.07	.56
7	8.0	1.5	.77	.93	.61
8	8.3	1.4	.75	.85	.65
11	8.4	1.5	.79	.93	.63

The results in Table 3.6 indicate that the standard errors for the overall ELA score

estimates are well-calibrated; average standard errors within each grade closely resemble the RMSE values. There are discrepancies between the average standard errors and the RMSE values for the claim scores, with the average standard error consistently smaller than the RMSE. This result is consistent with the earlier findings concerning average bias in the claim score estimates and the confidence interval miss rates (Table 3.5).

Marginal reliability was computed from the RMSE and observed variance in the scale score estimates, as described in Chapter 2. For the overall score, marginal reliability ranged from 0.91 to 0.93. Marginal reliability for the claim scores ranged from 0.74 to 0.80 for Claim 1 (Reading), 0.77 to 0.83 for Claim 2 (Writing), 0.55 to 0.63 for Claim 3 (Speaking/Listening), and 0.56 to 0.67 for Claim 4 (Research). The lower levels of marginal reliability for Claims 3 and 4 are expected, given that these scores are based on fewer items than the scores for Claims 1 and 2.

Table 3.7 summarizes the average standard errors for the overall ELA score within true score deciles. The averages in deciles 4-10 (i.e., for all examinees above the 30th percentile) range from 0.27 to 0.35 for all grade levels. Average standard errors are higher in the lowest deciles and have a range of 0.43 to 0.52 in decile 1. This is consistent with the fact that the item pools tend to an average level of difficulty that is higher than the average proficiency of the population of examinees (as seen in Table 3.4).

Table 3.7. Average Standard Errors by True Proficiency Decile – ELA ASL Pool

Grade	Deciles										Overall
	1	2	3	4	5	6	7	8	9	10	
3	.48	.33	.28	.27	.26	.25	.25	.24	.25	.28	.30
4	.47	.33	.30	.29	.28	.28	.27	.27	.28	.31	.32
5	.43	.31	.28	.28	.27	.27	.27	.27	.28	.31	.31
6	.50	.38	.32	.30	.28	.28	.28	.28	.28	.31	.33
7	.52	.40	.35	.33	.31	.30	.29	.29	.29	.31	.35
8	.49	.37	.33	.32	.30	.30	.30	.30	.30	.32	.34
11	.52	.43	.37	.35	.33	.32	.31	.30	.31	.34	.37

Table 3.8 presents, for each grade level, the correlation between the final score estimates (for overall ELA proficiency) and examinee true scores, as well as the correlation between the final score estimates and overall test difficulty. The overall test difficulty for an examinee is simply the average difficulty for items administered. The correlations between estimated and true proficiencies are quite high (0.96), indicating that the administered items are successful in recovering the rank ordering of students.

Correlations between estimated proficiency and overall test difficulty range from 0.59 to 0.76. These correlations may serve as a crude measure of the extent to which the CAT algorithm has tailored the difficulty of the test to examinee, within the constraints of the blueprint and given the properties of the available pool of items.

Table 3.8. Correlations between True and Estimated Proficiency, and between Estimated Proficiency and Overall Test Difficulty – ELA ASL Pool

Grade	r($\hat{\theta}, \theta$)	r($\hat{\theta}$,overall test difficulty)
3	.96	.71
4	.96	.72
5	.96	.76
6	.96	.68
7	.96	.59
8	.96	.64
11	.96	.60

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

Tables 3.9-3.12 present results concerning the extent to which simulated tests in each grade level fulfilled requirements of the summative test blueprint. These tables identify the particular blueprint specification (including the page of the blueprint document on which the specification is described) and the range of items that are required in order to fulfill the specification.

Tables 3.9 and 3.10 provide counts of the test instances (out of the 1,000 simulated within the grade level) that violated a specification. For the CAT portion of the test, violations were identified in grades 3, 4, 5, 6, 7, 8 and 11. Each of these violations was a failure to include the minimum number of items for a given target or at (or above) a given depth of knowledge (DOK) level. Several violations were present in greater than 5% of test instances. In grade 3 a failure to include the minimum number of DOK 3 items in Claim 2 was violated in 5.7% of test instances. In grade 4, a failure to include the minimum number of DOK 2 items in Claim 2 was violated in 9.6% of test instances. In grade 7, a failure to include the minimum number of Target 9 items in Claim 2 was violated in 8.4% of test instances. In grade 8, a failure to include the minimum number of DOK 2 items in Claim 2 was violated in 7.8% of cases.

Fully 100% of tests across all grade levels did not meet the requirements of including: at least 1 O/P item, at least 1 E/E item, exactly 3 O/P and E/E items, exactly 1 write brief

text, and only 2 brief revise texts.

Table 3.9. Tests with Blueprint Violations, CAT Component – ELA ASL Pool

Grade	Specification	Requirement			Number of Tests		
		Page	Min	Max	Total	Below	Above
3	Claim 1, DOK=2	4	7	NA	3	3	0
3	Claim 1, DOK>=3	4	2	NA	5	5	0
3	Claim 1 (Literary), Target 2	4	1	2	4	4	0
3	Claim 1 (Literary), Target 4	4	1	2	48	48	0
3	Claim 1 (Literary), Target 9	4	1	2	11	11	0
3	Claim 1 (Informational), Target 11	4	1	2	8	8	0
3	Claim 2, DOK=2	5	5	5	18	18	0
3	Claim 2, DOK>=3	5	1	NA	57	57	0
3	Claim 2 (at least 1 O/P item), Target 1a/3a/6a & 1b/3b/6b	5	1	2	1000	1000	0
3	Claim 2 (at least 1 E/E item) Target 1a/3a/6a & 1b/3b/6b	5	1	2	1000	1000	0
3	Claim 2 (exactly 3 O/P and E/E items), Target 1a/3a/6a & 1b/3b/6b	5	3	3	1000	1000	0
3	Claim 2 (exactly 1 write brief text), Target 1a/3a/6a	5	1	1	1000	1000	0
3	Claim 2 (only 2 revise brief text), Target 1b/3b/6b	5	2	2	1000	1000	0
3	Claim 2 (Conventions), Target 9	5	5	5	4	4	0
4	Claim 1, DOK=2	4	7	NA	40	40	0
4	Claim 1 (Literary), Target 4	4	1	2	11	11	0
4	Claim 1 (Informational), Target 9	4	1	2	10	10	0
4	Claim 1 (Informational), Target 11	4	1	2	2	2	0
4	Claim 2, DOK=2	5	4	NA	96	96	0
4	Claim 2 (at least 1 O/P item), Target 1a/3a/6a & 1b/3b/6b	5	1	2	1000	1000	0
4	Claim 2 (at least 1 E/E item) Target 1a/3a/6a & 1b/3b/6b	5	1	2	1000	1000	0
4	Claim 2 (exactly 3 O/P and E/E items), Target 1a/3a/6a & 1b/3b/6b	5	3	3	1000	1000	0
4	Claim 2 (exactly 1 write brief text), Target 1a/3a/6a	5	1	1	1000	1000	0
4	Claim 2 (only 2 revise brief text), Target 1b/3b/6b	5	2	2	1000	1000	0
4	Claim 2 (Conventions), Target 9	5	5	5	1	1	0

Table 3.9. Tests with Blueprint Violations, CAT Component – ELA ASL Pool - Continued

Grade	Specification	Requirement			Number of Tests		
		Page	Min	Max	Total	Below	Above
5	Claim 1, DOK=2	4	7	NA	3	3	0
5	Claim 1, DOK>=3	4	2	NA	1	1	0
5	Claim 1 (Literary), Target 2	4	1	2	6	6	0
5	Claim 1 (Literary), Target 4	4	1	2	22	22	0
5	Claim 1 (Informational), Target 9	4	1	2	13	13	0
5	Claim 1 (Informational), Target 11	4	1	2	3	3	0
5	Claim 2, DOK>=3	5	1	NA	13	13	0
5	Claim 2 (at least 1 O/P item), Target 1a/3a/6a & 1b/3b/6b	5	1	2	1000	1000	0
5	Claim 2 (at least 1 E/E item) Target 1a/3a/6a & 1b/3b/6b	5	1	2	1000	1000	0
5	Claim 2 (exactly 3 O/P and E/E items), Target 1a/3a/6a & 1b/3b/6b	5	3	3	1000	1000	0
5	Claim 2 (exactly 1 write brief text), Target 1a/3a/6a	5	1	1	1000	1000	0
5	Claim 2 (only 2 revise brief text), Target 1b/3b/6b	5	2	2	1000	1000	0
5	Claim 2 (Evidence/Elaboration), Target 8	5	2	2	1	1	0
5	Claim 2 (Conventions), Target 9	5	5	5	7	7	0
6	Claim 1 (Literary), Target 2	7	1	1	4	4	0
6	Claim 1 (Literary), Target 4	7	1	1	16	16	0
6	Claim 1 (1-2 machine scored), Target 1 and Target 4	7	1	2	16	16	0
6	Claim 2, DOK>=2	8	5	NA	5	5	0
6	Claim 2, DOK>=3	8	1	NA	7	7	0
6	Claim 2 (at least 1 O/P item), Target 1a/3a/6a & 1b/3b/6b	8	1	2	1000	1000	0
6	Claim 2 (at least 1 E/E item) Target 1a/3a/6a & 1b/3b/6b	8	1	2	1000	1000	0
6	Claim 2 (exactly 3 O/P and E/E items), Target 1a/3a/6a & 1b/3b/6b	8	3	3	1000	1000	0
6	Claim 2 (exactly 1 write brief text), Target 1a/3a/6a	8	1	1	1000	1000	0
6	Claim 2 (only 2 revise brief text), Target 1b/3b/6b	8	2	2	1000	1000	0
6	Claim 2 (Conventions), Target 9	8	5	5	1	1	0

Table 3.9. Tests with Blueprint Violations, CAT Component – ELA ASL Pool - Continued

Grade	Specification	Requirement			Number of Tests		
		Page	Min	Max	Total	Below	Above
7	Claim 1, DOK>=3	7	2	NA	3	3	0
7	Claim 1 (Literary), Target 2	7	1	1	40	40	0
7	Claim 1 (Informational), Target 4	7	1	1	26	26	0
7	Claim 1 (1-2 machine scored), Target 1 and Target 4	7	1	2	21	21	0
7	Claim 2 (at least 1 O/P item), Target 1a/3a/6a & 1b/3b/6b	8	1	2	1000	1000	0
7	Claim 2 (at least 1 E/E item) Target 1a/3a/6a & 1b/3b/6b	8	1	2	1000	1000	0
7	Claim 2 (exactly 3 O/P and E/E items), Target 1a/3a/6a & 1b/3b/6b	8	3	3	1000	1000	0
7	Claim 2 (exactly 1 write brief text), Target 1a/3a/6a	8	1	1	1000	1000	0
7	Claim 2 (only 2 revise brief text), Target 1b/3b/6b	8	2	2	1000	1000	0
7	Claim 2 (Evidence/Elaboration), Target 8	8	2	2	2	2	0
7	Claim 2 (Conventions), Target 9	8	5	5	84	84	0
8	Claim 1 (Literacy), Target 2	7	1	1	1	1	0
8	Claim 1 (Informational), Target 4	7	1	1	3	3	0
8	Claim 1 (1-2 machine scored), Target 1 and Target 4	7	2	1	3	3	0
8	Claim 2, DOK>=2	8	5	NA	78	78	0
8	Claim 2, DOK>=3	8	1	NA	5	5	0
8	Claim 2 (at least 1 O/P item), Target 1a/3a/6a & 1b/3b/6b	8	1	2	1000	1000	0
8	Claim 2 (at least 1 E/E item) Target 1a/3a/6a & 1b/3b/6b	8	1	2	1000	1000	0
8	Claim 2 (exactly 3 O/P and E/E items), Target 1a/3a/6a & 1b/3b/6b	8	3	3	1000	1000	0
8	Claim 2 (exactly 1 write brief text), Target 1a/3a/6a	8	1	1	1000	1000	0
8	Claim 2 (only 2 revise brief text), Target 1b/3b/6b	8	2	2	1000	1000	0
8	Claim 2 (Evidence/Elaboration), Target 8	8	2	2	1	1	0
8	Claim 2 (Conventions), Target 9	8	5	5	2	2	0

Table 3.9. Tests with Blueprint Violations, CAT Component – ELA ASL Pool - Continued

Grade	Specification	Requirement			Number of Tests		
		Page	Min	Max	Total	Below	Above
11	Claim 2, DOK>=2	10	5	NA	24	24	0
11	Claim 2, DOK>=3	10	1	NA	15	15	0
11	Claim 2 (at least 1 O/P item), Target 1a/3a/6a & 1b/3b/6b	10	1	2	1000	1000	0
11	Claim 2 (at least 1 E/E item) Target 1a/3a/6a & 1b/3b/6b	10	1	2	1000	1000	0
11	Claim 2 (exactly 3 O/P and E/E items), Target 1a/3a/6a & 1b/3b/6b	10	3	3	1000	1000	0
11	Claim 2 (exactly 1 write brief text), Target 1a/3a/6a	10	1	1	1000	1000	0
11	Claim 2 (only 2 revise brief text), Target 1b/3b/6b	10	2	2	1000	1000	0
11	Claim 2 (Evidence/Elaboration), Target 8	10	2	2	1	1	0
11	Claim 2 (Conventions), Target 9	10	5	5	41	41	0

In grade 4, 4.8% of tests violated the blueprint specification for the PT component (Table 3.10).

Table 3.10. Tests with Blueprint Violations, PT Component – ELA ASL Pool

Grade	Specification	Requirement			Number of Tests		
		Page	Min	Max	Total	Below	Above
4	Claim 4	6	2	3	48	48	0
4	Claim 4, DOK>=3	6	2	3	48	48	0
4	Claim 4 (Research), Target 2, Target 3, Target 4	6	2	3	48	48	0

Tables 3.11 and 3.12 present the percentage of test instances that met the blueprint requirements for the total number of items administered within each claim for the CAT and PT components, respectively. As seen in Table 3.11, all tests met the requirements specific to the CAT component. As seen in Table 3.12, 4.8% of tests in grade 4 failed to include the minimum number of Claim 4 items.

Table 3.11. Percentage of CAT Test Administration Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered – ELA ASL Pool

Grade	Claim	Requirement			% of Tests		
		Page	Min	Max	Under	Match	Above
3	Claim 1: Reading	1	14	16	0.0	100.0	0.0
3	Claim 2: Writing	1	10	10	0.0	100.0	0.0
3	Claim 3: Speaking/Listening	1	8	9	0.0	100.0	0.0
3	Claim 4: Research	1	6	6	0.0	100.0	0.0
4	Claim 1: Reading	1	14	16	0.0	100.0	0.0
4	Claim 2: Writing	1	10	10	0.0	100.0	0.0
4	Claim 3: Speaking/Listening	1	8	9	0.0	100.0	0.0
4	Claim 4: Research	1	6	6	0.0	100.0	0.0
5	Claim 1: Reading	1	14	16	0.0	100.0	0.0
5	Claim 2: Writing	1	10	10	0.0	100.0	0.0
5	Claim 3: Speaking/Listening	1	8	9	0.0	100.0	0.0
5	Claim 4: Research	1	6	6	0.0	100.0	0.0
6	Claim 1: Reading	2	13	17	0.0	100.0	0.0
6	Claim 2: Writing	2	10	10	0.0	100.0	0.0
6	Claim 3: Speaking/Listening	2	8	9	0.0	100.0	0.0
6	Claim 4: Research	2	6	6	0.0	100.0	0.0
7	Claim 1: Reading	2	13	17	0.0	100.0	0.0
7	Claim 2: Writing	2	10	10	0.0	100.0	0.0
7	Claim 3: Speaking/Listening	2	8	9	0.0	100.0	0.0
7	Claim 4: Research	2	6	6	0.0	100.0	0.0
8	Claim 1: Reading	2	13	17	0.0	100.0	0.0
8	Claim 2: Writing	2	10	10	0.0	100.0	0.0
8	Claim 3: Speaking/Listening	2	8	9	0.0	100.0	0.0
8	Claim 4: Research	2	6	6	0.0	100.0	0.0
11	Claim 1: Reading	3	15	16	0.0	100.0	0.0
11	Claim 2: Writing	3	10	10	0.0	100.0	0.0
11	Claim 3: Speaking/Listening	3	8	9	0.0	100.0	0.0
11	Claim 4: Research	3	6	6	0.0	100.0	0.0

Table 3.12. Percentage of PT Test Administration Meeting Blueprint Requirements for Each Claim – ELA ASL Pool

Grade	Claim	Requirement			% of Tests		
		Page	Min	Max	Under	Match	Above
3	Claim 2: Writing	1	3	3	0.0	100.0	0.0
3	Claim 4: Research	1	2	3	0.0	100.0	0.0
4	Claim 2: Writing	1	3	3	0.0	100.0	0.0
4	Claim 4: Research	1	2	3	4.8	95.2	0.0
5	Claim 2: Writing	1	3	3	0.0	100.0	0.0
5	Claim 4: Research	1	2	3	0.0	100.0	0.0
6	Claim 2: Writing	2	3	3	0.0	100.0	0.0
6	Claim 4: Research	2	2	3	0.0	100.0	0.0
7	Claim 2: Writing	2	3	3	0.0	100.0	0.0
7	Claim 4: Research	2	2	3	0.0	100.0	0.0
8	Claim 2: Writing	2	3	3	0.0	100.0	0.0
8	Claim 4: Research	2	2	3	0.0	100.0	0.0
11	Claim 2: Writing	3	3	3	0.0	100.0	0.0
11	Claim 4: Research	3	2	3	0.0	100.0	0.0

Item exposure rates for CAT items are summarized in Table 3.13. Across all grades, at least 95% of all items were administered to fewer than 40% of the simulees, and for all grades, over 90% of items were administered to fewer than 20% of examinees. The number of unused items ranged from 0 to about 3%.

Table 3.13. Item Exposure Rates – ELA ASL Pool

Grade	Total Items	Exposure Rate					
		Unused	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
3	595	1.34	94.45	4.03	0.00	0.00	0.17
4	592	0.00	95.95	3.89	0.00	0.00	0.17
5	591	2.20	94.42	3.05	0.00	0.00	0.34
6	556	2.88	90.65	6.12	0.36	0.00	0.00
7	532	1.88	90.98	7.14	0.00	0.00	0.00
8	520	0.38	91.35	8.08	0.19	0.00	0.00
11	1349	0.30	99.18	0.37	0.15	0.00	0.00

Histograms of exposure rates for the range 0-20% are presented in Figures 3.1 (for grades 3-6) and 3.2 (for grades 7, 8, and 11). In most grades, the exposure rates are fairly dispersed. The exception is grade 11, in which the exposure rate peaks around 5%. This more peaked distribution (and lower average exposure) is due to the fact that the number of available items is much greater in grade 11 than in the other grade levels.

Figure 3.1. Item Exposure Rates (ELA ASL Pool, Grades 3-6)

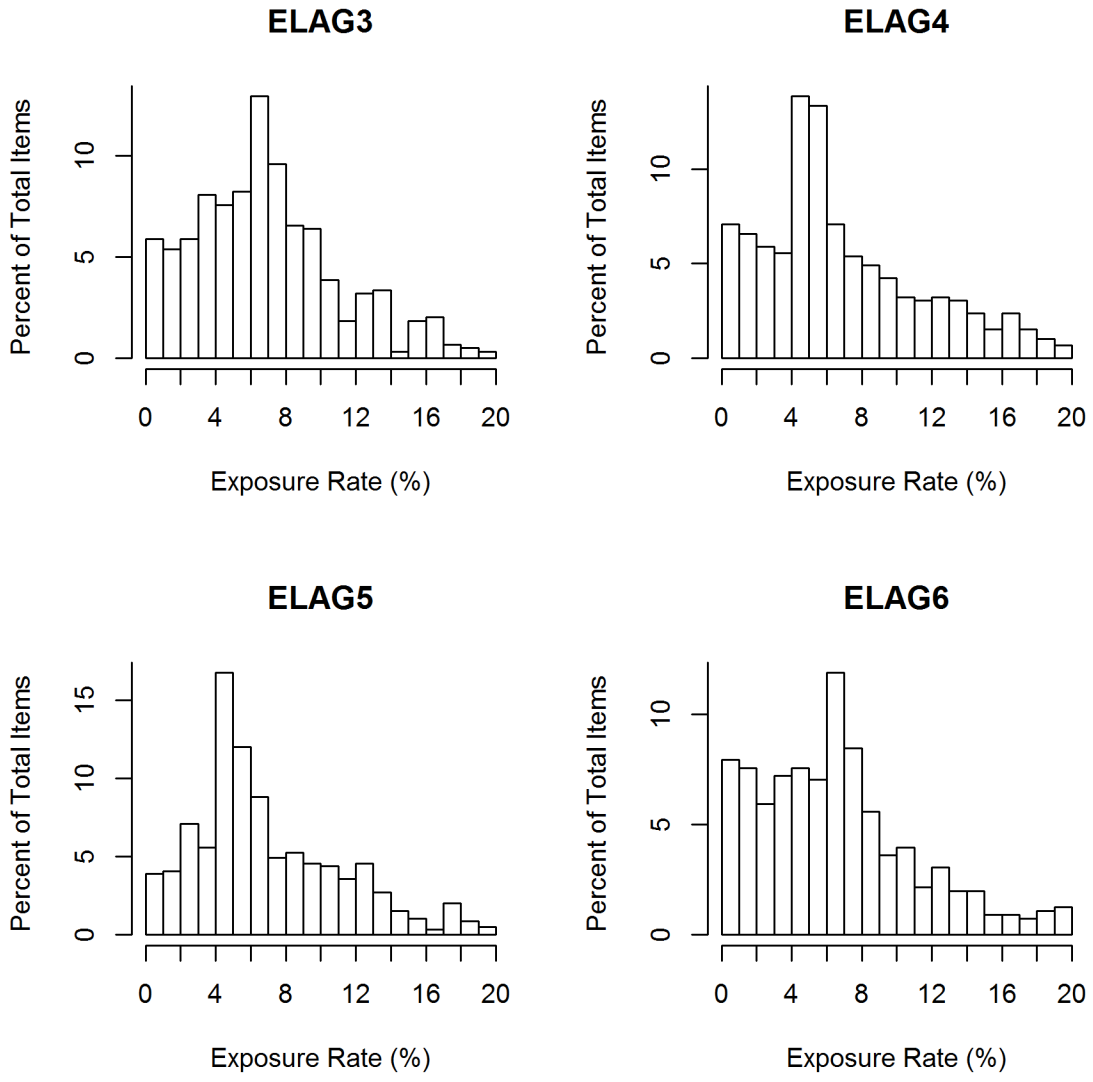
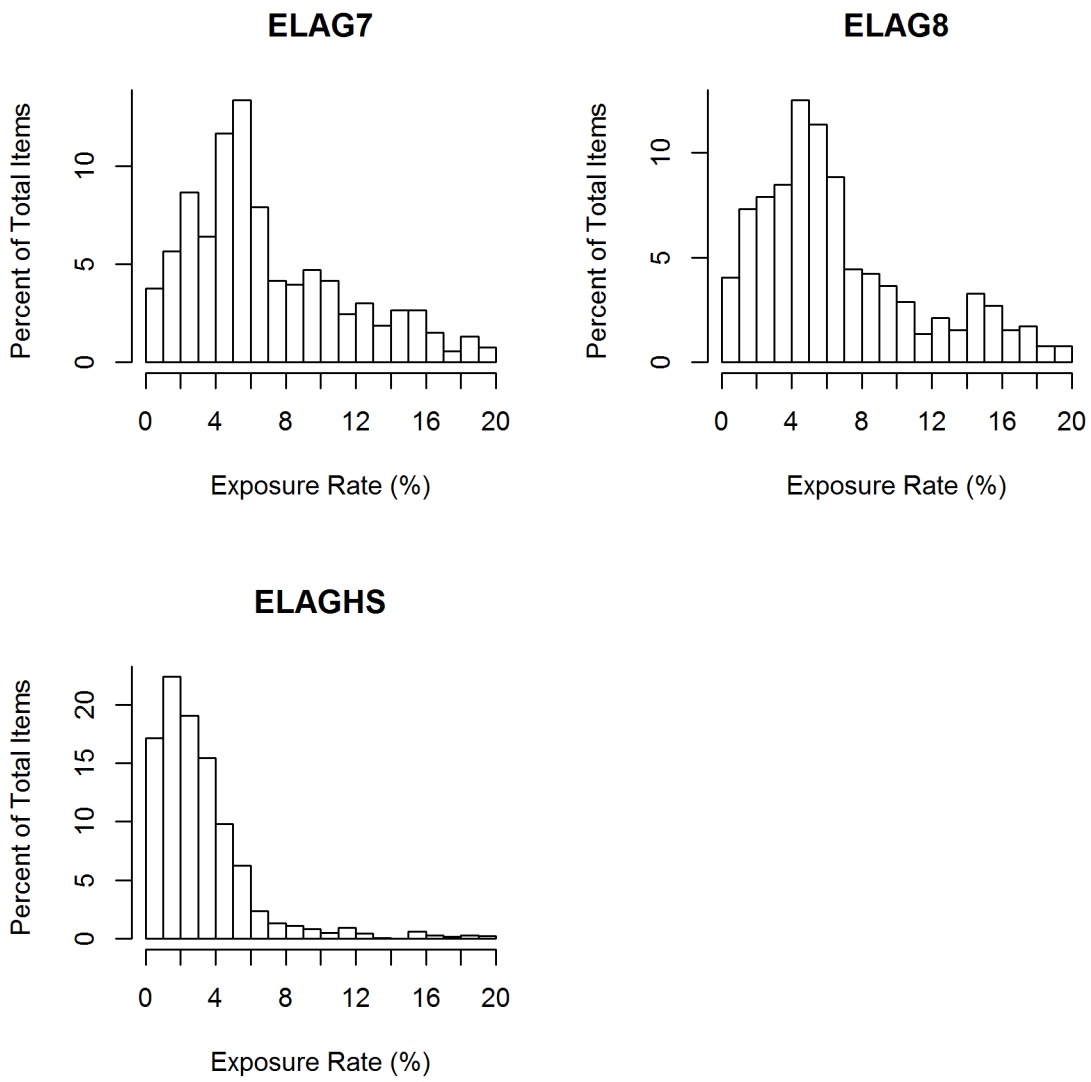


Figure 3.2. Item Exposure Rates (ELA ASL Pool, Grades 7-11)



Chapter 4. Results for the Mathematics ASL Pool

In this chapter, we present the results of the simulated administration of the Mathematics ASL pool. Within each of the seven grade levels, true values for student proficiency (θ) were drawn for 1,000 simulated examinees from a normal distribution with population parameters shown in Table 5.1, below. At the completion of the simulated test administration, some examinee score estimates were infinite (due to having achieved the minimum score on all items or achieving the maximum score on all items) or outside the specified range of obtainable scores. These estimates were assigned the lowest or highest obtainable T-score (LOT and HOT, respectively). Table 4.1 shows the percentage of cases assigned to the LOT or HOT.

Table 4.1. Characteristics of Simulated and Estimated Math Proficiencies

Grade	Population Parameters		Obtainable Range		% Infinite ML Scores		% Outside Range	
	Mean	SD	LOT	HOT	LOT	HOT	LOT	HOT
3	-1.29	0.97	-4.11	1.33	0.2	0.0	0.5	0.6
4	-0.71	1.00	-3.92	1.82	0.4	0.0	0.2	1.2
5	-0.35	1.08	-3.73	2.33	0.6	0.0	0.7	1.2
6	-0.10	1.19	-3.53	2.95	0.5	0.0	0.7	1.1
7	0.01	1.33	-3.34	3.32	0.7	0.0	1.5	1.3
8	0.18	1.42	-3.15	3.63	0.3	0.0	2.6	1.4
11	0.51	1.52	-2.96	4.38	0.7	0.0	3.4	1.0

The numbers of operational CAT and PT items in each grade are summarized in Tables 4.2 and 4.3. Separate counts are provided in Table 4.2 of items in grade levels 6 and above for which use of a calculator either is or is not allowed.

Table 4.2. Number of Operational Adaptive Items - Math ASL Pool

Grade	Calculator	Total	Claim 1	Claim 2	Claim 3	Claim 4
3	No	350	213	44	55	38
4	No	347	207	44	57	39
5	No	377	197	45	72	63
6	Yes	185	84	26	48	27
	No	167	161	0	6	0
7	Yes	245	142	27	48	28
	No	91	91	0	0	0
8	Yes	251	146	18	58	29
	No	77	77	0	0	0
11	Yes	460	253	48	107	52
	No	57	40	0	17	0

Note. Item counts current as of 2015-05-12.

The number of operational CAT items in grades ranges from 57 to 460. The Grade 11 CAT item pool is substantially larger (with 537 items). Across grade levels, Claim 1 items represent 52.3-69.6% of the CAT pool. In operational scoring, it should be noted that items from Claims 2 and 4 are combined to obtain a combined Claim 2/4 score.

The numbers of PT stimuli ranges from 5 (with 27 items) in grade 7 to 9 (with 48 items) in grade 3. As seen in Table 4.3, PT items do not contribute to scores for Claim 1.

Table 4.3. Number of Operational Performance Task Items - Math ASL Pool

Grade	Number of Items					Number of Stimuli (Across Claims)
	Total	Claim 1	Claim 2	Claim 3	Claim 4	
3	48	0	19	15	14	9
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 2016-05-12.

The overall performance of the adaptive test algorithm depends, in part, on the availability of items that are informative at the levels of proficiency found in the examinee population. Table 4.4 shows the means and standard deviations of item difficulty for the CAT and PT portions in each grade band. The means and standard deviations of the final proficiency estimates for the simulated examinees are also shown.

Table 4.4. Math ASL Pool - Item Difficulty and Estimated Student Math Proficiency

Grade	CAT Items		PT Items		Proficiency	
	Mean	SD	Mean	SD	Mean	SD
3	-0.80	1.10	-0.50	0.78	-1.34	1.01
4	0.11	1.03	-0.04	0.89	-0.77	1.06
5	0.70	1.09	0.98	0.74	-0.43	1.17
6	1.11	1.27	0.82	0.84	-0.18	1.27
7	1.88	1.27	1.54	1.20	-0.09	1.41
8	2.35	1.44	2.20	0.56	0.10	1.47
11	2.99	1.54	2.29	0.85	0.38	1.62

The mean difficulty of items increases across grade levels. However, it is notable that average item difficulty is consistently higher than the average proficiency of the

examinees, which could affect the efficiency of the adaptive test. Table 4.5 presents a summary of average bias in the overall and claim score estimates, along with the 95% and 99% confidence interval miss rates.

Table 4.5. Bias of the Estimated Proficiencies – Math ASL Pool

Grade	Bias	SE(bias)	*p*-value	MSE	95% CI Miss Rate	99% CI Miss Rate
			Overall Mathematics			
3	0.00	0.03	0.97	0.07	5.0	1.2
4	0.01	0.03	0.72	0.08	4.6	0.4
5	0.03	0.03	0.43	0.14	5.3	1.0
6	0.01	0.04	0.72	0.12	4.3	1.1
7	0.03	0.04	0.54	0.20	4.7	1.5
8	0.00	0.05	0.97	0.20	4.9	1.4
11	0.05	0.05	0.35	0.29	4.9	1.2
			Claim 1: Concepts and Procedures			
3	-0.02	0.03	0.60	0.13	5.6	0.4
4	0.02	0.03	0.45	0.16	5.3	0.8
5	0.06	0.03	0.08	0.25	4.9	1.1
6	0.02	0.04	0.67	0.21	5.4	0.6
7	0.05	0.04	0.28	0.33	5.8	1.7
8	0.03	0.05	0.48	0.36	4.9	1.3
11	0.08	0.05	0.10	0.57	5.9	2.1
			Claim 2/4: Problem Solving/Modeling and Data Analysis			
3	0.10	0.03	0.00	0.40	8.9	4.2
4	0.17	0.03	0.00	0.71	12.5	7.2
5	0.25	0.04	0.00	0.93	14.9	8.4
6	0.28	0.04	0.00	1.15	17.8	9.7
7	0.39	0.04	0.00	1.61	19.7	10.5
8	0.42	0.05	0.00	1.86	23.3	11.5
11	0.47	0.05	0.00	1.83	19.1	10.2
			Claim 3: Communicating Reasoning			
3	0.23	0.03	0.00	0.74	16.3	11.5
4	0.23	0.03	0.00	0.80	15.3	10.3
5	0.19	0.03	0.00	0.78	11.6	6.3
6	0.30	0.04	0.00	1.15	15.6	8.5
7	0.50	0.05	0.00	2.02	22.4	13.6
8	0.29	0.05	0.00	1.40	11.9	5.6
11	0.17	0.05	0.00	1.09	8.0	3.3

Mean bias in the overall Mathematics score estimates is fairly small for all grade levels (with a range of 0.00 to 0.05 on the logit scale), and the null hypothesis that the mean bias in the overall scores is equal to zero in the population cannot be rejected (*p*-values

are 0.35 or greater).

On the other hand, there is evidence of bias in claim score estimates. This bias appears to be due to the assignment of the LOT and HOT values for examinees with extreme score estimates for a given claim—in particular, those examinees with an infinite ML score estimate due to a perfect score patterns (i.e., achieving either the minimum score for all items or the maximum for all items). Such score patterns are of course far more likely within a claim (based on a relatively small number of items) than for the full test. Importantly, patterns in which all item scores received the minimum were far more frequent than patterns with all items receiving the maximum score. The fact that more infinite scores were replaced with the LOT than with the HOT value resulted in the observed average bias in the claim score results. It should be noted, however, that the assignment to LOT or HOT values would have little impact on the claim-level classifications currently used in practice (below standard, at or near standard, above standard).

Confidence interval miss rates for overall scores are very close to their expected levels. The overall score miss rate for the 95% confidence interval—expected to be 5%—ranges from 4.3% to 5.3%, while the miss rate for the 99% confidence interval—expected to be 1%—ranges from 0.4% to 1.5%. Taken together with the results concerning average bias, these confidence interval miss rates suggest that the standard errors of measurement for the overall score estimates are well-calibrated (i.e., correctly reflecting the level of score uncertainty).

The confidence interval miss rates for the claim scores are less consistent and—for Claim 2/4 and Claim 3, in particular—show evidence of poor calibration. This is not surprising, however, given the bias observed in these claim score estimates. It is likely that the deviations of the miss rates from their expected values are due to the assignment of the LOT and HOT for examinees with perfect item score patterns. Because such patterns are relatively common for the small number of items in a claim, the LOT or HOT is a rather poor estimate of the true score for many examinees. This makes it less likely that the confidence interval around the LOT/HOT will include the true score, increasing the miss rate.

Table 4.6 summarizes the standard deviation in score estimates, average standard error, actual error (RMSE), and marginal reliability for the overall and claim scores.

Table 4.6. Overall Score and Claim Score Precision/Reliability – Math ASL Pool

Grade	mean # Items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$
		Overall Mathematics			
3	39.3	1.0	.25	.26	.94
4	38.7	1.1	.29	.28	.93
5	39.9	1.2	.35	.37	.90
6	38.7	1.3	.36	.35	.92
7	39.4	1.4	.45	.45	.90
8	38.8	1.5	.48	.45	.91
11	41.8	1.6	.55	.54	.89
		Claim 1: Concepts and Procedures			
3	20.0	1.0	.35	.36	.88
4	20.0	1.1	.38	.40	.87
5	20.0	1.2	.48	.50	.84
6	19.0	1.3	.46	.46	.88
7	20.0	1.5	.56	.58	.85
8	20.0	1.5	.62	.60	.85
11	22.0	1.7	.74	.75	.80
		Claim 2/4: Problem Solving/Modeling and Data Analysis			
3	9.7	1.2	.51	.64	.72
4	9.4	1.4	.62	.85	.64
5	9.9	1.6	.65	.96	.62
6	9.7	1.7	.68	1.07	.61
7	10.0	1.9	.82	1.27	.57
8	9.2	2.1	.81	1.36	.56
11	9.5	2.1	.89	1.35	.59
		Claim 3: Communicating Reasoning			
3	9.7	1.4	.60	.86	.61
4	9.4	1.5	.59	.89	.63
5	10.0	1.5	.63	.89	.63
6	10.0	1.7	.80	1.07	.60
7	9.4	1.9	.95	1.42	.45
8	9.7	1.8	1.07	1.18	.58
11	10.3	1.9	1.02	1.04	.69

The standard errors for the overall Mathematics score estimates are well-calibrated; average standard errors within each grade closely resemble the RMSE values. However, there are discrepancies between the average standard errors and the RMSE values for the claim scores. This result is consistent with the earlier findings concerning average bias in the claim score estimates and the confidence interval miss rates (Table 4.5).

Marginal reliability was computed from the RMSE and observed variance in the scale score estimates, as described in Chapter 2. For the overall score, marginal reliability ranged from 0.89 to 0.94. Marginal reliability for the claim scores ranged from 0.80 to 0.88 for Claim 1 (Concepts and Procedures), 0.59 to 0.72 for Claim 2/4 (Problem Solving/Modeling and Data Analysis), and 0.45 to 0.69 for Claim 3 (Communicating Reasoning). The lower levels of marginal reliability for Claim 2/4 and Claim 3 are expected, given that these scores are based on fewer items than the scores for Claim 1.

Table 4.7 summarizes the average standard errors for the overall Mathematics score within true score deciles. The averages in deciles 6-10 (i.e., for all examinees above the median) range from 0.21 and 0.41 for all grade levels. Average standard errors are higher in the lowest deciles and are particularly large in decile 1 for the upper grade levels. This is consistent with the fact that the item pools tend to an average level of difficulty that is higher than the average proficiency of the population of examinees (as seen in Table 4.4). As a result, the administered items contribute less information about examinees with the lowest true scores.

Table 4.7. Average Standard Errors by True Proficiency Decile – Math ASL Pool

Grade	Deciles										Overall
	1	2	3	4	5	6	7	8	9	10	
3	.38	.28	.25	.24	.23	.22	.21	.21	.21	.23	.25
4	.48	.34	.28	.25	.24	.23	.22	.21	.21	.24	.29
5	.60	.43	.34	.31	.29	.26	.24	.22	.21	.22	.35
6	.62	.43	.37	.33	.30	.28	.26	.25	.24	.25	.36
7	.77	.58	.49	.43	.38	.32	.29	.26	.23	.24	.45
8	.80	.61	.52	.46	.41	.37	.33	.30	.26	.25	.48
11	.91	.73	.61	.55	.47	.41	.36	.31	.27	.27	.55

Table 4.8 presents, for each grade level, the correlation between the final score estimates (for overall Math proficiency) and examinee true scores, as well as the correlation between the final score estimates and overall test difficulty (average difficulty for items administered). The correlations between estimated and true proficiencies are quite high (0.94-0.97), indicating that the administered items are successful in recovering the rank ordering of students. Correlations between estimated proficiency and overall test difficulty range from 0.72 to 0.86. These correlations may serve as a crude measure of the extent to which the CAT algorithm tailored the difficulty of the test to examinee, within

the constraints of the blueprint and given the properties of the available pool of items.

Table 4.8. Correlations between True and Estimated Math Proficiency, and between Estimated Proficiency and Overall Test Difficulty – Math ASL Pool

Grade	r($\hat{\theta}, \theta$)	r($\hat{\theta}$,overall test difficulty)
3	.97	.86
4	.96	.85
5	.95	.81
6	.96	.82
7	.95	.72
8	.95	.79
11	.94	.72

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

Tables 4.9-4.12 present results concerning the extent to which simulated tests in each grade level fulfilled requirements of the summative test blueprint. These tables identify the particular blueprint specification (including the page of the blueprint document on which the specification is described) and the range of items that are required in order to fulfill the specification.

Tables 4.9 and 4.10 provide counts of the test instances (out of the 1,000 simulated within the grade level) that violated a specification. As noted in Table 4.9, no violations were identified in the CAT portion of the test.

Table 4.9. Tests with Blueprint Violations, CAT Component – Math ASL Pool

Grade Specification	Requirement			Number of Tests		
	Page	Min	Max	Total	Below	Above
All CAT Specifications Met						

Violations were identified in the PT component for all grade levels except grade 6. These violations are summarized in Table 4.10. The violation included both exceeding the maximum number of items specified in the blueprint and failing to include the minimum number. Specifically, in most cases, the nature of the violation for Claim 2 was exceeding the maximum number of items specified in the blueprint. Its maximum was not met in 12.2% of tests in grade 4, 27.9% of tests in grade 5, and 19.0% of tests in grade 7. The nature of the violation for Claim 4 was failing to include the minimum number of items. This minimum was not met in 11.1% of tests in grade 3, 12.5% of tests in grade 4,

16.8% of tests in grade 8, and 16.3% of the tests in grade 11.

Table 4.10. Tests with Blueprint Violations, PT Component – Math ASL Pool

Grade	Specification	Requirement			Number of Tests		
		Page	Min	Max	Total	Below	Above
3	Claim 2 (Problem Solving) or Claim 4 (Modeling and Data Analysis)	5	2	5	336	112	224
3	Claim 2 (Problem Solving)	5	1	2	459	112	347
3	Claim 4 (Modeling and Data Analysis)	5	1	3	111	111	0
3	Claim 3 (Communicating Reason)	5	0	2	111	0	111
4	Claim 2 (Problem Solving)	7	1	2	122	0	122
4	Claim 4 (Modeling and Data Analysis)	7	1	3	125	125	0
5	Claim 2 (Problem Solving)	9	1	2	279	0	279
7	Claim 2 (Problem Solving)	13	1	2	190	0	190
8	Claim 4 (Modeling and Data Analysis)	15	1	3	168	168	0
11	Claim 4 (Modeling and Data Analysis)	17	1	3	163	163	0
11	Claim 3 (Communicating Reason)	17	0	2	163	0	163

Tables 4.11 and 4.12 present the percentage of test instances that met the blueprint requirements for the total number of items administered within each claim for the CAT and PT components, respectively. As seen in Table 4.11, all tests met the requirements specific to the CAT component. Violations were observed in the PT component for grade 3 to 8, and grade 11. In grade 3, the blueprint was met in 66.4% of tests for Claim 2/4, 54.1% of tests for Claim 2, 88.9% of tests for Claim 4, and 88.9% of tests for Claim 3. In grade 4, the blueprint was met in 87.8% of tests for Claim 2, 87.5% of tests for Claim 4. In grade 5, the blueprint was met in 72.1% of tests for Claim 2. In grade 7, the blueprint was met in 81.0% of tests for Claim 2. In grade 8, the blueprint was met in 83.2% of tests for Claim 4. In grade 11, the blueprint was met in 83.7% of tests for Claim 4, 83.7% of tests for Claim 3.

Table 4.11. Percentage of CAT Test Administration Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered – Math ASL Pool

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

Table 4.12. Tests with Blueprint Violations, PT Component – Math ASL Pool

Grade	Specification	Requirement			% of Tests		
		Page	Min	Under	Under	Match	Above
3	Claim 2 (Problem Solving) or Claim 4 (Modeling and Data Analysis)	5	2	5	11.2	66.4	22.4
3	Claim 2 (Problem Solving)	5	1	2	11.2	54.1	34.7
3	Claim 4 (Modeling and Data Analysis)	5	1	3	11.1	88.9	0.0
3	Claim 3 (Communicating Reason)	5	0	2	0.0	88.9	11.1
4	Claim 2 (Problem Solving)	7	1	2	0.0	87.8	12.2
4	Claim 4 (Modeling and Data Analysis)	7	1	3	12.5	87.5	0.0
5	Claim 2 (Problem Solving)	9	1	2	0.0	72.1	27.9
7	Claim 2 (Problem Solving)	13	1	2	0.0	81.0	19.0
8	Claim 4 (Modeling and Data Analysis)	15	1	3	16.8	83.2	0.0
11	Claim 4 (Modeling and Data Analysis)	17	1	3	16.3	83.7	0.0
11	Claim 3 (Communicating Reason)	17	0	2	0.0	83.7	16.3

Item exposure rates for CAT items are summarized in Table 4.13. Across all grades, at least 94% of all items were administered to fewer than 20% of the simulees; only a very small percentage of the items appeared on more than 40% of the tests. Overall, CAT item exposure was good for grade 3-7, with relatively few items either completely unused or overexposed. However, for grade 8, 28% of the items have an exposure of more than 60%, and for grade 11, 18% of the items have an exposure rate of more than 80%.

Table 4.13. Item Exposure Rates – Math (ASL Pool)

Grade	Total Items	Exposure Rate					
		Unused	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
3	398	0.00	93.72	6.03	0.25	0.00	0.00
4	385	0.00	91.43	8.05	0.52	0.00	0.00
5	418	0.00	93.06	6.46	0.48	0.00	0.00
6	386	0.00	91.97	7.77	0.26	0.00	0.00
7	363	0.00	88.43	10.47	1.10	0.00	0.00
8	357	0.00	88.80	9.80	1.12	0.28	0.00
11	552	0.00	94.20	4.89	0.72	0.00	0.18

Histograms of exposure rates for the range 0-20% are presented in Figures 4.1 (for grades 3-6) and 4.2 (for grades 7, 8, and 11). These histograms make it clear that most items are administered to fewer than 10% of examinees.

Figure 4.1. Exposure Rates (Math ASL Pool, Grades 3-6)

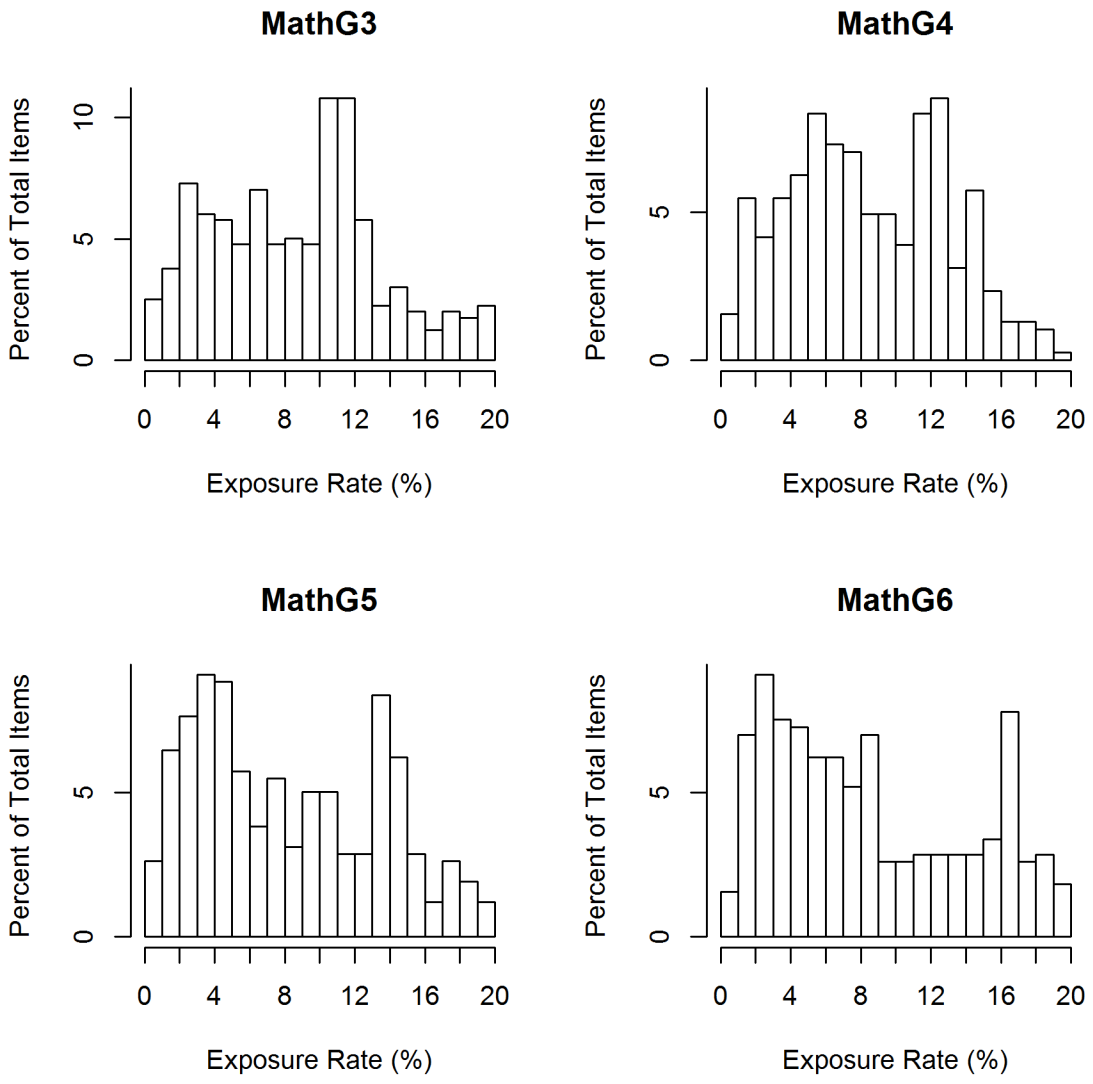
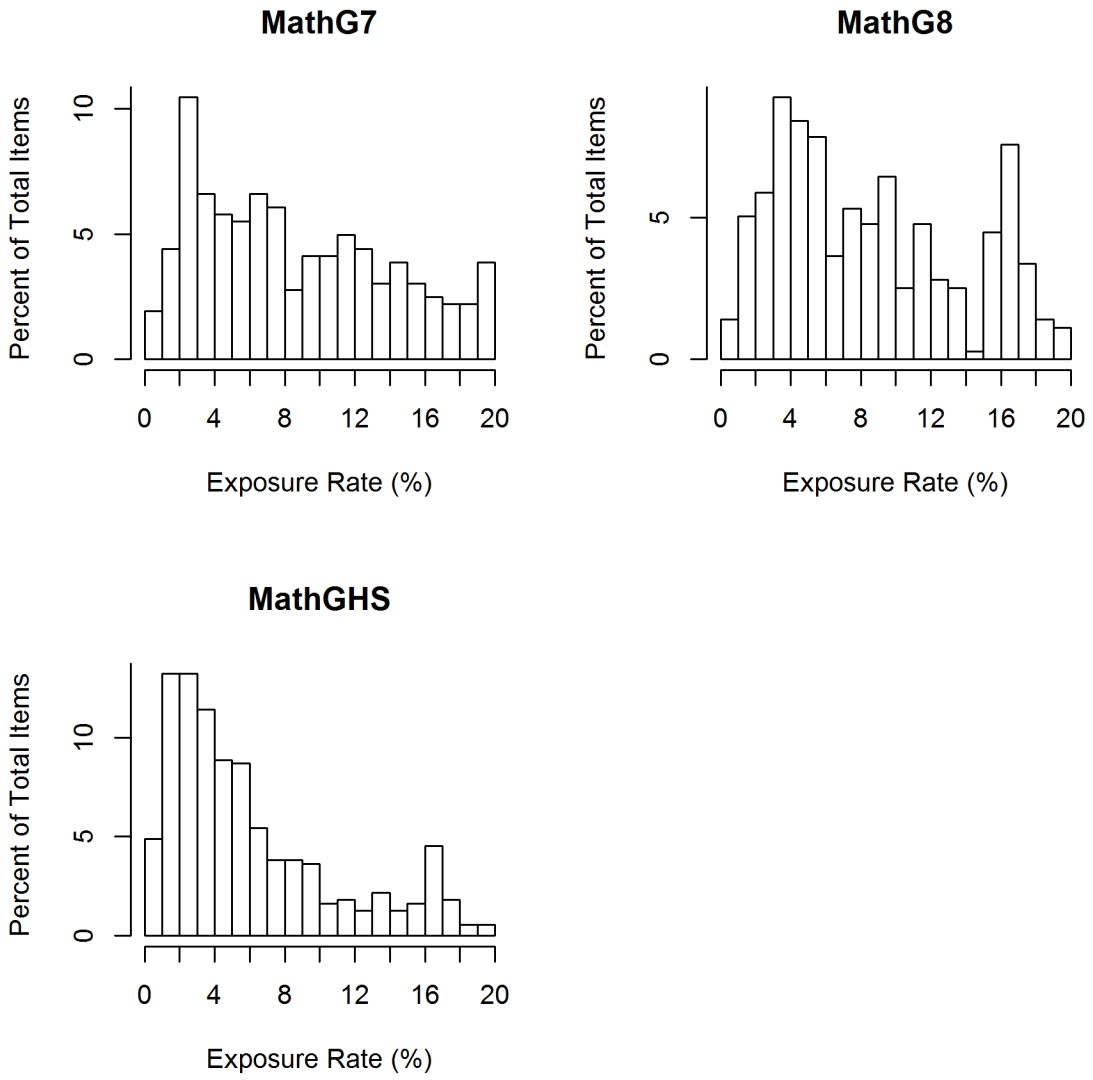


Figure 4.1. Exposure Rates (Math ASL Pool, Grades 7, 8, and HS)



Chapter 5. Results for the Mathematics Translated Glossary Pool

In this chapter, we present the results of the simulated administration of the Mathematics Translated Glossaries pool. Within each of the seven grade levels, true values for student proficiency (θ) were drawn for 1,000 simulated examinees from a normal distribution with population parameters shown in Table 5.1, below. At the completion of the simulated test administration, some examinee score estimates were infinite (due to having achieved the minimum score on all items or achieving the maximum score on all items) or outside the specified range of obtainable scores. These estimates were assigned the lowest or highest obtainable T-score (LOT and HOT, respectively). Table 5.1 shows the percentage of cases assigned to the LOT or HOT.

Table 5.1. Characteristics of Simulated and Estimated Math Proficiencies

Grade	Population Parameters		Obtainable Range		% Infinite ML Scores		% Outside Range	
	Mean	SD	LOT	HOT	LOT	HOT	LOT	HOT
3	-1.29	0.97	-4.11	1.33	0.2	0.0	0.5	0.4
4	-0.71	1.00	-3.92	1.82	0.5	0.0	0.2	1.2
5	-0.35	1.08	-3.73	2.33	0.3	0.0	1.0	1.2
6	-0.10	1.19	-3.53	2.95	0.5	0.0	1.0	1.1
7	0.01	1.33	-3.34	3.32	0.4	0.0	1.9	1.4
8	0.18	1.42	-3.15	3.63	0.5	0.0	2.7	1.2
11	0.51	1.52	-2.96	4.38	0.4	0.0	3.4	0.8

The numbers of operational CAT and PT items in each grade are summarized in Tables 5.2 and 5.3. Separate counts are provided in Table 5.2 of items in grade levels 6 and above for which use of a calculator either is or is not allowed.

Table 5.2. Number of Operational Adaptive Items - Math Translated Glossary Pool

Grade	Calculator	Total	Claim 1	Claim 2	Claim 3	Claim 4
3	No	369	220	49	56	44
4	No	362	210	45	60	47
5	No	367	190	40	72	65
6	Yes	204	86	37	51	30
	No	178	171	0	7	0
7	Yes	247	128	35	55	29
	No	96	96	0	0	0
8	Yes	222	128	14	57	23
	No	60	60	0	0	0
11	Yes	443	236	48	107	52
	No	55	38	0	17	0

Note. Item counts current as of 2016-05-12.

The number of operational CAT items in grades ranges from 55 to 443. The Grade 11 CAT item pool is substantially larger (with 443 items). Across grade levels, Claim 1 items represent 52 -67% of the CAT pool. In operational scoring, it should be noted that items from Claims 2 and 4 are combined to obtain a combined Claim 2/4 score.

The numbers of PT stimuli ranges from 6 (with 32 items, 24 items and 29 items) in grade 6, 7 and 11 to 10 (with 36 items) in grade 8. As seen in Table 5.3, PT items do not contribute to scores for Claim 1.

Table 5.3. Number of Operational Performance Task Items - Math Translated Glossary 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	32	0	11	12	9	6
7	24	0	11	6	7	6
8	36	0	10	13	13	10
11	29	0	9	11	9	6

Note. Item counts current as of 2016-05-12.

The overall performance of the adaptive test algorithm depends, in part, on the availability of items that are informative at the levels of proficiency found in the examinee population. Table 5.4 shows the means and standard deviations of item difficulty for the CAT and PT portions in each grade band. The means and standard deviations of the final proficiency estimates for the simulated examinees are also shown.

Table 5.4. Math Translated Glossary Pool - Item Difficulty and Estimated Student Math Proficiency

Grade	CAT Items		PT Items		Proficiency	
	Mean	SD	Mean	SD	Mean	SD
3	-0.77	1.06	-0.50	0.79	-1.35	1.02
4	0.07	1.02	-0.04	0.89	-0.77	1.06
5	0.70	1.09	0.98	0.74	-0.44	1.17
6	1.11	1.25	0.76	0.83	-0.18	1.27
7	1.88	1.26	1.33	1.30	-0.07	1.38
8	2.38	1.49	2.01	0.80	0.09	1.48
11	2.99	1.54	2.45	0.81	0.38	1.62

The mean difficulty of items increases across grade levels. However, it is notable that average item difficulty is consistently higher than the average proficiency of the examinees, which could affect the efficiency of the adaptive test.

Table 5.5 presents a summary of average bias in the overall and claim score estimates, along with the 95% and 99% confidence interval miss rates.

Table 5.5. Bias of the Estimated Proficiencies – Math (Translated Glossary Pool)

Grade	Bias	SE(bias)	*p*-value	MSE	95% CI Miss Rate	99% CI Miss Rate
			Overall Mathematics			
3	0.01	0.03	0.69	0.07	4.9	1.1
4	0.01	0.03	0.76	0.08	5.6	0.8
5	0.03	0.03	0.34	0.13	4.8	0.9
6	0.02	0.04	0.62	0.13	4.3	0.6
7	0.00	0.04	0.92	0.17	4.0	1.0
8	0.00	0.05	0.92	0.21	5.2	1.1
11	0.04	0.05	0.38	0.30	5.0	0.9
			Claim 1: Concepts and Procedures			
3	0.00	0.03	0.92	0.13	5.4	0.7
4	0.02	0.03	0.49	0.14	4.9	0.6
5	0.07	0.03	0.05	0.25	4.4	1.1
6	0.02	0.03	0.64	0.22	4.8	0.9
7	0.03	0.04	0.48	0.32	5.6	2.0
8	0.04	0.05	0.43	0.33	4.3	0.7
11	0.09	0.05	0.07	0.57	6.3	1.8
			Claim 2/4: Problem Solving/Modeling and Data Analysis			
3	0.10	0.03	0.00	0.40	10.7	5.2
4	0.17	0.03	0.00	0.72	13.2	6.8
5	0.28	0.04	0.00	1.03	16.2	9.4
6	0.31	0.04	0.00	1.22	18.3	10.6
7	0.31	0.04	0.00	1.35	15.9	8.6
8	0.32	0.05	0.00	1.58	18.9	10.3
11	0.46	0.05	0.00	1.93	20.0	10.1
			Claim 3: Communicating Reasoning			
3	0.25	0.03	0.00	0.72	15.6	10.9
4	0.18	0.03	0.00	0.69	13.5	8.0
5	0.18	0.03	0.00	0.69	10.4	5.1
6	0.30	0.04	0.00	0.13	14.0	8.7
7	0.30	0.04	0.00	1.52	14.7	9.1
8	0.27	0.05	0.00	1.59	12.3	7.7
11	0.18	0.05	0.00	1.19	8.6	3.8

Mean bias in the overall Mathematics score estimates is fairly small for all grade levels (with a range of 0.00 to 0.04 on the logit scale), and the null hypothesis that the mean bias in the overall scores is equal to zero in the population cannot be rejected (*p*-values are 0.34 or greater).

On the other hand, there is evidence of bias in claim score estimates. This bias appears to be due to the assignment of the LOT and HOT values for examinees with extreme score estimates for a given claim—in particular, those examinees with an infinite ML score estimate due to a perfect score patterns (i.e., achieving either the minimum score for all items or the maximum for all items). Such score patterns are of course far more likely within a claim (based on a relatively small number of items) than for the full test. Importantly, patterns in which all item scores received the minimum were far more frequent than patterns with all items receiving the maximum score. The fact that more infinite scores were replaced with the LOT than with the HOT value resulted in the observed average bias in the claim score results. It should be noted, however, that the assignment to LOT or HOT values would have little impact on the claim-level classifications currently used in practice (below standard, at or near standard, above standard).

Confidence interval miss rates for overall scores are very close to their expected levels. The overall score miss rate for the 95% confidence interval—expected to be 5%—ranges from 4.0% to 5.6%, while the miss rate for the 99% confidence interval—expected to be 1%—ranges from 0.6% to 1.1%. Taken together with the results concerning average bias, these confidence interval miss rates suggest that the standard errors of measurement for the overall score estimates are well-calibrated (i.e., correctly reflecting the level of score uncertainty).

The confidence interval miss rates for the claim scores are less consistent and—for Claim 2/4 and Claim 3, in particular—show evidence of poor calibration. This is not surprising, however, given the bias observed in these claim score estimates. It is likely that the deviations of the miss rates from their expected values are due to the assignment of the LOT and HOT for examinees with perfect item score patterns. Because such patterns are relatively common for the small number of items in a claim, the LOT or HOT is a rather poor estimate of the true score for many examinees. This makes it less likely that the

confidence interval around the LOT/HOT will include the true score, increasing the miss rate. Table 5.6 summarizes the standard deviation in score estimates, average standard error, actual error (RMSE), and marginal reliability for the overall and claim scores.

Table 5.6. Overall Score and Claim Score Precision/Reliability – Math Translated Glossary Pool

Grade	mean # Items	SD($\hat{\theta}$)	mean SE($\hat{\theta}$)	RMSE	$\bar{\rho}$
		Overall Mathematics			
3	39.9	1.0	.26	.26	.93
4	38.7	1.1	.29	.28	.93
5	39.9	1.2	.34	.36	.91
6	38.3	1.3	.36	.36	.92
7	38.0	1.4	.43	.41	.91
8	37.6	1.5	.48	.46	.91
11	40.8	1.6	.56	.55	.89
		Claim 1: Concepts and Procedures			
3	20.0	1.1	.35	.37	.88
4	20.0	1.1	.38	.38	.88
5	20.0	1.2	.48	.50	.84
6	19.0	1.3	.46	.47	.87
7	20.0	1.4	.55	.56	.85
8	20.0	1.5	.60	.57	.86
11	22.0	1.7	.75	.76	.80
		Claim 2/4: Problem Solving/Modeling and Data Analysis			
3	10.0	1.2	.50	.63	.73
4	9.4	1.4	.61	.85	.64
5	9.9	1.6	.64	1.01	.60
6	9.3	1.8	.68	1.10	.60
7	9.0	1.8	.82	1.16	.60
8	8.3	2.0	.86	1.26	.59
11	9.0	2.1	.91	.39	.58
		Claim 3: Communicating Reasoning			
3	9.9	1.4	.61	.85	.62
4	9.4	1.4	.59	.83	.66
5	10.0	1.4	.64	.83	.66
6	10.0	1.7	.80	1.06	.59
7	9.0	1.8	.96	1.23	.53
8	9.3	1.9	1.12	1.26	.54
11	9.8	1.9	1.05	1.09	.66

The standard errors for the <u>overall</u> Mathematics score estimates are well-calibrated; average standard errors within each grade closely resemble the RMSE values. However, there are discrepancies between the average standard errors and the RMSE values for

the claim scores. This result is consistent with the earlier findings concerning average bias in the claim score estimates and the confidence interval miss rates (Table 5.5).

Marginal reliability was computed from the RMSE and observed variance in the scale score estimates, as described in Chapter 2. For the overall score, marginal reliability ranged from 0.89 to 0.93. Marginal reliability for the claim scores ranged from 0.80 to 0.88 for Claim 1 (Concepts and Procedures), 0.58 to 0.73 for Claim 2/4 (Problem Solving/Modeling and Data Analysis), and 0.53 to 0.66 for Claim 3 (Communicating Reasoning). The lower levels of marginal reliability for Claim 2/4 and Claim 3 are expected, given that these scores are based on fewer items than the scores for Claim 1.

Table 5.7 summarizes the average standard errors for the overall Mathematics score within true score deciles. The averages in deciles 6-10 (i.e., for all examinees above the median) range from 0.21 and 0.42 for all grade levels. Average standard errors are higher in the lowest deciles and are particularly large in decile 1 for the upper grade levels. This is consistent with the fact that the item pools tend to an average level of difficulty that is higher than the average proficiency of the population of examinees (as seen in Table 5.4). As a result, the administered items contribute less information about examinees with the lowest true scores.

Table 5.7. Average Standard Errors by True Proficiency Decile – Math Translated Glossary Pool

Grade	Deciles										Overall
	1	2	3	4	5	6	7	8	9	10	
3	0.39	0.28	0.25	0.24	0.23	0.22	0.21	0.21	0.21	0.23	0.26
4	0.47	0.34	0.28	0.25	0.24	0.23	0.22	0.22	0.22	0.24	0.29
5	0.59	0.43	0.34	0.30	0.28	0.26	0.24	0.22	0.21	0.22	0.34
6	0.61	0.42	0.37	0.33	0.31	0.29	0.27	0.25	0.24	0.25	0.36
7	0.70	0.54	0.46	0.42	0.37	0.33	0.30	0.26	0.24	0.25	0.43
8	0.78	0.60	0.52	0.46	0.42	0.38	0.35	0.31	0.28	0.27	0.48
11	0.91	0.73	0.62	0.56	0.48	0.42	0.37	0.32	0.28	0.27	0.56

Table 5.8 presents, for each grade level, the correlation between the final score estimates (for overall Math proficiency) and examinee true scores, as well as the correlation between the final score estimates and overall test difficulty (average difficulty for items administered). The correlations between estimated and true proficiencies are quite high

(0.94-0.97), indicating that the administered items are successful in recovering the rank ordering of students. Correlations between estimated proficiency and overall test difficulty range from 0.73 to 0.86. These correlations may serve as a crude measure of the extent to which the CAT algorithm tailored the difficulty of the test to examinee, within the constraints of the blueprint and given the properties of the available pool of items.

Table 5.8. Correlations between True and Estimated Math Proficiency, and between Estimated Proficiency and Overall Test Difficulty – Math Translated Glossary Pool

Grade	$r(\hat{\theta}, \theta)$	$r(\hat{\theta}$,overall test difficulty)
3	0.97	0.86
4	0.96	0.84
5	0.95	0.81
6	0.96	0.83
7	0.96	0.77
8	0.95	0.77
11	0.94	0.73

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

Tables 5.9-5.12 present results concerning the extent to which simulated tests in each grade level fulfilled requirements of the summative test blueprint. These tables identify the particular blueprint specification (including the page of the blueprint document on which the specification is described) and the range of items that are required in order to fulfill the specification.

Tables 5.9 and 5.10 provide counts of the test instances (out of the 1,000 simulated within the grade level) that violated a specification. As noted in Table 5.9, no violations were identified in the CAT portion of the test.

Table 5.9. Tests with Blueprint Violations, CAT Component – Math Translated Glossary Pool

Grade Specification	Requirement			Number of Tests		
	Page	Min	Max	Total	Below	Above
All CAT Specifications Met						

Violations were identified in the PT component for all grade levels except grade 6. These

violations are summarized in Table 5.10. The violation included both exceeding the maximum number of items specified in the blueprint and failing to include the minimum number. Specifically, in grades 3, 4, 5 and 7, the nature of the violation for Claim 2 was exceeding the maximum number of items specified in the blueprint. Its maximum was exceeded in 38.7% of tests in grade 3, 12.2% of tests in grade 4, and 27.9% of tests in grade 5, and 15.6% of the tests in grade 7. For grades 8 and 11, the nature of the violation for Claim 2 was failing to include the minimum number of items specified in the blueprint. Its minimum was not met in 40.5% of tests in grade 8, and 16.3% of tests in grade 11. The nature of the violation for Claim 4 was failing to include the minimum number of items. This minimum was not met in 12.4%% of tests in grade 3, 12.5% of tests in grade 4, 33.2% of tests in grade 7, 19.3% of tests in grade 8, and 16.3% of the tests in grade 11.

Table 5.10. Tests with Blueprint Violations, PT Component – Math Translated Glossary Pool

Grade	Specification	Requirement			Number of Tests		
		Page	Min	Max	Total	Below	Above
3	Claim 2 (Problem Solving) or Claim 4 (Modeling and Data Analysis)	5	2	5	274	0	274
3	Claim 2 (Problem Solving)	5	1	2	387	0	387
3	Claim 4 (Modeling and Data Analysis)	5	1	3	124	124	0
3	Claim 3 (Communicating Reason)	5	0	2	124	0	124
4	Claim 2 (Problem Solving)	7	1	2	122	0	122
4	Claim 4 (Modeling and Data Analysis)	7	1	3	125	125	0
5	Claim 2 (Problem Solving)	9	1	2	279	0	279
7	Claim 2 (Problem Solving) or Claim 4 (Modeling and Data Analysis)	13	2	5	176	176	0
7	Claim 2 (Problem Solving)	13	1	2	156	0	156
7	Claim 4 (Modeling and Data Analysis)	13	0	2	332	332	0
8	Claim 2 (Problem Solving)	15	1	2	405	405	0
8	Claim 4 (Modeling and Data Analysis)	15	1	3	193	193	0
11	Claim 2 (Problem Solving)	17	1	2	163	163	0
11	Claim 4 (Modeling and Data Analysis)	17	1	3	163	163	0

Tables 5.11 and 5.12 present the percentage of test instances that met the blueprint requirements for the total number of items administered within each claim for the CAT and PT components, respectively. Note that the main difference in these results from those presented in Tables 5.9 and 5.10 are that separate requirements for Claims 2 and 4 are not considered (the focus here is on the required number of items for producing the claim score). As seen in Table 5.11, all tests met the requirements specific to the CAT

component. Violations in the number of PT items administered in Claims 2/4 and 3 were observed in grades 3 and 7. In grade 3, 27.4% of tests exceeded the specified number of Claim 2/4 items and 12.4% exceeded the requirements for Claim 3. In grade 7, 17.6% of tests failed to meet the requirement for Claim 2/4.

Table 5.11. Percentage of CAT Test Administration Meeting Blueprint Requirements for Each Claim and the Number of Passages Administered – Math Translated Glossary Pool

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

Table 5.12. Percentage of PT Test Administration Meeting Blueprint Requirements for Each Claim – Math (Translated Glossary Pool)

Grade	Claim	Requirement			% of Tests		
		Page	Min	Max	Under	Match	Above
3	Claim 2/4: Problem Solving/Modeling and Data Analysis	5	2	5	0.0	72.6	27.4
3	Claim 2 (Problem Solving)	5	1	2	0.0	61.3	38.7
3	Claim 4 (Modeling and Data Analysis)	5	1	3	12.4	87.6	0.0
3	Claim 3: Communicating Reasoning	5	0	2	0.0	87.6	12.4
4	Claim 2 (Problem Solving)	7	1	2	0.0	87.8	12.2
4	Claim 4 (Modeling and Data Analysis)	7	1	3	12.5	87.5	0.0
5	Claim 2 (Problem Solving)	9	1	2	0.0	72.1	27.9
7	Claim 2 (Problem Solving) or Claim 4 (Modeling and Data Analysis)	13	2	5	17.6	82.4	0.0
7	Claim 2 (Problem Solving)	13	1	2	0.0	84.4	15.6
7	Claim 4 (Modeling and Data Analysis)	13	0	2	33.2	66.8	0.0
8	Claim 2 (Problem Solving)	15	1	2	40.5	59.5	0.0
8	Claim 4 (Modeling and Data Analysis)	15	1	3	19.3	80.7	0.0
11	Claim 2 (Problem Solving)	17	1	2	16.3	83.7	0.0
11	Claim 4 (Modeling and Data Analysis)	17	1	3	16.3	83.7	0.0

Item exposure rates for CAT items are summarized in Table 5.13. Across all grades, at least 86% of all items were administered to fewer than 20% of the simulees; only a very small percentage of the items appeared on more than 40% of the tests. Overall, CAT item exposure was good for grade 3-7, with relatively few items either completely unused or overexposed. However, for grade 8, 31% of the items have an exposure of more than 60%, and for grade 11, 19% of the items have an exposure rate of more than 80%.

Table 5.13. Item Exposure Rates – Math Translated Glossary Pool

Grade	Total Items	Exposure Rate					
		Unused	0%-20%	21%-40%	41%-60%	61%-80%	81%-100%
3	416	0	94.23	5.77	0	0	0
4	400	0	92.75	7.00	.25	0	0
5	408	0	93.14	6.37	.49	0	0
6	414	0	93.48	6.52	.00	0	0
7	367	0	90.74	8.99	.27	0	0
8	318	0	85.53	12.89	1.26	0	.31
11	527	0	93.93	5.12	.76	0	.19

Histograms of exposure rates for the range 0-20% are presented in Figures 5.1 (for grades 3-6) and 5.2 (for grades 7, 8, and 11). These histograms make it clear that most items are administered to fewer than 10% of examinees.

Figure 5.1. Exposure Rates (Math Translated Glossary Pool, Grades 3-6)

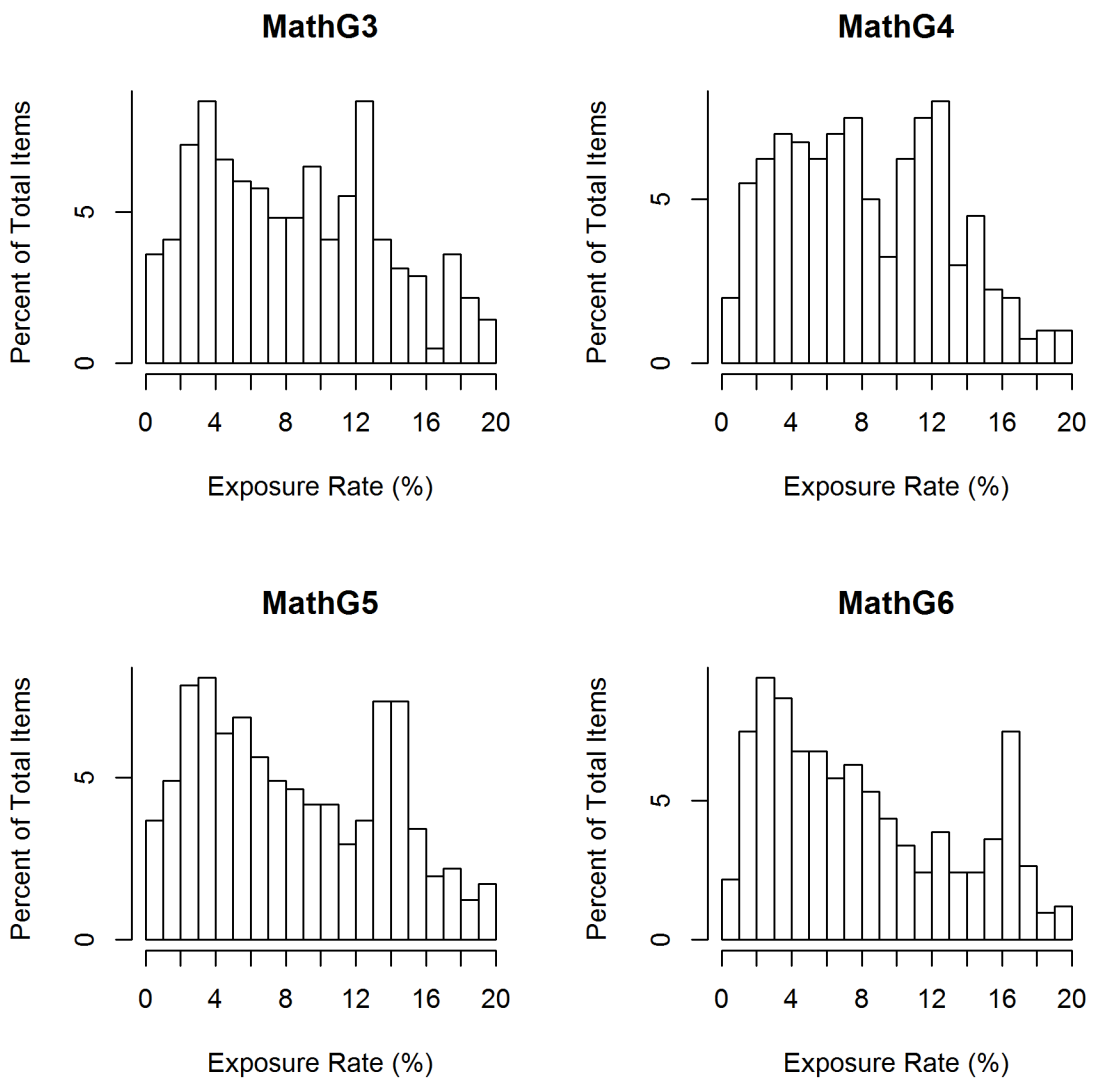
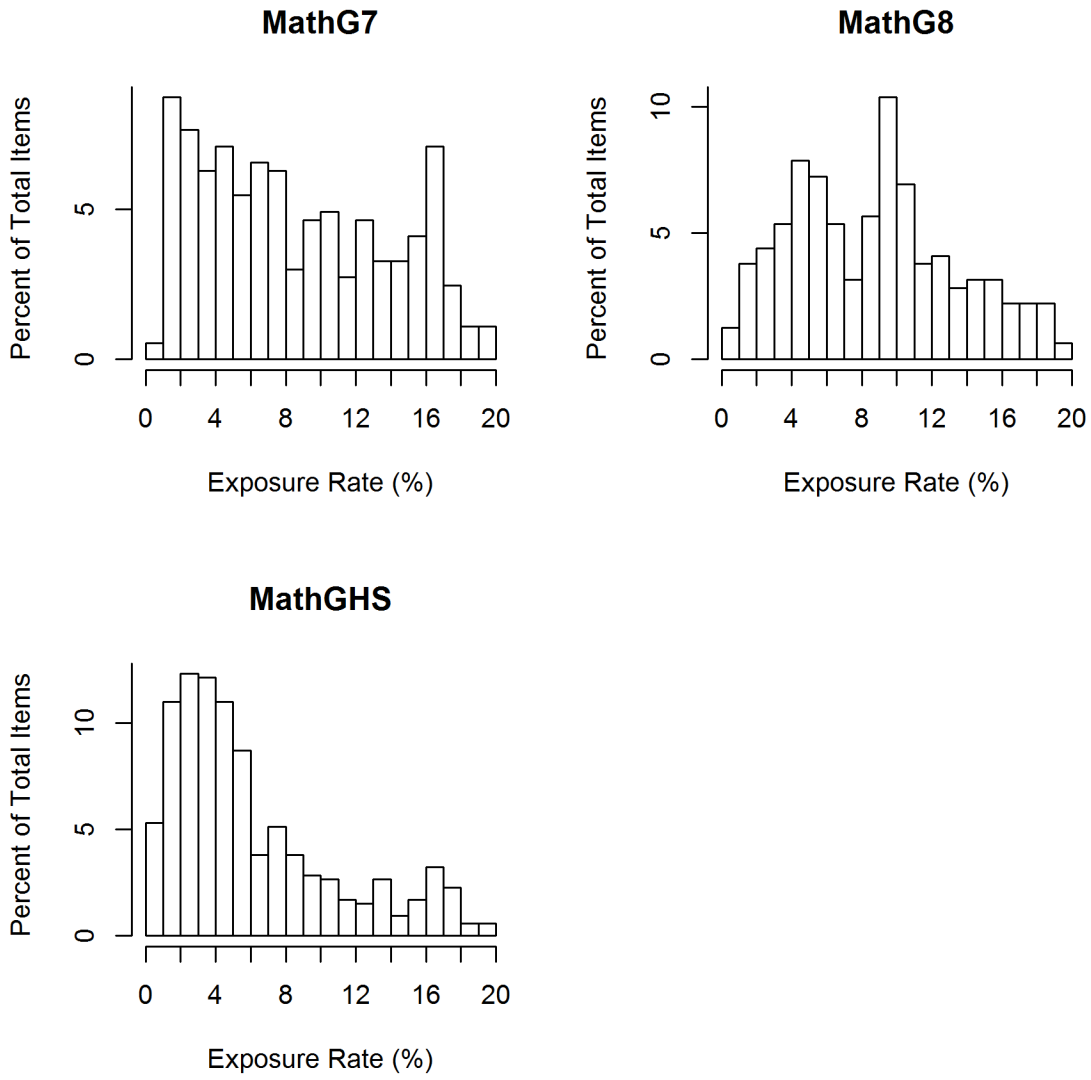


Figure 5.2. Exposure Rates (Math Translated Glossary Pool, Grades 7, 8, and HS)



References

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